On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism

by

LUIS ALBERTO BARRÓN CEDEÑO

Departmento de Sistemas Informáticos y Computación

Universitat Politècnica de València

THESIS

submitted for the degree of

Philosophiæ Doctor

(Computer Science)

Under the Supervision of

Dr. Paolo Rosso

June 2012

Members of the examining committee

Ricardo Baeza-Yates (Yahoo Research!, Spain)

Benno Stein (Bauhaus-Universität Weimar, Germany)

Paul Clough (University of Sheffield, UK)

Fabio Crestani (Università della Svizzera italiana, Switzerland)

José Miguel Benedí Ruiz (Universitat Politècnica de València, Spain)

On the book cover: the word 2-grams of this dissertation (2-grams with two stopwords were discarded). The size of a 2-gram corresponds to its frequency in the text. Word cloud created with Wordle (`http://www.wordle.net/`).

To Silvia and Cristina

Many students seem to almost reflexively embrace a philosophy rooted in the subculture of computer hackers: that all information is, or should be, free for the taking

Mark Fritz

Acknowledgements

I want to dedicate a couple of pages to express my gratitude to all the people that supported me during the development of this research work.

My gratitude to Paolo Rosso for investing hours and hours of hard work in advising me and teaching me. Paolo is the responsible for turning me from a student into a researcher. Thank you for all the opportunities, discussions, and nice time together. You made this outcome possible.

My thanks to the researchers, students, and managers at the Department of Information Systems and Computation. In particular to David Pinto, who instructed me from the perspective of the mature PhD student, explaining the Maths and codes in a way much simpler than in class. Thanks to José Miguel Benedí and Alfons Juan for teaching me the statistical point of view of NLP. Also to Gustavo Rovelo, Parth Gupta, Santiago Mola, Sandra García, Enrique Flores, and Lidia Moreno.

I also thank Paul Clough for receiving me at the University of Sheffield during the summer of 2010, allowing me to spend four inspiring months in South Yorkshire. I learned a lot from Paul, Evangelos Kanoulas, and Monica Paramita. Also thanks to the members of the Accurat Project for allowing me to interact with them.

My gratitude to Benno Stein, Martin Potthast, and Andreas Eiselt for giving me the opportunity to work together. Thank you for your support in practically every different stage of this work and for the precise feedback, which has made me a more mature researcher and person.

Thanks to the other organisers of the PAN initiatives: International Competition on Plagiarism Detection and Cross-Language !ndian Text Re-Use. Helping in the organisation of such events is not always possible and I am happy to be part of them. Thanks to the participants in both competitions, specially to Diego Rodríguez Torrejón; your hard work has given a meaning to our efforts.

I also want to acknowledge the people that have allowed me to remain inside of a multi-disciplinary world. On the one side, Marta Vila and Toni Martì took me back to the linguistic point of view of NLP, where text re-use is seen as paraphrasing. On the other side, with the call of “let’s do some Maths”, Mirko Degli Esposti and Chiara Basile took me to the abstract side of NLP, where everything can become a number. I thank the Know-Center GmbH and particularly to Michael Granitzer for his support during

the last months. Also thanks to my research links with Mexico: Grigori Sidorov, Adelina Escobar, and Alfonso Medina.

Thanks to Paul Clough, Benno Stein, and Ricardo Baeza-Yates, members of the dissertation committee, for their valuable feedback. The three of them, together with José Miguel Benedí and Fabio Crestani, are the members of my examining committee. Thank you all for being with me during the culmination of this stage of my life. Also thanks to Alistair Moffat and Noriko Kando for their advice during the SIGIR 2010 Doctoral Consortium.

All my gratitude to Silvia for staying with me during this time. Thank you for believing in me and leaving it all behind to share this adventure. Your love, support, commitment, and complicity have been what I needed to keep going during these years. This stage is coming to an end and I hope we can stay together in the next one. My thanks to my Mum for being permanently with me regardless of the spatial distance. It is amazing how the feelings can easily cross such a big ocean. I hope the sacrifice of being away has been worth and that you feel proud of me. Also thanks to all our family members and friends in America (the continent) for your support; particularly to Rosa, Hesiquio, and Patrick.

Thanks to those who have shared with us the experience of living abroad: Gustavo, Andreas, Leticia, Antonio, David and family, Valeria, Alessandra, Jan, Armando, Carlos, Adriana and German, Ihab, Diego, Hedy, and Tad. And also to those who, without having been (precisely) abroad, opened their doors to share a part of their lives with us: Paolo and family, Afif and family, Ana, Isabel and César, José Ramón and family, Maria Ángeles, and Edu. I apologise if I forgot to include some name. Fortunately you are many and it is hard to remember everybody.

Finally, my thanks to the National Council of Science and Technology of Mexico for funding my PhD studies through the 192021/302009 scholarship, and to the Ministry of Education of Spain for supporting my internship in the University of Sheffield through the TME2009-00456 grant. This research has been carried out in the framework of the following projects: (*i*) MICINN research project TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 and (*ii*) VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems

Valencia, June 2012

Alberto Barrón Cedeño

Agradecimientos

Quiero utilizar un par de páginas para expresar mi gratitud a todas las personas que han estado conmigo durante el desarrollo de esta investigación.

Agradezco a Paolo Rosso por invertir horas y horas de intenso trabajo asesorándome y enseñándome. Paolo es el responsable de llevarme de ser un estudiante a un investigador. Gracias por todas las oportunidades, discusiones, y los buenos tiempos. Tú has hecho esto posible.

Mi agradecimiento a los investigadores, estudiantes y administradores del Departamento de Sistemas Informáticos y Computación. En particular a David Pinto, quien me instruyó desde la perspectiva del estudiante de doctorado experimentado, explicando las matemáticas y los códigos de una manera mucho más simple que en clase. Gracias a José Miguel Benedí y Alfons Juan por enseñarme el punto de vista estadístico del PLN. También a Gustavo Rovelo, Parth Gupta, Santiago Mola, Sandra García, Enrique Flores y Lidia Moreno.

Igualmente agradezco a Paul Clough por recibirme en la Universidad de Sheffield durante el verano de 2010, permitiéndome pasar cuatro inspiradores meses en el sur de Yorkshire. Aprendí mucho de Paul, Evangelos Kanoulas y Monica Paramita. También gracias a los miembros del proyecto Accurat por permitirme interactuar con ellos.

Mi gratitud a Benno Stein, Martin Potthast y Andreas Eiselt por darme la oportunidad de trabajar juntos. Gracias por el respaldo en prácticamente todas las etapas de este trabajo y por las opiniones precisas, que me han hecho ser un investigador (y persona) más maduro.

Gracias a los otros organizadores de las iniciativas del PAN: *International Competition on Plagiarism Detection* y *Cross-Language !ndian Text Re-Use*. Ayudar en la organización de estos eventos no es siempre posible y estoy feliz de haber sido parte de ellos. Gracias a los participantes de ambas competencias, especialmente a Diego Rodríguez Torrejón; su trabajo ha dado significado a nuestros esfuerzos.

También quiero agradecer a las personas que me han permitido seguir dentro de un mundo multidisciplinario. Por un lado, Marta Vila y Toni Martì me llevaron de regreso al punto de vista lingüístico del PLN, donde la reutilización de texto es vista como una paráfrasis. Por el otro, al grito de "vamos a hacer matemáticas", Mirko Degli Esposti y Chiara Basile me llevaron al lado abstracto del PLN, donde todo puede volverse un

número. Gracias también al *Know-Center GmbH*, en particular a Michael Granitzer por su apoyo durante los últimos meses. Y, por supuesto, a mis enlaces con la investigación en México: Grigori Sidorov, Adelina Escobar y Alfonso Medina.

Gracias a Paul Clough, Benno Stein y Ricardo Baeza-Yates, miembros del comité evaluador de la tesis, por sus valiosos comentarios y sugerencias. Ellos tres, así como José Miguel Benedí y Fabio Crestani, constituyen el comité examinador en mi defensa. Gracias a todos por estar conmigo en la culminación de esta etapa de mi vida. También gracias a Alistair Moffat y Noriko Kando por sus consejos en el consorcio doctoral del SIGIR 2010.

Toda mi gratitud a Silvia por estar conmigo durante este tiempo. Gracias por creer en mí y dejar todo atrás para compartir esta aventura. Tu amor, apoyo, compromiso y complicidad han sido lo que necesitaba para seguir durante estos años. Esta etapa está llegando a su final y espero que podamos seguir juntos en la próxima. Gracias a mi mamá por estar siempre conmigo, sin importar la distancia espacial. Es impresionante cómo los sentimientos pueden cruzar un océano tan grande con tanta facilidad. Espero que el sacrificio de estar lejos haya valido la pena y que te sientas orgullosa de mí. Mi agradecimiento a todos los miembros de nuestra familia y a nuestros amigos en América; particularmente a Rosa, Hesiquio y Patrick.

Gracias a aquellos que han compartido con nosotros la experiencia de vivir en el extranjero: Gustavo, Andreas, Leticia, Antonio, David y su familia, Valeria, Alessandra, Jan, Armando, Carlos, Adriana y German, Ihab, Diego, Hedy y Tad. También a aquellos que, sin haber estado (precisamente) en el extranjero, abrieron sus puertas para compartir una parte de sus vidas con nosotros: Paolo y su familia, Afif y su familia, Ana, Isabel y César, José Ramón y su familia, Maria Ángeles, y Edu. Pido disculpas si he olvidado incluir algún nombre. Afortunadamente son muchos y es difícil recordar a todos.

Finalmente, agradezco al Consejo Nacional de Ciencia y Tecnología de México por financiar mis estudios de doctorado a través de la beca 192021/302009, así como al Ministerio de Educación de España por apoyarme para realizar una estancia en la Universidad de Sheffield, por medio de la beca TME2009-00456. Esta investigación se ha llevado a cabo en el marco de los siguientes proyectos: (*i*) Proyecto MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 y (*iii*) VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Valencia, junio de 2012

Alberto Barrón Cedeño

Abstract

Automatic text re-use detection is the task of determining whether a text has been produced by considering another as its source. Plagiarism, the unacknowledged re-use of text, is probably the most famous kind of re-use. Favoured by the easy access to information through electronic media, plagiarism has raised in recent years, requesting for the attention of experts in text analysis.

Automatic text re-use detection takes advantage of technology on natural language processing and information retrieval in order to compare thousands of documents, looking for the potential source of a presumably case of re-use. Machine translation technology can be used in order to uncover cases of cross-language text re-use. By exploiting such technology, thousands of exhaustive comparisons are possible, also across languages, something impossible to do manually.

In this dissertation we pay special attention to three types of text re-use, namely: (*i*) cross-language text re-use, (*ii*) paraphrase text re-use, and (*iii*) mono- and cross-language re-use within and from Wikipedia.

In the case of cross-language text re-use, we propose a cross-language similarity assessment model based on statistical machine translation. The model is exhaustively compared to other available models up to date, showing to be one of the best options when looking for exact translations, regardless they are automatically or manually created.

In the case of paraphrase, the core of plagiarism, we investigate what types of paraphrase plagiarism cases are most difficult to detect. Our analysis of plagiarism detection from the perspective of paraphrasing represents something never done before. Our insights include that the most common paraphrasing strategies when plagiarising are lexical changes. These findings should be integrated in the future generation of plagiarism detectors.

Finally, in the case of Wikipedia we explore the encyclopedia as a multi-authoring framework, where texts are re-used within versions of the same article and across languages. Our analysis of multilingualism shows that Wikipedia editions in less-resourced languages tend to be better related to others. We also investigate the feasibility of extracting parallel fragments from the Wikipedia in order to (*i*) detect cases of cross-language re-use within the encyclopedia and (*ii*) enriching our cross-language similarity assessment model.

In order to empirically prove our models, we perform millions of mono- and cross-language text comparisons on the basis of different representations and measurement models. In many cases we make it on corpora generated by ourselves, which now are freely available for the interested researcher.

Resumen

La detección automática de texto reutilizado consiste en determinar si un texto ha sido producido considerando otro como fuente. Quizás el plagio, la reutilización de texto sin el crédito adecuado, sea el tipo más famoso. Los casos de plagio se han incrementado de manera dramática en los últimos años, en parte debido a la facilidad con la que es posible acceder a la información a través de medios electrónicos. Ello ha motivado que expertos en análisis de textos presten atención a este fenómeno.

Con base en las tecnologías de procesamiento de lenguaje natural y recuperación de información, los métodos de detección automática de texto reutilizado comparan miles de documentos en busca de la posible fuente de un texto presumiblemente reutilizado. En aquellos casos en los que se desea descubrir casos de reutilización entre lenguas, es posible utilizar técnicas de traducción automática. Gracias a toda esta tecnología es posible realizar miles de comparaciones exhaustivas incluso entre documentos en distintas lenguas, algo imposible de llevar a cabo de manera manual.

En esta tesis nos enfocamos principalmente en tres tipos de reutilización:
(*i*) reutilización de texto translingüe, (*ii*) reutilización de texto con paráfrasis, y (*iii*) reutilización monolingüe y translingüe dentro y desde Wikipedia.

En el caso de la reutilización de texto translingüe, proponemos un modelo para medir la similitud entre textos basado en traducción automática estadística. El modelo es comparado con algunos otros de los disponibles a la fecha de manera exhaustiva, mostrando ser una de las mejores opciones en aquellos casos en los que se buscan traducciones exactas, sin importar si éstas han sido generadas automática o manualmente.

En el caso de la paráfrasis, el núcleo del plagio, investigamos los tipos de paráfrasis que son más difíciles de detectar. Nuestro análisis de la tarea de detección de plagio desde la perspectiva de la paráfrasis representa una investigación que nunca antes se había llevado a cabo. Entre nuestros descubrimientos, cabe destacar que las estrategias de paráfrasis que más comúnmente se aplican son cambios léxicos. Dichos descubrimientos deberían ser considerados en la próxima generación de detectores de plagio.

Finalmente, exploramos a Wikipedia como un marco en el que interactúan infinidad de autores; en el que los contenidos son reutilizados en la generación de nuevas versiones de un artículo y también saltan de una lengua a otra. Nuestro análisis de plurilingüismo muestran que aquellas ediciones de Wikipedia en lenguas con menos recursos tienden a estar mejor relacionadas con otras. También investigamos qué tan factible es extraer fragmentos paralelos de Wikipedia con el objetivo de (*i*) detectar casos de reutilización translingüe en la enciclopedia y (*ii*) enriquecer nuestro modelo para medir la similitud de textos en distintas lenguas.

Con el objetivo de probar nuestros modelos empíricamente, realizamos millones de comparaciones, tanto monolingües como translingües, con base en diversas técnicas de representación y medidas de similitud. En muchos casos, nuestros experimentos son realizados considerando corpus desarrollados por nosotros mismos, los cuales están ya disponibles para cualquier investigador interesado de manera gratuita.

Resum

La detecció automàtica de text reutilitzat consisteix a determinar si un text ha estat produït considerant-ne un altre com a font. El plagi, la reutilització de text sense citar-ne l'autor, és potser el tipus de text reutilitzat més famós. Els casos de plagi han incrementat considerablement en els últims anys, en part, a causa de la facilitat amb què es pot accedir a la informació a través de mitjans electrònics. Això ha fet que experts en anàlisi de textos parin atenció a aquest fenomen.

Basant-se en tecnologies de processament del llenguatge natural i recuperació d'informació, els mètodes de detecció automàtica de text reutilitzat comparen milers de documents, a la recerca de la possible font d'un text presumiblement reutilitzat. Quan es volen trobar casos de reutilització entre llengües diferents, es poden utilitzar tècniques de traducció automàtica. Gràcies a tota aquesta tecnologia, és possible realitzar milers de comparacions exhaustives, fins i tot entre documents en llengües diferents, cosa impossible de dur a terme manualment.

En aquesta tesi ens centrem principalment en tres tipus de reutilització: (*i*) reutilització de text entre llengües diferents, (*ii*) reutilització de text amb paràfrasis, i (*iii*) reutilització monolingüe i entre llengües a dins i des de la Wikipedia.

En el cas de la reutilització de text entre llengües, proposem un model per mesurar la similitud entre textos basat en traducció automàtica estadística. El model es compara de manera exhaustiva amb altres models disponibles actualment. Aquesta comparació mostra que és una de les millors opcions per tractar aquells casos en què es busquen traduccions exactes, sense tenir importància si aquestes han estat generades automàticament o manual.

En el cas de la reutilització de text amb paràfrasis, que constitueixen el nucli del plagi, investiguem els tipus de paràfrasi que són més difícils de detectar pels sistemes. L'anàlisi de la detecció de plagi des de la perspectiva de la paràfrasi és pionera, en el sentit que mai abans s'havia dut a terme. Dels resultats del nostre treball, destaca el fet que les estratègies de paràfrasi més utilitzades són els canvis lèxics. Això caldria tenir-ho en compte en la creació de la propera generació de detectors de plagi.

Finalment, explorem la Wikipedia com un entorn on interactuen infinitat d'autors; on els continguts són reutilitzats en la generació de noves versions d'un article i també salten d'una llengua a una altra. La nostra anàlisi del plurilingüisme mostra que aquelles Wikipedies en llengües amb menys recursos tendeixen a estar més ben enllaçades amb les altres. També investiguem fins a quin punt és factible extreure fragments paral·lels de la Wikipedia amb l'objectiu de (i) detectar casos de reutilització entre llengües a l'enciclopèdia i (ii) enriquir el nostre model per tal de poder mesurar la similitud de textos en diferents llengües.

Amb l'objectiu de provar els nostres models empíricament, fem milions de comparacions, tant monolingües com entre llengües, basant-nos en diverses tècniques de representació i mesures de similitud. En molts casos, els nostres experiments es realitzen considerant corpus desenvolupats per nosaltres mateixos, els quals estan ja disponibles de manera gratuïta per a qualsevol investigador interessat.

Contents

List of Figures

List of Tables

Notation Summary

Documents, authors and text

$\mathcal{A}$	author	a	Wikipedia article
d	document	d_q	query document
D / D_q	set of documents d / d_q	q	query
s	text fragment	t	term
w	graphic word		

Weighting and similarity

$\vec{d}$	d's vectorial representation	$tf_{t,d}$	term frequency of t in d
idf_t	inverse doc. frequency of term t	$tp_{t,d}$	transition point of term t in d
$sim(a, b)$	similarity between a and b		

Evaluation

fn	false negative	fp	false positive
gran	granularity	HFM	highest false match
F-measure	harmonic mean of prec and rec	plagdet	plagiarism detection
prec	precision	rec	recall
sep	separation	tn	true negative
tp	true positive		

Languages (ISO 639-1)

de	German	el	Greek
en	English	es	Spanish
et	Estonian	eu	Basque
fr	French	hi	Hindi
hr	Croatian	lt	Lithuanian
lv	Latvian	nl	Dutch
pl	Polish	ro	Romanian
sl	Slovenian		
L	a given language		

Chapter 1

Introduction

> Their writings are thoughts stolen from us by anticipation.
>
> Alexis Piron

Nowadays technology offers the facility to copy text (and other kinds of information) easier than ever before. A mouse click, followed by a couple of keystrokes are enough to save a document from the Web and, as some people think, to get its property. This issue was highlighted already in the 1980s by Mallon (2001)[1], who mentioned that the *Save as* button opened a window to define a new name for a file and, therefore, change its identity. Very simple operations are necessary to edit the document contents in order to adapt them to a given writer interests; to re-use them. Indeed, both Clough (2003) and Comas and Sureda (2008b) stress that cyberspace has thinned the author's ownership over a given material. Whereas current technology has made information easy to reach, its re-use has become easy as well.

As described in Chapter 2, many kinds of text re-use exist, such as *co-derivation* (when a document is derived from another one). Still plagiarism represents probably the most famous kind of re-use, as texts contained in other documents are used when generating another one, but no proper reference about the source is included. The deceit that this act implies is well captured by Samuelson's (1994, p. 24) conception of plagiarism:

> *The wrong in plagiarism lies in misrepresenting that a text originated from the person claiming to be its author when that person knows very well that it was derived from another source, knows also that the reader is unlikely to know this, and hopes to benefit from the reader's ignorance.*

Plagiarism, and in general re-use, does not occur within texts only. Music, images, videos

[1]Our attention to Mallon's book was originally drawn by Clough (2003), whom references the 1989 edition, entitled *Stolen Words: Forays into the Origins and Ravages of Plagiarism*. We had access to the edition published in 2001: *Stolen Words. The Classic Book on Plagiarism*. This edition constitutes a revised version of the 1989 book and includes a new afterword. All of the references to Mallon's book in this thesis refer to the 2001 edition, but most of its contents date back to the 1980s.

and even ideas are often subject of re-use.[2] Nevertheless, this research work is focussed on text re-use and its unfair commitment: text plagiarism. Human beings are the best in detecting a case of re-use. However, the explosion in the amount of information available causes keeping track of every available resource unaffordable. As a result, methods that assist the human in detecting this kind of phenomenon are mandatory. Indeed, similar technology to that which has caused the dramatic increase in cases of plagiarism can be exploited to detect it. In this document, our efforts on the development of technology that assists the human expert in the detection of text re-use and plagiarism are discussed.

The reader could be wondering why it is worth carrying out research on plagiarism and its automatic detection. It might be thought that plagiarism is a problem limited to academia and that prevention is better (and easier) than detection. Unfortunately, this is far to be truth. Many cases of plagiarism occur in academic environments, but other circles are not extent. Text re-use (sometimes plagiarism) happens when a blog entry is re-published in a different website, when a candidate borrows speeches for her political campaign, when fiction authors base large parts of their books on previous literature, or when musicians take authors writings and other plays into their own repertory.[3] More importantly: it is not necessary to borrow an entire document, picture, video or song to (potentially) commit plagiarism. Just a fragment, for instance a text sentence, is enough to further investigate whether there is a case of re-use at hand and, if no proper credit is provided, plagiarism.

We have made some efforts in increasing the interest on the development of models for automatic text re-use detection. We have worked on the standardisation of evaluation frameworks for automatic plagiarism detection. We have also developed models for assisting in text re-use and plagiarism detection with special emphasis on cross-language cases. At the end of the day, our aim is providing with the technology necessary to uncover cases of re-use. When using this kind of technology, an important fact should not be forgotten: text re-use and plagiarism detectors must be considered as text matching systems (Jones, Reid, and Bartlett, 2008). The final responsibility belongs to the expert (professor, reviewer, forensic linguist or another expert).

Proposition 1.1 *Determining whether a text fragment has been plagiarised, or even re-used, is a decision that concerns to human judge. Automatic systems are aimed at assisting such an expert to uncover a potential case and, if possible, to take an informed decision. Claiming that a person is culpable of plagiarism is the responsibility of the expert, not of a computer program.*

The aim of automatic plagiarism detectors is, first of all detecting misconduct cases. However, in long term their aim is to discourage people from being tempted to plagiarise.[4]

[2] In many cases unfair re-use does not imply (only) plagiarism, but copyright violation as well. There is a narrow line between plagiarism and copyright, but we do not approach copyright issues in this document, a topic more related to legal aspects.

[3] A broader overview exemplifying some interesting cases is provided later, in Section 2.4.

[4] A study developed by Pupovac, Bilić-Zulle, and Petrovečki (2008) to Croatian students showed that the plagiarism rate decreases as soon as students know that their texts will be automatically analysed by means of a plagiarism detection software (cf. Section 2.5.1).

1.1 Motivation and Objectives

Information re-use at the scale we are witnessing nowadays represents a new phenomenon and challenge. Whereas in some cases information is re-used with the aim of enriching it, in many others it does not imply so good intentions. The simple act of cutting & pasting a text fragment, without any other rational processing, should be disapproved. For the specific case of plagiarism, considered by some people in academic life as a "cardinal sin" (Clough, Gaizauskas, and Piao, 2002, p. 1678), the long-term aim is not detecting the most cases of borrowing and punishing the "guilty". The aim is discouraging it. Jones *et al.* (2008) states that deterring cheating is far more effective than detecting it. Teaching good research strategies and delimiting clear academic rules have shown to work well, but still not enough. Nevertheless, while students do not receive the proper instruction about cheating, systems that assist in the detection of plagiarism and other kinds of misconduct are necessary.

Maurer, Kappe, and Zaka (2006, p. 1056) consider that the most effective approach against plagiarism must include prevention, surveillance, and response. This research work is intended to generate models that assist in detecting cases of text re-use and plagiarism (the surveillance stage). We are particularly interested in those plagiarism cases where paraphrasing and translation mechanisms are involved. Indeed, some scholars consider that paraphrasing and translation are "intimately related" (Callison-Burch, 2007, p. 1). For Milićević (2007, pp. 4, 56–57), translation is a particular case of *interlingual paraphrase*.

Automatic plagiarism detectors assist to search for potential cases of plagiarism and provide evidence to help in making the final decision. They do so by reducing the time invested in analysing and comparing texts. As a result, considering large amounts of documents and locating sources of potentially plagiarised texts becomes feasible —as soon as they are available in electronic resources— (Clough, 2003). Fortunately, the same technological principles exploited when re-using text can be applied to uncover it.

1.1.1 General Objectives

The general objectives of this research work are the following:

1. To review the problem that text plagiarism implies nowadays, establishing its particularities with respect to the broader concept of text re-use.
2. To review state of the art technology for text re-use detection, identifying strengths and drawbacks.
3. To design models for detecting cases of text re-use that employ paraphrasing and translation.

1.1.2 Specific Objectives

The specific objectives are the following:

1. To analyse the students' plagiarism attitudes and commitment in order to understand the current state of the plagiarism phenomenon in academia.
2. To analyse state of the art plagiarism detection models' drawbacks respect to cases of plagiarism with a high level of paraphrasing.
3. To design models for modified plagiarism analysis paying special attention to translated plagiarism, a problem nearly approached in the literature.
4. To analyse re-use practices in Wikipedia from both mono- and cross-language perspectives.

1.2 Research Questions

During the last twenty years, increasing efforts have been invested in the development of more and better models for the automatic detection of text re-use and plagiarism. Still many interesting gaps have been identified that we aim at filling at some extent with this dissertation.

Among the open issues in this area, Clough (2003) and Maurer *et al.* (2006) identified six problems whose solution would benefit automatic plagiarism detection:

1. Multilingual detection. As writers have access to a plethora of resources in different languages, it is plausible that a plagiarist would re-use text in a given language by translating it into another one. Clough (2003) proposed the application of cross-language information retrieval (CLIR) and multilingual copy detection (MCD) systems. Indeed, Maurer *et al.* (2006) considered that this kind of re-use was to be challenged in short time.
2. Detection of cases where extensive paraphrasing is applied. The most dishonest case could be the use of automatic synonymising tools for hiding plagiarism.[5] Maurer *et al.* (2006) identified this as a key problem in plagiarism detection.
3. Creation of text collections for plagiarism detection. As plagiarism represents a misconduct, ethical issues have difficult the generation of a —freely available— standard collection with cases of plagiarism. As a result, comparisons among different approaches to plagiarism detection are nearly impossible. Alike other research areas, benefited by initiatives such as the Text REtrieval Conference (TREC) and the Cross-Language Evaluation Forum (CLEF), which generate collections and test-beds for different tasks, plagiarism detection requires such an impulse.
4. Use of natural language processing (NLP). Text re-use often implies different levels of rewriting. NLP techniques, such as paraphrase and morphological analysis, may be worth considering when aiming at detecting plagiarism and text re-use.
5. Use of techniques from machine learning (ML). Plagiarism detection can be seen as a classification problem. Labels can be, for instance, *original* versus *non-original*.

[5]For instance, consider the *Anti-Anti-Plagiarism System*, which promises to reformulate an English input text with "[...] as many textual changes as possible, while maintaining grammar and spelling". (cf. `http://sourceforge.net/projects/aaps/`; last visited, Aug. 12 2011).

6. Detection within single texts. Professors consider that a change in the writing style of a document are enough to signal a case of plagiarism (72% of the surveyed people by Bull, Collins, Coughlin, and Sharp (2001) said so). Therefore, computational stylometry should be considered when looking for plagiarism. This kind of technique can be applied when the source of a borrowed text is not available electronically (a problem pointed out by Maurer *et al.* (2006)).

Interestingly, paraphrase re-use is still identified as an open issue. Burrows, Potthast, and Stein (2012) point that plagiarism detection is "a relevant paraphrase recognition task that, in particular, deals with passage-level text re-use." Indeed, they consider that "verbatim copying is easy to detect, whereas manually paraphrased plagiarism cases are quite difficult".

Moreover, we have identified the main difficulties when approaching automatic text re-use detection are threefold:

(a) plagiarism implies an infringement and, due to ethical aspects, no standard collection of plagiarism cases is available;
(b) the source of a re-used text may be hosted on large collections of documents; and
(c) plagiarism often implies modifications such as paraphrasing and, if possible even more interesting, translation, perhaps the most drastic obfuscation strategy.

These issues lead us to the following research questions:

1. How to build a standard collection of documents for the study and development of automatic plagiarism detection?
   (a) What are the reasons behind the lack of freely available collections of documents with actual cases of plagiarism?
   (b) Under what circumstances currently available corpora of text re-use are useful for plagiarism detection?
   (c) How to build a corpus with artificial cases of plagiarism?
   (d) How valid is a synthetic corpus for plagiarism detection?
2. What models perform best to detect cases of re-use with high level of paraphrasing?
   (a) Are simple —syntax-based— models enough for detecting this kind of borrowing?
   (b) How can paraphrases analysis techniques support text re-use and plagiarism detection?
   (c) What are the paraphrasing operations that most mislead automatic detectors?
3. How can we detect cases of text re-use across languages?
   (a) How can we build a collection of cross-language text re-use cases?
   (b) How well do (adapted) models for CLIR perform when aiming at detecting text re-use?

(c) How well do (adapted) models for machine translation perform when detecting re-use?

Our research work is focussed to address most of the aforementioned questions and open issues. We have created models for text re-use and plagiarism detection with special emphasis in cross-language cases. Moreover, in the framework of the activities of the competitions of PAN organised at CLEF and FIRE, we have created a standard evaluation framework for this area both at monolingual and cross-language level. These contributions fill important research gaps for the international community including three disciplines: natural language processing, information retrieval, and forensic linguistics.

1.3 Main contributions

The main contributions of this thesis are summarised below.

A novel model for cross-language similarity assessment, Cross-Language Alignment-based Similarity Analysis, is proposed. This model is compared to state-of-the-art models on different steps and scenarios of cross-language plagiarism detection considering different language pairs. The obtained results show that this model performs best when looking for nearly exact translations at document level and is competitive when looking for plagiarism fragments within documents. (Chapters 3, 6, and 7.)

The phenomenon of (simulated) plagiarism is studied from the point of view of paraphrases. Though a key factor, paraphrases are rarely approached in the literature on text re-use detection. Here we go beyond existing research with a pioneering analysis that opens the door to the development of better models. The most common paraphrase phenomena when re-using texts are identified and of state-of-the-art systems for automatic plagiarism detection are analysed against instances of plagiarism with different types of paraphrasing. (Chapter 8.)

Wikipedia is analysed as a mono- and cross-language text re-use environment. Re-use across revisions of an article (monolingual) and between comparable articles in different languages of Wikipedia (cross-language) is studied. The simulation and detection of cross-language re-use from Wikipedia, which is identified as a preferred source when plagiarising, is also explored. (Chapter 9.)

Evaluation frameworks are developed for the analysis and automatic detection of text re-use and plagiarism. Special emphasis has been made on cross-language text re-use from Wikipedia in the PAN initiative on Cross-Language !ndian Text Re-use. (Chapters 4, 7, and 9.)

1.4 Outline of the Thesis

This document consists of 10 chapters and 3 appendices, describing our efforts to contribute in solving the questions exposed in Section 1.2. The contents are described

following.[6] Chapters 2 and 3 intend to be an overall introduction of the topics here covered. Chapter 2 offers a overview of text-reuse with special emphasis on plagiarism. Chapter 3 gives an introduction to those information retrieval and natural language processing concepts that are used through the rest of the thesis. The experienced reader can safely skip these chapters.

Chapter 2 Plagiarism and Text Re-Use.

This chapter gives an overview of the text re-use phenomenon. Special attention is paid to signal why plagiarism is such a particular kind of re-use. A brief history of plagiarism is provided and some interesting cases are reviewed. The (still thin) bridge between computational linguistics and forensic linguistics in this task is discussed. Our cutting edge contribution comes in the form of a survey we recently held in different Mexican universities. Our aim was to assess how often plagiarism is committed across languages and students' attitudes respect to paraphrase plagiarism (factors never before analysed). An overview of the problem of text re-use and plagiarism detection implications is offered, including what the lacks for the developments of these techniques are. Finally, an overview of some commercial systems for automatic plagiarism detection currently available is provided.

Chapter 3 Text Analysis for Text Re-Use and Plagiarism Detection.

This chapter kicks off with an overview of text representation models, useful for characterising documents for analysis and comparison. It continues with a description of some of the most common models for text similarity estimation. These models are valuable when comparing a set of documents looking for re-used fragments. Some of the most successful measures to represent stylistic and complexity text features are then discussed. These measures are exploited when looking for suspicious text fragments inside of a document, without considering any reference. The publications supporting the contributions of this chapter are:

- Potthast, Barrón-Cedeño, Stein, and Rosso (2011a)
- Pinto, Civera, Barrón-Cedeño, Juan, and Rosso (2009)
- Barrón-Cedeño, Rosso, Pinto, and Juan (2008)

Chapter 4 Corpora and Evaluation Measures.

This chapter is divided in two parts. The former one describes some of the most interesting corpora for (automatic) analysis of text re-use and plagiarism available up to date. It includes corpora with manually and algorithmically created cases. In the latter part some metrics used for evaluating this task are described. Some of the measures included are well known in information retrieval and related areas. Some other have been just proposed, and specially designed for this task. Participation in the construction of three corpora —co-derivatives, CL!TR, and to a smaller extent PAN-PC— are cutting edge contributions that belong to the framework in

[6]Every chapter opens with a quote related to its contents. Quotes in the dissertation opening and Chapters 2, 3, 5, and 10 belong to different people and have been compiled by Mallon (2001, pp. 245, 2, 221, 96, 249), Chapter 1 (Piron, 1846, p. 56), Chapter 4 (Chomsky, 1957) (as seen in McEnery and Wilson (2001, p. 10)), Chapter 6 Van Gogh Museum (Amsterdam), Chapter 7 (Cerf, 2011), Chapter 8 (de Montaigne, 1802, p. 162), Chapter 9 (Wikipedia, 2011q), and References Albert Einstein.

which this research work has been carried out. The publications supporting the contributions of this chapter are:

- Stein, Potthast, Rosso, Barrón-Cedeño, Stamatatos, and Koppel (2011a)
- Potthast, Eiselt, Barrón-Cedeño, Stein, and Rosso (2011b)
- Potthast, Stein, Barrón-Cedeño, and Rosso (2010a)
- Barrón-Cedeño, Potthast, Rosso, Stein, and Eiselt (2010a)
- Barrón-Cedeño and Rosso (2010)
- Potthast, Barrón-Cedeño, Eiselt, Stein, and Rosso (2010d)
- Potthast, Stein, Eiselt, Barrón-Cedeño, and Rosso (2009)
- Barrón-Cedeño, Eiselt, and Rosso (2009a)

Chapter 5 Monolingual Detection of Text Re-Use and Plagiarism Detection.

The discussion opens with the definition of the two main approaches to text re-use detection: intrinsic and external. Literature available on the topic, from the monolingual point of view, is reviewed. The definition of a prototypical plagiarism detection architecture composes the preamble for describing some of our contributions to external —monolingual— plagiarism detection. These contributions include: (a) the evaluation of a previously proposed model based on word n-grams and (b) a model for retrieving those related documents to the suspicious one, hence reducing the load when performing the actual plagiarism detection process. Such an approach is often neglected in the plagiarism detection literature, which often assumes that either the step is not necessary or it is already solved. The publications supporting the contributions of this chapter are:

- Barrón-Cedeño (2010)
- Barrón-Cedeño, Basile, Degli Esposti, and Rosso (2010d)
- Barrón-Cedeño and Rosso (2009a)
- Barrón-Cedeño, Rosso, and Benedí (2009b)
- Barrón-Cedeño and Rosso (2009b)

Chapter 6 Cross-Language Detection of Text Re-Use and Plagiarism Detection.

This chapter represents one of the most novel research work included in this dissertation. The few approaches available for cross-language plagiarism detection are reviewed. The model we have proposed to approach this problem (CL-ASA) is then described and compared to other state-of-the-art models over common test-beds. We analyse the expressiveness of the different cross-language similarity assessment models in different sub-tasks of the cross-language plagiarism detection process. The languages considered in the experiments are both well-resourced (e.g. English, German) and under-resourced (e.g. Polish, Basque). This variety of languages is considered in order to analyse the strengths and weaknesses of the different models. We show quantitative results to appreciate the advantages of CL-ASA over other state-of-the-art models, mainly when dealing with nearly-exact translations. The publications supporting the contributions of this chapter are:

- Potthast, Barrón-Cedeño, Stein, and Rosso (2011a)
- Barrón-Cedeño, Rosso, Agirre, and Labaka (2010c)
- Barrón-Cedeño (2010)

- Sidorov, Barrón-Cedeño, and Rosso (2010)
- Pinto, Civera, Barrón-Cedeño, Juan, and Rosso (2009)
- Barrón-Cedeño, Rosso, Pinto, and Juan (2008)

Chapter 7 PAN International Competition on Plagiarism Detection.

With the aim of promoting the development of more and better systems for text re-use and, in particular, plagiarism detection, we have been running a competition during the last three years. This chapter offers an overview of such a competition. Some of the most successful approaches applied by the competitors are analysed. We also experiment with our mono- and cross-language text re-use models and discuss the obtained results. The publications supporting the contributions of this chapter are:

- Potthast, Eiselt, Barrón-Cedeño, Stein, and Rosso (2011b)
- Stein, Potthast, Rosso, Barrón-Cedeño, Stamatatos, and Koppel (2011a)
- Barrón-Cedeño (2010)
- Potthast, Stein, Barrón-Cedeño, and Rosso (2010a)
- Barrón-Cedeño, Basile, Degli Esposti, and Rosso (2010d)
- Barrón-Cedeño, Potthast, Rosso, Stein, and Eiselt (2010a)
- Sidorov, Barrón-Cedeño, and Rosso (2010)
- Barrón-Cedeño and Rosso (2010)
- Potthast, Barrón-Cedeño, Eiselt, Stein, and Rosso (2010d)
- Pinto, Civera, Barrón-Cedeño, Juan, and Rosso (2009)
- Potthast, Stein, Eiselt, Barrón-Cedeño, and Rosso (2009)

Chapter 8 Plagiarism meets Paraphrasing.

In this chapter the relationship between plagiarism and paraphrasing is analysed, and the potentials of such a relationship in automatic plagiarism detection are set out. It starts with the definition of a recently proposed paraphrases typology. The typology has been used to annotate, at paraphrase level, a sample of the manually generated cases of the corpus used at the competition on plagiarism detection in 2010. Afterwards, we analyse how poorly different models for plagiarism detection perform when facing cases of re-use with different types of paraphrasing. Our findings should provide useful insights to take into account for the development of the next generation of plagiarism detection systems. The publication supporting the contributions of this chapter are:

- Barrón-Cedeño, Vila, and Rosso (2010b)

Chapter 9 Detection of Text Re-Use in Wikipedia.

Wikipedia is perhaps the environment with most cases of text re-use publicly available. Firstly, we analyse the phenomenon of monolingual co-derivation among revisions of Wikipedia articles. Secondly, we analyse the phenomenon of cross-language text re-use among editions of Wikipedia in different languages. Related to the latter issue, we offer a preliminary discussion on a challenge on cross-language text re-use we recently organised, where the potentially re-used documents were written in Hindi and the potential source documents were written in English. The publications supporting the contributions of this chapter are:

- Barrón-Cedeño, Rosso, Lalitha Devi, Clough, and Stevenson (2011)
- Silvestre-Cerdà, García-Martínez, Barrón-Cedeño, and Rosso (2011)
- Barrón-Cedeño (2010)
- Barrón-Cedeño, Eiselt, and Rosso (2009a)

Chapter 10 Conclusions.

It offers a final summary of the contributions of this research work. It includes suggestions for further work and a brief overview of the research we are carrying out currently: the analysis and exploitation of Wikipedia as a multilingual resource for text re-use detection.

Appendix A Generation of Dictionaries for CL-ASA.

Our cross-language similarity assessment model requires a statistical bilingual dictionary. In this appendix we describe the process we have followed to generate the instances.

Appendix B Related Publications.

The scientific publications generated during this research work are listed. It includes articles in journals, conferences and workshops as well as book chapters. We include an overview of the scientific events we have collaborated in organising.

Appendix C Media Coverage.

The research work on cross-language plagiarism detection has attracted a certain attention from media. Reports on television, radio and press are listed here.

Chapter 2

Plagiarism and Text Re-Use

> If being charged with plagiarism could be compared to running a red light, the defenses have, for centuries now, been on the order not of "I didn't run it" and "It hadn't turned red" but of "What exactly do you mean by a light?" and "Define 'run.'"
>
> Thomas Mallon

When a new idea is created, it stands in all kind of previous knowledge. Re-use is inherent to innovation and, in the case of text generation, there is no exception. However, when re-use is not reported, a potential case of plagiarism exists. In this chapter, we discuss text re-use, starting with an analysis of the factors behind its commitment in Section 2.1. We focus on an interesting case of text re-use in Section 2.2: plagiarism. After understanding the phenomenon, we offer an overview of a highly related area to computational linguistics, particularly when facing automatic plagiarism detection and authorship attribution:[1] forensic linguistics in Section 2.3. The increase in the amount of cases of plagiarism in different circles is discussed in Section 2.4, including an overview of interesting cases and surveys reflecting students and professors attitudes respect to academic plagiarism and cheating, one of them conducted by ourselves. Afterwards, we discuss the tasks of text-reuse and plagiarism detection in Section 2.6. An overview of commercial systems for plagiarism detection is offered in Section 2.7.

Key contributions Testing of some standard plagiarism detection models in a set of real (publicly available) plagiarism cases (Section 2.4.1). Report on a new survey recently applied to students from Mexican universities, particularly focussed on cross-language plagiarism behaviour; a kind of plagiarism which had never been surveyed before (Section 2.5.2).

[1] Authorship attribution can be divided into two main tasks: (*a*) authorship identification implies determining the author of a text from a set of candidate authors and (*b*) authorship verification consists of determining whether one specific author wrote a text or not (Argamon and Juola, 2011).

2.1 Text Re-Use

Text re-use is defined as "the situation in which pre-existing written material is consciously used again during the creation of a new text or version" (Clough, 2010). Indeed, it is considered to be present when summarising, translating and re-writing a document, among other text manipulation processes (Clough, 2003). Co-derivation, that stands for a document being derived from another one (Bernstein and Zobel, 2004) is yet another kind of re-use. Even text simplification[2] represents a case of text re-use. Clough *et al.* (2002) point out that text re-use "stretches from verbatim, or literal word-for-word re-use, through varying degrees of transformation involving substitutions, insertions, deletions and re-orderings, to a situation where the text has been generated completely independently, but where the same events are being described by another member of the same linguistic and cultural community (and hence where one can anticipate overlap of various sorts)."

Not every kind of re-use is considered as a fault. It is extremely hard to come out with an absolutely novel idea without considering previous related work. As a result, original works (more academic rather than literary) stand on re-use of both: texts and ideas. *Press rewriting* (also known as *journalistic text re-use*) is a particular case where re-use is considered benign (Wilks, 2004): information is generated by agencies intending to have newspapers and other media re-using it. News agencies[3] cover events and generate news stories about them. Media pay fees to gather access to the contents, acquiring the right for re-using them. The subscription grants them the right to publishing notes verbatim or to re-write them as they fit to their editorial style or interests.[4]

Another environment where re-use is not considered unfair is *collaborative authoring*. On the Web, many projects where contents are generated by multiple authors exist; for instance, software manuals. One of the most interesting co-authoring frameworks is Wikipedia[5]. Wikipedia is self-defined as "a free, Web-based, collaborative, multilingual encyclopedia project" (Wikipedia, 2011p). As a result, contents in specific entries (articles) of the Wikipedia can be re-used when generating other related concepts, also crossing languages. From our point of view, the most interesting case of borrowing in Wikipedia is *cross-language text re-use*; when the contents of an entry are used for generating the corresponding article in another language. The references cited in the source text are also included in the target one in most cases, and no reference to the source Wikipedia article is necessary.[6] Using contents from a Wikipedia article when generating another one is considered fair. Wikipedia has been identified as one of the favourite

[2]For instance *Text Adaptor*, a system developed by the *Educational Testing Service* (ETS) with the aim of automatically simplifying a text for comprehension by a given target population (Burstein, 2009).

[3]Such as the British PA, the American AP, and the Spanish EFE.

[4]Indeed, corpora of journalistic text re-use have been used "simulating" cases of plagiarism. The only difference between them is that for newspapers the text is adapted to agree a house's style, while plagiarists edit the contents in order to hide their fault. While psychologically (and ethically), the purposes are clearly different, from the point of view of natural language processing (NLP) and information retrieval (IR) they seem not to be (cf. Section 4.2.1 for a review of one of the most important corpora for text re-use analysis).

[5]`http://www.wikipedia.org`

[6]This is a common phenomenon in Wikipedia. A preliminary analysis about it is included in Section 9.4.

sources for plagiarists (Martínez, 2009), extracting contents from the encyclopedia for their own documents. Technology aimed at massifying access to information is misused as a short-cut.

2.1.1 Methods for Text Re-Use Commitment

Many different ways for re-using text exist. Martin (1994) identifies six methods for plagiarising. Nevertheless, most of them are actually methods for text re-use. Therefore, we discuss following only methods for text re-use, whereas factors that become a re-use into a case of plagiarism are discussed in Section 2.2.2. Five of the methods identified by Martin (1994) are the following:

Word-for-word re-use It is the case where text fragments are copied without further modification. It is known as copy-paste (Maurer *et al.*, 2006) and verbatim copy (Clough *et al.*, 2002; Taylor, 1965) as well.

Re-use of ideas It occurs when an idea is re-used, but without any dependence in terms of words or form to the source.

Paraphrasing When the contents borrowed by an author $\mathcal{A}$ are changed or rephrased; i.e., paraphrased (cf. Chapter 8).

Maurer *et al.* (2006) consider two more kinds of text re-use (albeit they call it plagiarism):

Translated re-use implies translating some content and re-using it, even after further modification (cf. Chapter 6).

Re-use of source code implies using a piece of programming code (e.g. a function or class).

We consider that the most likely cases to be detected automatically are those created by word-for-word copy. Paraphrase and translated re-use are harder to deal with. Automatic models for detecting plagiarism of ideas seem to be still far to exist.

2.2 Plagiarism

Probably one of the best known and widely studied kind of text re-use is *plagiarism*.[7] One of the main reasons for this apparently deserved "renown" is the explosion in number of cases during the last decades. As Jones *et al.* (2008, p. 19) consider, "scholarship is built on other people's works and ideas"; therefore, students require to have instruction enough to differentiate between scholarship and cheating. Jude Carroll, in turn, considers that "being 'original' does not mean having novel ideas never before expressed by a human. It simply means doing the work for yourself" (BBC, 2011). The border between plagiarism and fair re-use is not always clear. One of the main problems about plagiarism

[7] In fact, text re-use can be considered as the hypernym of plagiarism, as plagiarism represents a specific case of the broader phenomenon.

is that students (and other writers) do not have a clear idea about what it really is and represents. The role of the "educator" (cf. page 19) is somehow failing in many cases.

Different kinds of cheating related to plagiarism have been identified by Wood (2004):

Collusion The collaboration among students without preliminary approval.[8]
Falsification A student presents another work as his own.
Replication A student submits the same work once and again (a kind of self-plagiarism)[9].

The punishment, even the level of culpability, a person deserves when falling into the temptation of plagiarism is not clear. For instance, Dr. John Olsson, from the *Forensic Linguistics Institute*, declared at BBC (2011) that "first year students should be allowed a little leeway [...] but for postgraduate students, [...] there is no excuse". Nevertheless, as Cavanillas (2008) establishes, plagiarism represents a twofold issue: (*i*) it is an illicit appropriation of the work of another author; and (*ii*) it is a fraud for its target audience.

Given this bunch of different conceptions and points of view, determining what exactly plagiarism is would be necessary. Text plagiarism occurs if a text written by another person is included in the self writing without proper credit or citation. It is, in words of Maurer *et al.* (2006), "theft of intellectual property". Other scholars consider it *borrowing* although others, such as Charles Gayley, *stealing* (Mallon, 2001, p. 75). Being a broadly discussed concept, the act of plagiarising has deserved multiple definitions. The most interesting ones from our point of view are the following.

Definition Plagiarism stands for:

(a) to steal and pass off the ideas or words of another as one's own (Merriam-Webster, 2011);
(b) the re-use of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source (IEEE, 2008);
(c) giving incorrect information about the source of a quotation (Plagiarism.org, 2011); and
(d) taking the thought or style of another writer whom one has never, never read (Bierce, 1911).

Definition (a) refers to the unacknowledged re-use of ideas, independently of the words they are expressed through. The scope of Def. (b) is broader, as it includes processes, such as algorithms, and even the results obtained from them. As expected, the kinds of plagiarism described may well be embodied in form of written text. Precisely related to the latter, Def. (c) stresses that even if a borrowed text is quoted, plagiarism may exist, mainly in those cases where the corresponding source is not properly cited. Finally, Def. (d) resembles two special aspects: plagiarism could be the result of (*i*) a harmful

[8]This kind of misconduct could be uncovered by means of authorship attribution or intrinsic plagiarism detection (a single document is analysed, without considering other texts) models (cf. Section 3.4).

[9]Self-plagiarism is probably one of the most polemic kinds of this misconduct as no clear trends exist about whether a person has the right to re-submit a work she actually wrote (cf. Collberg and Kobourov (2005) for an overview of self-plagiarism in computer science research).

process, where the writer is conscious of his ethical failure; or (*ii*) cryptomnesia (Taylor, 1965). Cryptomnesia occurs when author $\mathcal{A}$ does not remember that she has previously had access to another's idea and writes it down assuming it as original. Yet another possibility exists: a writer that lacks any knowledge about plagiarism and commits it inadvertently.

2.2.1 A History of Plagiarism

In order to better understand this interesting phenomenon, a brief analysis of its development throughout the years is necessary. Probably one of the first cases of plagiarism (that by the time was not so called) occurred in the 5th century BC. Irribarne and Retondo (1981)[10] mention that texts extracted from the Library of Alexandria had been presented for a poetry contest, arguing they had been just written. The offenders were judged as thieves. According to Wikipedia (2011l), the Latin word *plagiarius* was first denoted in the 1st century AD. It was a Roman poet named *Martial* who applied it. The reason: he was complaining that another poet had "kidnapped his verses." The word plagiarius was indeed related to kidnapping a slave or a child (Mallon, 2001, p. 6).

We are not aware about the development of the concept or the occurrence of other "famous" cases by that time. During the centuries to come, the most of the information and stories were transmitted orally, making them free for modification and self-interpretation. As mentioned by Wilks (2004), texts were rewritten and rewritten, causing the ownership of a document to be unclear. As a result, plagiarism did not demand much attention. Mallon (2001, p. 4) identifies the rise of printing as the trigger of change. Within printed material authors were able to capture their stories; to own them. Now cases of misrepresentation could be easily exposed, but probably at a high cost, as texts were now exposed to stealing.[11]

Whereas printers caused some change in authoring attitudes, it definitively did not represent an actual change in writing behaviour. For instance, according to the *Books for Lawyers* section of the Ruml (1952, p. 584), Alexander Lindey (1952) found that "Shakespeare was a notorious borrower". Lynch (2006) is more precise on the description of Lindey's work mentioning that Shakespeare's *The Tempest* is claimed to be plagued of passages originally written by Montaigne. This is an old case of translated re-use.[12] It was *Ben Jonson* whom by 1601, already in the last part of Shakespeare's life, introduced the term plagiarism into the English language, by describing a *plagiary* as a person guilty of literary theft (Lynch, 2006). The current concept of plagiarism was coined in the period between the 17th and 18th centuries. Two opposing points of view were present by the time. The first one supported, and encouraged, the re-use of previous work. This fact is reflected by *Alexander Pope*, whom proclaimed that "We have no choice but steal from the classics because 'To copy Nature is to copy them' "(Lynch, 2006). On the other side of the balance, probably one of the first formal definitions of plagiarism was written down. Samuel Johnson (1755) included in his dictionary the

[10] As seen at (Girón Castro, 2008).

[11] In the 20th century, it was not the printers, but the computers which changed the panorama one more time. Nowadays plagiarism is easier than ever; at anyone's fingertips.

[12] cf. Chapter 6 for detection models for this kind of "borrowing".

following two definitions:[13]

> PLA'GIARISM [from *plagiary*] Theft; literary adoption of the thoughts or works of another.
>
> PLAGIARY [from plagium, Lat.] A thief in literature, one who steals the thoughts or writings of another. The crime of literary theft.

Only four years later Edward Young (1759, p. 10) favoured original work from imitation by considering that...

> *Originals are, and ought to be, great Favourites, for they are great Benefactors; they extend the Republic of Letters, and add a new province to its dominion: Imitators only give us a fort [sort] of Duplicates of what we had, poffibly [possibly] much better, before.*[14]

Probably the biggest increase in the interest for plagiarism (and its relationship to copyright issues) was given by the beginning of the 18th century. The *Statute of Anne* was proclaimed in 1710 with a core idea: registered writing could not be printed by anyone with a press (Mallon, 2001, p. 39)[15]. Despite a long "prehistory" of text borrowing, the sense of plagiarism we understand nowadays was to born during this period. The Statute of Anne, together with the establishment of the concept, caused cases of plagiarism to be avoided and, in some cases, uncovered and frowned. As expected, people subject to be plagiarised were authors of plays, novels and other kinds of texts. The plagiary, in most of the cases, was another author, trying to publish similar material. After a few centuries, the panorama is dramatically different.

A big change occurred in the second half of the 20th century, yet before access to computers had spread (not saying Internet). By the 1970s a new industry emerged: paper mills.[16] Companies such as *Research Assistance*, *Research Unlimited*, or *Authors' Research Services Inc.* came out, drawing attention through magazines and newspapers ads (Mallon, 2001). The service these companies offered was a big collection of research documents on the most diverse academic topics; in the most of the cases including the corresponding references. What was sold as material for assisting the student, was indeed an industry for borrowing. By the 1970s the client had to wait days for the material to arrive by mail. Today, a few seconds after payment, the material arrives at the buyer's desktop. On-line paper mills sell essays and even theses (some of them guarantee that the text will be free of plagiarism and therefore would not be detected by any system as such), making the problem even worst (Pupovac *et al.*, 2008).[17]

Obviously, paper mills are not the only resource when plagiarising. All the technology related to computers is being exploited for easing text re-use nowadays. Dr. Samuel

[13] Originally seen at (Mallon, 2001, p. xii, 11).

[14] Quotes from Ben Jonson, Alexander Pope, and Samuel Johnson were originally seen at (Lynch, 2006). Fortunately, many of the original sources are now available in electronic format.

[15] While this statute did not establish the perpetuity of such a restriction, modern laws do. Most of them even consider *post-mortem* property.

[16] Also known as essay mill. It is a selling service of essays, reports, theses and other material (very common on the Web). Such activity is known as ghostwriting.

[17] cf. `http://www.coastal.edu/library/presentations/mills2.html` for an extensive list of on-line paper mills, in English, on the Web (last accessed: June 28th 2011).

Johnson (a different character from that of the 18th century), noted by the end of the 1980s how word processing was "incrementally more helpful" to the production of stolen papers (Mallon, 2001, p. 98). In the early 1990s, Internet was identified as being involved in the majority of cases of plagiarism (Mallon, 2001, p. 245), and it maintains its leading position today. As a result, the concept of ownership of a text is fuzzier than ever before, generating a necessity to apply detection mechanisms (Wilks, 2004).

2.2.2 Plagiarism Commitment and Prevention

As aforementioned, plagiarism is a special kind of re-use in which no proper citation to the source is provided. All of the kinds of text re-use described in Section 2.1 may become also cases of plagiarism under different circumstances. We identify the following (partially adapted from Martin (1994) and Maurer *et al.* (2006)):

Plagiarism by reference omission When any kind of re-use —exact copy, paraphrasing, or translation— is made without citing the source (probably the most common kind).

Plagiarism of authorship When $\mathcal{A}$ simply kicks out the original author of a given material and puts herself. It also occurs in a different way: $\mathcal{A}$ could ask some other person to write the document in her stead.[18]

Plagiarism of secondary sources It occurs when $\mathcal{A}$ includes proper citations but without looking them up. It is known as references plagiarism as well (Iyer and Singh, 2005).

Plagiarism of the form of a source It occurs when the plagiarist looks up the cited reference, but does not stress the dependence with respect to the secondary source.[19]

Inappropriate use of quotation marks When the limit of the borrowed parts are not properly identified.

Plagiarism by amount of borrowed text It occurs when proper citations are included, but the most of the resulting text is borrowed, without further contributions.

Word-for-word re-use becomes word-for-word plagiarism and paraphrase re-use becomes paraphrase plagiarism if references are omitted. For many scholars, omitting citation does not imply any fault if the idea belongs to the common knowledge, though. Martin (1994) considers that the kind of re-use detected more often is word-for-word. However, he points out, it is not necessarily because this is the most common kind of re-use (among students), but because it is the easiest to detect. Surprisingly, he considers that paraphrase plagiarism occurs when the borrowed text is not modified enough.

[18] This phenomenon is known as *ghostwriting*, also common in politician speechwriting. A similar case is that of the "honorary authorship", where a person who has done none of the work to produce a document, a scientific paper for instance, is listed as an author (Martin, 1994). See (Martin, 1994) for an interesting comparison between institutionalised plagiarism, that occurs very often in the form of ghostwriting, and competitive plagiarism, more common among academics.

[19] Indeed, we originally studied this classification of kinds of plagiarism in (Clough, 2003). It would have been very easy to include the classification citing (Martin, 1994) directly, without actually reading that paper. Detecting whether I looked at Martin's paper seems to be a difficult task, even for humans.

Nevertheless, re-use of ideas represents precisely that kind of borrowing, where source and target texts are independent in terms of vocabulary or style.

At a different level, some other kinds of plagiarism exist. For instance, Clough (2003) points the attention to "patchwork plagiarism", where $\mathcal{A}$ copies text fragments from different electronic sources to come out with a whole document. As Comas and Sureda (2008b) mention, a survey held at various Spanish universities showed that 61% of students acknowledged having borrowed fragments from the Web at least once. In fact, 3.3% admitted buying documents from paper mills (cf. Section 2.5). They differentiate the factors that foster plagiarism as *intra-system* and *extra-system*. The former kind refers to situations provoked by the educational system, while the latter refers to the external environment, beyond school. We identify a fuzzy division among the intra-system factors proposed by Comas and Sureda (2008b) as follows:

Teacher oriented The problem resides on teaching strategies and assignments request models. In particular a lack of commitment is reflected by the following aspects:

- Teachers requesting the same works over and over again through the years;
- Teachers failing to explain the work or justifying the relevance of the addressed topic;
- Teachers poorly reviewing students work;
- Lack of collaboration among teachers;
- Assignments commended with hard time constraints; and
- Lack of understanding of the concept of academic plagiarism.

Student oriented The problem can be found in students' attitudes to school and the learning process:

- Students have a deficient training in documentary strategies, including resources management and citation techniques;
- Students fail to design appropriate schedules for their own activities, causing workloads to get concentrated short time before deadline (also referred by Pupovac *et al.* (2008));
- Competitiveness among students cause them to consider the evaluation result more important than the learning process;
- Students lack of commitment, aiming at investing the least possible effort;
- Lack of understanding of the concept of academic plagiarism; and
- Easiness of plagiarism, mainly from electronic media, as handling of Internet resources is "virtually anonymous" (Pupovac *et al.*, 2008).

Educational system oriented The problem is in a lack of clear rules, politics and instructions from the educational institution.

- The decrease in number of tests and quizzes and its substitution by projects and essays has caused an increase in the workload;
- As nowadays education is crowded, the relationship between teachers and students has impoverished;

- Lack of clear rules on cheating policies;[20] and
- The standardisation of an evaluation system interested in assessing the final result, ignoring all the necessary process.

Evidently, these three categories are not mutually exclusive. For instance, if a student lacks of a clear notion of plagiarism, this is caused by the misinformation of the school she studies at.

The extra-system factors include the following:

- The mistake in considering that, as the contents on the Web are publicly and freely available, they can be re-used without any acknowledgement;[21] and
- An environment full of deceptive behaviour, such as frauds, corruption, piracy, or simply imitation.[22]

Interestingly, Comas and Sureda (2008b) consider the first extra-system factor as internal. Following with the discussion, they identify three roles a person —or institution— can assume when dealing with plagiarism. Such roles depend on the stage the problem of plagiarism is approached in:

The educator Her aim is instructing the student with good documentation and citation principles. Among other roles, she must encourage the student to be original, critical and authentic when solving a problem or writing a document.

The policeman This character acts when there is suspicion of plagiarism. She aims at watching whether a case of plagiarism occurs on a daily basis.

The judge This person acts when a fault has occurred. She is supposed to sentence a student who committed plagiarism. Such a sentence depends on the institution policies (if they exist), and could go from withdrawing some mark to suspending the student.[23]

[20] Comas and Sureda stress that this problem is particularly sensitive in the Spanish educational system; this frame can be easily extended to many Latin American countries, though. Maurer *et al.* (2006, pp. 1052–1056), in contrast, describe how well defined policies exist in United States universities, where specific rules are delivered to the student when she enrols an institution. They consider that plagiarism is more serious in developing countries, where lack of guidance, poor knowledge on plagiarism issues and little institutional commitment makes the problem even worse. This seems to be the case of Spain as well.

[21] About that, Jones *et al.* (2008) consider that "copying, or plagiarism, from the Internet may not be 'cheating' in the eyes of students – the material is seen as being in the public domain and without ownership". This fact is supported by the results in the survey we recently ran in Mexico (cf. Section 2.5.2). We let for a final comment from the surveyed students and many of them claimed that the contents on the Web are published to be shared, and therefore they can be borrowed.

[22] Some curious cases of deceptive imitation in the last years are starred by Chinese motor companies, out of academy. Nice examples can be seen when considering the pair of original versus copied car models: *Mini Cooper* versus *Lifan 320*, *Scion XB* versus *Great Wall Coolbear*, or *Smart Fortwo* versus *Shuanghuan Noble*.

[23] The regulations applicable to a case of plagiarism depend on the framework and situation it occurs in. Cf. Cavanillas (2008, p. 8) for an overview on civil, administrative and legal responses to plagiarism. Cf. Maurer *et al.* (2006, pp. 1053–1054) for a review on actions against academic misconduct, in particular plagiarism, in different universities around the globe.

The educator is probably the most convenient and influential of the roles, as it takes place before any fault is committed; she can avoid plagiarism beforehand. Nevertheless, the policeman and the judge are still necessary, until the educator succeeds completely.

2.3 Computational Linguistics Meets Forensic Linguistics

Automatic plagiarism detection aims at supporting the work carried out by people interested in uncovering cases of unfair re-use; for instance: teachers, project reviewers, paper reviewers or *forensic linguists*. In this section we pay special attention to the latter expert. One of the reasons is that we have found interesting similarities between the two worlds: computational linguistics and forensic linguistics. Some researchers from both fields show mutual interest in looking at the other side of the fence.

2.3.1 Forensic Linguistics

According to Jackson and Jackson (2008, p. 1), in a broad sense *forensic science* is "any science that is used in the service of the justice system". Its aim is providing evidence to legal investigations by applying scientific techniques (Tilstone, Savage, and Clarck, 2006, p. 1). It can be divided into many specialities, such as *forensic anthropology*, *forensic chemistry* or *forensic dactyloscopy*.

We are interested in two sub-areas: *forensic document examination* and *forensic linguistics* (FL). In the former, a questioned document is analysed either with respect to a set of other documents or with respect to different components of the document itself.[24] Such analysis is carried out on the basis of diverse techniques, such as handwriting analysis. Nevertheless, with the advent of electronic media, more and more documents are generated by means of computers and other devices, leaving handwriting analysis out of the play in many cases. FL "deals with issues in the legal system that require linguistic expertise" (Wikipedia, 2011c). According to Turell and Coulthard (2011), it is divided in three main branches: (*i*) *language of the law*: analysing and making understand the language used in laws; (*ii*) *language of the court*: analysing the language used by judges, witnesses, and police speech, particularly during judgements; and (*iii*) *language as evidence*: grouping the techniques through which written and speech language are used as a variable in the analysis of a legal process.

We focus our attention on (*iii*), which includes sub-areas such as discourse analysis, forensic phonetics and forensic dialectology (indeed, forensic document examination could be safely included among these sub-areas). Following, we pay special attention to the problems of authorship identification and plagiarism detection from the point of view of FL. Figure 2.1 represents the typical distribution of forensic material considered in

[24] This is related to the two main approaches to automatic plagiarism detection: external and intrinsic. In the former, stylometric features are considered to identify fragments of a document that could be plagiarised. In the latter case a text is compared to a set of documents looking for re-used fragments (cf. Section 5).

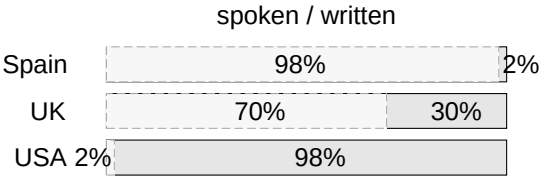


Figure 2.1: Distribution of spoken (light) and written (dark) forensic material in Spain, UK, and USA (as estimated by Fitzgerald (2010)).

three different countries. We are interested in the analysis of written text; be expressed in books, notes, letters, or, more recently, emails, SMSes, and blogs.

When analysing a text, the expert's aim is to answer questions of the kind "*Who wrote this text?*" (e.g., a suicide note, a menace or a will). Therefore, the aim of the linguist is to determine the most likely author of a string of words. In both cases, the problem can be approached as that of authorship attribution and plagiarism detection: determining who is the writer that produced a text (or text fragment). Either for authorship attribution or plagiarism detection, a key factor in the forensic linguist's labour is the concept of *idiosyncratic idiolectal style* (Turell, 2011). According to Turell and Coulthard (2011), language is used in a "distinctive way by an individual". The interesting fact is that *style*, the result of a person's background, seems to be quite stable throughout time (but not so through genre). As a result, a good part of the forensic process is based on the concept of *uniqueness*, that refers to the fact that every person is linguistically unique (Coulthard and Alison, 2007): no two people exist that express their ideas in the exact same way.

In authorship attribution, a *dubious document*[25] (that for which authorship is uncertain) is analysed by the linguist aiming at determining its most likely author among a closed set of candidates. On the basis of the uniqueness concept, having a sentence or a particular sequence of words or characters in common with a single person's writings, is reason enough to start building a hypothesis about the authorship of a dubious text. As a result, idioms and slang are highly relevant for this kind of study.

In plagiarism detection, a dubious document is presented which is claimed to be generated on the basis of another one. Finding two documents that share a common sentence (or sequence of words) is extremely unlikely if the documents are actually independent. Finding two, long enough, sequences of words in the dubious document and the potential source is, once again, reason enough to trigger suspicion.

Therefore for both, authorship attribution and plagiarism detection, cases of *hapax-legomena* and *hapax-dislegomena*[26] are of high interest. If a sequence of words appears only once or twice within a set of documents, it is relevant; either it could be an author's style marker (if it is legomena), one has borrowed from the other, or both consider a common external document as one of their sources (in case of dislegomena). As Coulthard points out: "the longer a phrase, the less likely you are going to find anybody use it". This is in agreement with the corpus linguistics point of view of McEnery and Wilson (2001, p. 8), who mentioned that "unless it is a very formulaic sentence (such as those

[25]We use here the term *dubious* because it is more common in FL literature. It is equivalent to *suspicious*.

[26]Hapax legomena (dislegomena) are phenomena, for instance, words, that appear only once (twice) in a text or collection of texts.

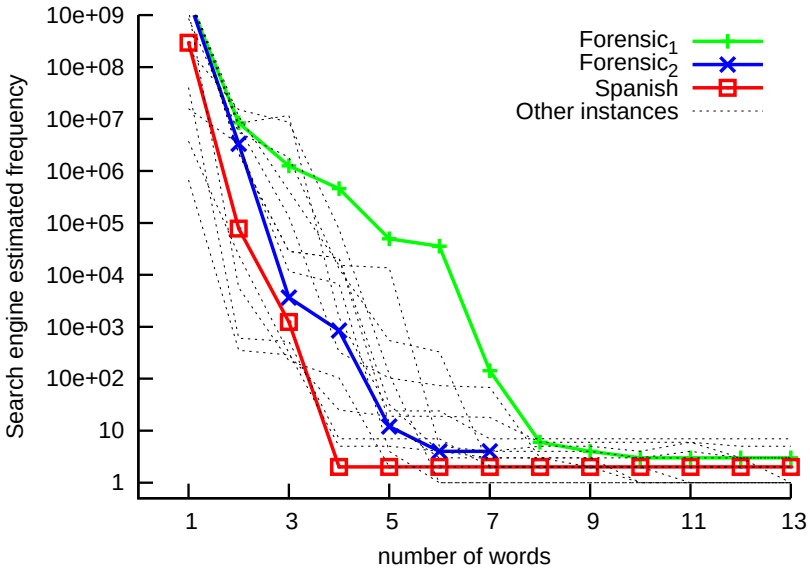


Figure 2.2: Instances of exact word sequences in a search engine considering increasing lengths. Fifteen phrases included; thirteen are the opening words in publications we have authored. Three curves are highlighted: "Spanish" corresponds to a publication written in Spanish and "Forensic$_{[1,2]}$" to sequences used in actual court cases (Coulthard, 2004). As expected, the number of instances (hits) decreases as longer sequences are searched against the search engine (Experiment carried out by querying quoted phrases against the *Google* search engine on Oct. 8th, 2011; process automatised with pygoogle: `http://code.google.com/p/pygoogle/`.)

appearing as part of a legal disclaimer at the beginning of a book), it is deeply unlikely that you will find it repeated in its exact form in any book, in any library, anywhere". Clough and Gaizauskas (2009) are more precise and consider that a match of ten consecutive words between two documents "highlights almost certain re-use". Regarding this issue, we performed an informal experiment: we queried a Web search engine with increasing length quoted chains of words[27], in order to look for exact matches. The results are displayed in Fig. 2.2. All of the curves show precisely the expected behaviour: the frequency of occurrence (i.e., the estimated number of websites containing the exact sequences), decreases as the length of the phrase increases. The highlighted curves correspond to the following three phrases:

Spanish *El plagio, el reuso no autorizado y sin referencia de texto, es un fenómeno* [...][28]
Forensic$_1$ I asked her if I could carry her bags and she said "yes"
Forensic$_2$ I picked something up like an ornament

The phrase labelled as *Spanish* is the one with the fastest decrease rate, until arriving to one single website containing its leading four words. Two facts may be the cause for such a behaviour: (*i*) less contents exist on the Web in Spanish compared to English (as a result *El plagio, el reuso* is already a hapax legomena); and (*ii*) this phrase opens a chapter in a book on Forensic Linguistics (Barrón-Cedeño *et al.*, 2010b), not

[27] In the following chapters, these sequences of words will be named word n-grams.
[28] Plagiarism, the unauthorised and without reference re-use of text, is a phenomenon [...]

	plagiarism studies/ literary authorship	**criminal authorship**
length	long texts	short texts
spontaneity	non-spontaneous	incidental and spontaneous
target	big audience	limited audience

Table 2.1: Linguistic material involved in literary and criminal studies. Characteristics of the typical analysed piece of text. Adapted from Turell and Coulthard (2011).

available on-line (we offer a draft version of this specific chapter on our website only). Interestingly, there is no other document in the whole Web —indexed by Google— with the same phrase. Forensic$_1$ and Forensic$_2$ are two sentences used in a real forensic case in order to prove that an interviewed record had been indeed invented, created on the basis of a previous statement (Coulthard, 2004). As in the current experiment, both sentences were queried to Google, returning less and less hits as the number of words augmented. None of the two plots converges to 1 as the previous one. The reason is that various websites include contents discussing the case and how the uniqueness concept was analysed precisely with the assistance of search engines.

From a linguistics point of view, words can be divided into two main groups: (*a*) *lexical words* refer to nouns, verbs and adjectives; whereas (*b*) *grammatical words* refer to prepositions, articles, auxiliaries and others. It is estimated that 40% of the words in a text are lexical. Coulthard (2010) considers that documents on the same topic could share around 25% of lexical words. However, if two documents contain circa 60% of lexical words in common, they can be considered related.

2.3.2 (Dis)similarities between Computational and Forensic Linguistics

Some important differences exist with respect to approaching plagiarism detection and authorship attribution from the point of view of computational and forensic linguistics. In the former one the main interest is automatising the most of the process in order to provide some evidence to the expert to help her to take a final decision. An important issue for this point of view is scale: the amount of documents that can be taken into account when analysing a suspicious document. This makes sense when considering that the potential source of a plagiarised text fragment could exist on the Web, where millions of documents are allocated. For the latter, scale in most cases is not an issue. The FL expert faces a specific (short set of) document(s) that must be manually analysed to provide the best possible evidence.

Different peculiarities characterise the typical documents to be handled in literary (and academic) authorship and forensic authorship studies. Turell stresses the differences depicted in Table 2.1. It is worth noting that the nature of forensic texts makes them be more difficult to analyse than literary and plagiarism cases. The reasons are simple: whereas forensic documents are short, spontaneous and written thinking in a specific target person, the others are, in general, much longer, better planned and created for a broader target audience.

In general, when studying the problem of plagiarism detection from a computational linguistics point of view, not enough attention is paid to actual problems of plagiarism

uncovering or authorship attribution. This problem has to do with an expert in language analysis (also written) within the context of courts: the forensic linguist. The potential co-operation between the computational and the forensic linguist is necessary, but how to strengthen it remains unclear.

Although the two worlds seem to be still far away from each other, efforts recently started to bring them closer. At the *1st (In)formative Conference on Forensic Linguistics* (ICFL) (Garayzábal Heinze, Jiménez Bernal, and Reigosa Riveiros, 2010), a linguists forum where a few computational linguists took part, a discussion prevailed on whether technology could be used in the forensic linguists' activities, in particular when analysing text, and how.[29] Manuel de Juan Espinosa, director of the *Master in Forensic Sciences* at the *Autonomous University of Madrid* emphasised the relevance of technology in problems of (automatic) detection of personality or mood, as well as when approaching cases of cyber-terrorism and forensic informatics. In 2011, Turell and Coulthard identified a set of problems where forensic and computational linguists might converge. They identified the following for authorship:

1. Generating base rate population statistics;
2. Determining Bayesian likelihood ratios for written texts;
3. Identifying the first language of non-native writers;
4. Identifying impersonation; and
5. Automatically analysing SMSes.

The issues they identified for plagiarism detection (and therefore related to the topic of this thesis) are:

1. Determining the plagiarism directionality between contemporary texts;
2. Detecting plagiarism of meaning;
3. Detecting cross-language plagiarism; and
4. Detecting paraphrasing.

Grozea and Popescu (2010b) already started working on issue 1. More works have recently focussed on issue 3, e.g. (Corezola Pereira, Moreira, and Galante, 2010a; Potthast, Stein, and Anderka, 2008a). Our work on cross-language plagiarism detection is discussed in Chapter 6. Issue 4 remains nearly approached, but seminal works start to appear (Burrows *et al.*, 2012). Our efforts on analysing paraphrase plagiarism are described in Chapter 8.

2.4 Plagiarism: An Explosion of Cases

As aforementioned, the current concept of plagiarism has lasted for nearly 300 years. Nevertheless, the amount of cases occurring nowadays is unprecedented, mainly in academia.

[29] Obviously this is not the case of forensic phoneticians that use electronic devices in order to analyse and compare waves (just to give an example).

This explosion is sometimes attributed to the availability of resources on the Web. If the situation could be even more aggravated, some scholars consider that "[...] plagiarism is something people may do for a variety of reasons, but almost always something they do more than once" (Mallon, 2001, p. xii).[30]

2.4.1 Overview of Recent Cases of Plagiarism

We somehow focus our analysis on academic plagiarism, but plagiarism goes beyond academia. An example is that of government agencies and organisations commissioning studies and other material, paying for it (whether this actions can be considered as plagiarism —or ghostwriting— remains an open issue). Here we review some interesting cases of plagiarism commitment.

2.4.1.1 Cases of Plagiarism in Academia

Plagiarism in academia seems to be the most frequent. Whereas in most of the cases its repercussion stops at the classroom or institution they occur in, a few cases have claimed for external attention. In general, sounded academic cases imply facts happened during the academic life of current renowned people.

One of these cases implied *Karl-Theodor zu Guttenberg*, former German Defence Minister. The University of Bayreuth discovered that zu Guttenberg's 2006 doctorate dissertation had "whole lifted sections without attribution" (BBC, 2011). Being a political figure in Germany, the case attracted spotlights. In parallel to the "official" investigation by the University of Bayreuth, different analyses were carried out by Gipp, Meuschke, and Beel (2011) and the *Guttenplag Wiki* project[31], finding that most of the dissertation contents had been borrowed from several sources (Wikipedia, 2011h). As a result of the scandal, zu Guttenberg resigned his charge and his doctorate title.

Another plagiarism case implies *Martin Luther King, Jr.* According to Mallon (2001, p. 240), some of his writings contained borrowed material without citation, including his 1955 dissertation. Unlike the above case, the lifting was discovered after King's dead. The *King Papers Project*[32] analyses discovered that Luther King's *Boston University* doctoral dissertation included borrowed sections from *Jack Boozer*'s, presented in 1952 (Wikipedia, 2011j). Many other writings were found to be improperly re-used from different sources. About this, *Ralph E. Luker*, wrote: "the plagiarism in his dissertation seemed to be, by then, the product of his long-established practice" (Wikipedia, 2011j). In this case the doctoral title was not revoked, but a letter was added to the Boston University official copy stating the findings (The New York Times, 1991).

[30] Consider, for example, the cases of Gerónimo Vargas and Martin Luther King Jr., at Sections 2.4.1.1 and 2.4.1.3, respectively.

[31] `http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki/English`

[32] The *King Papers Project*, at *Stanford University*, was created aiming at publishing "a definitive fourteen-volume edition of King's most significant correspondence, sermons, speeches, published writings, and unpublished manuscripts" (The Martin Luther King, Jr. Research and Education Center, 2011).

Students are not the only plagiarists in academia. An interesting case regarding a professor and his former PhD student spawned in the Spanish press in January 2009. *Francisco José Alonso Espinosa*, professor of Mercantile Law at the *University of Murcia* published a book in 2006, whereas his student, María Isabel G. G. (as the name appears in the news note), had published her dissertation in 2001. According to court's sentence, the similarities between the published books were many, and the plagiary "used [his former student's] material, without citation, even at the bibliography, using conclusions and the structure of the other book" (Cabanes, 2009, Levante newspaper). Alonso had to pay a compensation and was forced to acknowledge the sentence in national media.

2.4.1.2 Cases of Plagiarism in Research

With research plagiarism we refer to that occurring within scientific publications. In particular, we describe a case concerning two research groups: one Chinese and one Spaniard. In this case we include the "reaction of the crowds", i.e., how people reacted and commented the event on the media websites.

In January 11, 2011, the *Journal of Chemical and Engineering Data* decided to retire two papers authored by a Galician research group. The reason was that the papers had been found to be "duplicated publications". In one of the papers the abstract, together with the first paragraph of the introduction, had been copied word-for-word from the original papers (Ingendaay, 2011). Juan Carlos Mejuto, full professor[33] at the Department of Physics and Chemistry and former dean of the Faculty of Natural Sciences, at the University of Vigo (Ourense), was one of the authors, together with three PhD students and two more academics. After the retirement, Mejuto declared to the media: "I recognise to be a bungler, but I am not a cheater; this has been an error, but not plagiarism" (Rivera, 2011).

The original papers had been published in 2007 and 2009 by a research group from the University of Hunan, China (one by Xinliang Yu, Bing Yi and Wang Xueye (2007) and another by Liu and Wanqiang Chenzhong Cao (2009)). Mejuto said that they used the papers in order to fill the author's lacks of ability to properly write in English; that a draft, containing the exact text formerly borrowed from the Chinese group, had been submitted instead of the "rephrased" —camera-ready— version. Interestingly, Mejuto claims that he "[...] did not have any intention to borrow ideas, processes, results or words from other people without the appropriate credit". This sentence is extremely similar to the IEEE definition of plagiarism (cf. Section 2.2). The offenders were banned for two years from the journal.[34] Surprisingly, all of the co-authors remain in their positions, without any penalty. For the German *Frankfurter Allgemeine Feuilleton* newspaper this situation is amazing, compared to the case of zu Guttenberg, which cost him his job and his title (Ingendaay, 2011).

The comments written after the note published by Rivera (2011) are indeed interesting. Some of them draw a claimed common story where the professor exploits the

[33] *Catedrático* in the Spanish terminology.

[34] Both papers are still available at the journal website (`http://pubs.acs.org/toc/jceaax/55/11`, last visited October, 2011) but they include the banner "This paper was withdrawn on January 11, 2011 [...]".

students to make all the work and at the end simply signs the paper, without paying attention enough to the research or the writing stage. Another comment considers that this would not have happened if the authors had written the paper in Galician, as foreign researchers would not have realised the fault. A more serious comment says that these papers would have never been accepted. If the program committee and the editors would have made their work properly, the papers had to be rejected from the beginning. The fact is that, with the plethora of texts available online, it is hard to detect a case of plagiarism by simply reading a document.[35]

These cases had an end, which is not always possible. In Mexico, the scholar *Myrna Soto* accused the PhD student *Paula Mues* for plagiarising her book "El Arte Maestra: un tratado de pintura novohispano"[36] (La Jornada, 2007). The interesting fact about this case is that, apparently, both implied writers had been working for long time on the same topic. Paula Mues has denied the plagiarism act, claiming that both researchers were working independently and arrived at similar (but not completely identical) conclusions.

We are aware of more cases of plagiarism in scientific publications, some of them caught in time to be rejected before publication. Ironically, these cases include papers about automatic plagiarism detection which plagiarise other papers; some of them could include self-plagiarism (we do not describe these cases further for ethical reasons).

2.4.1.3 Cases of Plagiarism in Journalism

Though journalism is one of few environments where re-use is allowed (cf. Section 2.1), still cases of unfair re-use exist.

In the mid-1990s *Ruth Shalit*, a political journalist, was accused of stealing passages for her notes. After being uncovered, Shalit claimed that "the material [...] on one side of her screen had somehow jumped, citation-less, into the text she was creating" (Mallon, 2001, p. 240). After being dismissed from her work at *The New Republic* for cases of plagiarism, she started working on advertising (Wikipedia, 2011n). Apparently, "once again" technology is guilty for the cases of plagiarism.

Another case mentioned by the BBC (2011) is that of *Maureen Dowd*, columnist of the New York Times and Pulitzer winner. She admitted using a paragraph virtually word-for-word from blogger Josh Marshall without attribution. Nevertheless, the borrowing was claimed an unintentional mistake as she was supposed to be "suggested" by a friend of hers. After realising the fault, Dowd apologised, the note was corrected, and Marshall was credited. The offence was sanitised.

Beyond cases of plagiarism, text re-use is common in this environment. Due to the business settings, not plagiarism, but text re-use is on the rise.[37]

[35] In recent years some journals include an automatic plagiarism detector in their review process. See for instance the agreement between iParadigms (the enterprise behind Turnitin) and IEEE: `http://www.ieee.org/publications_standards/publications/rights/crosscheckmain.html` (consulted: Nov. 26, 2011). However, most of these tools manage to detect only verbatim copies.

[36] The Master Art: Treatise of Novohispanic Painting.

[37] Once again, the border between fair and abusive re-use is thin. *Media Standards Trust* aims at "detecting press releases without much added" through their project *Churnalism*

2.4.1.4 Cases of Plagiarism in Literature

As already discussed, years ago it was expected that most of the cases of plagiarism had to do with literature. Rather than making an entire review of the endless list of literary plagiarism occurrences, we include here a few recent cases.

A sound case is related to the *Harry Potter* book series. In June 16th, 2009, the British *Daily Mail* published that *J.K. Rowling* had been sued for £500*m* (Kisiel, 2009). The reason: *Harry Potter and the Goblet of Fire* was claimed to be plagiarised from *Willy the Wizard*, by *A. Jacobs*. As reported by The Mail, "Adrian Jacobs [...] allegedly sent the manuscript to C. Little, the literary agent at Bloomsbury Publishing who went on to represent Miss Rowling, but it was rejected". Apparently both main characters, young magicians, aim at rescuing people locked in a bathroom by half-human creatures. Common settings, such as magic trains and prisons have been found. The difficulty in proving, or discarding, plagiarism in this case is that no common string of text exists between the books. If plagiarism exists, it was plagiarism of ideas. According to the Wikipedia (2011i) article about legal disputes over the Harry Potter series, the judge considered in 2011 that "[...] there simply was not enough similarity between the two books to make a case for plagiarism".

A similar case is that of the late Spaniard writer *Camilo José Cela*, Literature Nobel Prize. In October 2010, a magistrate opened trial against the publisher *Grupo Planeta*, for alleged plagiarism by Cela (La Vanguardia, 2010). The reason was simple: there were more than expected coincidences between his *Premio Planeta de Novela*[38] novel *La Cruz de San Andrés* and *María del Carmen Formoso Lapido*'s *Carmen, Carmela, Carmiña (Fluorescencia)*. The case starts when Formoso, a Galician writer, submitted her novel to the Planeta contest in 1994. Apparently, though, Cela was offered with the award beforehand and all he needed was a novel to submit; making the case even more intriguing.

Carmen Formoso relates that when she started reading Cela's book she was shocked by discovering her novel story was narrated in this book, containing strong references to her own personal life. The investigations suggest that Cela was out of ideas and Planeta itself provided him with Formoso's material, as a source for inspiration. After submitting his work right on the deadline, Cela won. Similar characters (with different names), situations and places exist in both books. One more time, no relationship between the text in the two books actually exist. According to *El País* (2009), signs had been found that suggested the plagiarism, namely the dates when the different authors submitted their works (Formoso in May 2nd and Cela in June 30th), and the report by the expert Luis Izquierdo, (professor of Spanish Literature, Universidad de Barcelona), concluding that Cela's novel was "an assumption of transformation, at least partial, of the original book" (Ríos, 2009, El País). Ríos (2010) reported that by September 2010 the judge considered that Formoso's piece had been transformed by Cela into "a play systematically different, with the author's [Cela] own stamp". Plagiarism was sentenced.

The former two cases have to do with plagiarism of ideas, where the texts are inde-

(`http://churnalism.com`).

[38] Prize Planeta.

pendent of each other. Nevertheless, cases exist of near-duplicates[39] as well. BBC (2011) mentions the case of *Kaavya Viswanathan*. Her novel *How Opal Mehta Got Kissed, Got Wild, and Got a Life* was reported to be plagiarised from *Megan McCafferty's Sloppy Firsts* and *Second Helpings* (Wikipedia, 2011f; Zhou, 2006). Nicely, Wikipedia includes some of the near-duplicate passages. Five of them are reproduced in Table 2.2, together with their estimated similarity. By analysing the numbers, the derivation becomes evident. For instance, consider the common chunks at the first fragments: "*[...]s my age and live[...]*", "*For the first [...] years of my life[...]*", and many more. The table includes a figure of the estimated similarities between the fragments, on the basis of the well-known cosine similarity measure, ranged in $[0, \ldots, 1]$ (0 means null similarity, 1 means exact match, cf. Section 3.3.1.2). The documents were represented as a vector of character 3-grams and word 1-grams.[40] For the former case, the entire vocabulary was considered, only discarding punctuation marks. For the latter, only lexical words were considered and grammatical words discarded (stopwords, in computational terms). By considering Coulthard's assertion ("if two documents contain circa 60% of lexical words in common, they can be considered related", cf. Section 2.3.1), the evidence in this case suggests that a case of plagiarism exists.

Cases of monolingual plagiarism may occur between translations as well. In 1993, *Manuel Vázquez Montalbán*, a late Catalan writer and philosopher, was found guilty of plagiarising his translation of *Julius Caesar*, from the one generated by *Ángel Luis Pujante* (Gibbons and Turell, 2008). Instead of performing his own translation, Vázquez exploited Pujante's when doing the commended work. Both texts were compared to previous translations and the offended one was noted much more original, than the plagiarised one. The uniqueness concept (cf. Section 2.3) was one of the variables considered to uncover the fault. Only 20% of the vocabulary in the Pujante and Vazquez's translations was unique with respect to each other, whereas it was higher for other, independent, translations (ranging between 27 and 45%) (Gibbons and Turell, 2008, p. 294). Some investigators estimated that around 40% of the piece was plagiarised (El País, 1990). An interesting fact about this case is that the Court considered that "[...] in such translations, such as a five-act dramatic work like Julius Caesar, the space for originality is much higher than in the case of short translations" (Gibbons and Turell, 2008, p. 292).

2.4.1.5 Cases of Plagiarism in Politics

Though cases of politicians' plagiarism when attending school have been discussed already (cf. Section 2.4.1.1), we now regard at cases occurring during their political life.

In 1987, Maureen Dowd (1987)[41] published a note claiming that *Joe Biden* "[...] lifted Mr. [Neil] Kinnok's closing speech with phrases, gestures and lyrical Welsh syntax intact for his own closing speech at a debate at the Iowa State Fair [...]". Biden had had access to a commercial, recorded in a tape, from the leader of the British Labour party

[39] Two documents are considered near-duplicates if their contents are almost identical; "if they share a very large part of their vocabulary" (Potthast and Stein, 2008).

[40] For instance, the character 3-grams of "example" are *exa*, *xam*, *amp*, *mpl*, and *ple*. Considering word 1-grams is equivalent to handle the text's vocabulary (cf. Section 3.1.3).

[41] This is the same Maureen Dowd implied in another case of plagiarism (cf. Section 2.4.1.3).

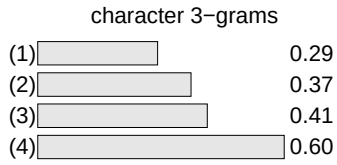
Table 2.2: Sample fragments of plagiarism in literature with estimated similarities. The similarities are shown at the bottom, for character 3-grams and word 1-grams. Text fragments borrowed from Wikipedia (2011f).

McCafferty's fragments	Viswanathan's fragments
(1) Bridget is my age and lives across the street. For the first twelve years of my life, these qualifications were all I needed in a best friend. But that was before Bridget's braces came off and her boyfriend Burke got on, before Hope and I met in our seventh grade Honors classes.	Priscilla was my age and lived two blocks away. For the first fifteen years of my life, those were the only qualifications I needed in a best friend. We had bonded over our mutual fascination with the abacus in a playgroup for gifted kids. But that was before freshman year, when Priscilla's glasses came off, and the first in a long string of boyfriends got on.
(2) Sabrina was the brainy Angel. Yet another example of how every girl had to be one or the other: Pretty or smart. Guess which one I got. You'll see where it's gotten me.	Moneypenny was the brainy female character. Yet another example of how every girl had to be one or the other: smart or pretty. I had long resigned myself to category one, and as long as it got me to Harvard, I was happy. Except, it hadn't gotten me to Harvard. Clearly, it was time to switch to category two.
(3) . . . but in a truly sadomasochistic dieting gesture, they chose to buy their Diet Cokes at Cinnabon.	In a truly masochistic gesture, they had decided to buy Diet Cokes from Mrs. Fields . . .
(4) He's got dusty reddish dreads that a girl could never run her hands through. His eyes are always half-shut. His lips are usually curled in a semi-smile, like he's in on a big joke that's being played on you but you don't know it yet.	He had too-long shaggy brown hair that fell into his eyes, which were always half shut. His mouth was always curled into a half smile, like he knew about some big joke that was about to be played on you.
(5) Tanning was the closest that Sara came to having a hobby, other than gossiping, that is. Even the webbing between her fingers was the color of coffee without cream. Even for someone with her Italian heritage and dark coloring, it was unnatural and alienlike.	It was obvious that next ro to casual hookups, tanning was her extracurricular activity of choice. Every visible inch of skin matched the color and texture of her Louis Vuitton backpack. Even combined with her dark hair and Italian heritage, she looked deep-fried.

by the time. As Dowd (1987) wrote, Biden had concurrently used "themes, phrases and concepts from Kinnok", giving him credit, though not in Iowa and some further events. Biden declared that "speeches are not copyrighted". Indeed, according to Wikipedia (2011g), Biden had been accused for plagiarism already during his first year at *Syracuse University*. By the time his reply was that he "did not know the proper rules of citation". In a similar fashion than the example of Viswanathan, Table 2.3 contains fragments of Kinnok and Biden's speeches. Evidently, as the campaign team accepted, this was a clear case of re-use and more concretely, as no citation had been provided, of plagiarism (even if no copyright is being violated).

Table 2.3: Sample fragments of plagiarism in politics with estimated similarities. Fragments copied verbatim are in bold, whereas re-written chunks are in italics. The estimated similarity between the texts is shown at the bottom for character 3-grams. Text fragments borrowed from Mallon (2001, p. 127).

	Kinnock's commercial	Biden's speech
(1)	**Why** *am I the first Kinnock in a thousand generations to be able to get* to **university**?...	**Why** i*s it that Joe Biden is the first in his family ever to go* **to** a **university?**...
(2)	*Was* **it because our** *predecessors* **were** *thick?*...	*Is* **it because our** *fathers and mothers* **were** *not bright?*...
(3)	*Was* **it because they** were weak, *those people who could work eight hours underground and then come up* **and play football**, weak?	*Is* **it because they** didn't work hard, *my ancestors who worked in the coal mines of Northeast Pennsylvania and would come up after 12 hours* **and play football** for four hours?...
(4)	*It was* **because** *there was no* **platform upon which** *they could* **stand**.	*It's* **because** *they didn't have a* **platform upon which** *to* **stand**.



An even more interesting case has *Gerónimo Vargas Aignasse*, an Argentinian Deputy, as the main character. According to the English Wikipedia (2011d), Vargas is a frequent borrower from Wikipedia itself when preparing his writings. The most ironic case was on a law proposal about plagiarism he presented. The same source mentions that 331 words were copied from the Spanish Wikipedia article on plagiarism. When interviewed by the *Clarín* newspaper about this fault, Vargas declared "*No tengo la obligación de citar las fuentes*"[42] (Arce, 2010).

Table 2.4 contains the nine most similar sentences between the two documents, together with their estimated similarity on the basis of the cosine measure for character 3-grams. Seven sentences (many of which are really long) were copied verbatim from the Wikipedia article.[43] Curiously, this practice deserved the Deputy his own article in both English and Spanish Wikipedia (Wikipedia, 2011d,e).

2.4.1.6 Cases of Plagiarism in Show Business

Another circle where plagiarism (and copyright) cases occur very often is show business. Songs, lyrics and stories are re-used once and again without citing the corresponding source. The line between inspiration, adaptation, and plagiarism is particularly thin in this case. With direct contact to the crowds and the large amounts of money implied, the cases are always more striking.

[42] "I have no obligation to cite sources"

[43] There was a possibility that the Wikipedia article could have been generated from Vargas's proposal, but the contents in discussion had been added to Wikipedia long time before.

Table 2.4: Sample fragments of plagiarism in politics from Wikipedia. Only the nine most similar sentences are included. The similarity (left column) is calculated for character 3-grams. For those cases of verbatim copy only one version is reproduced (as they are exactly the same). Text fragments borrowed from Wikipedia (2011m) and Vargas Aignasse (2010).

sim	Wikipedia article on *plagio* (plagiarism)	Vargas' proposal
1.00	En un sentido más amplio, generalmente se denomina plagio a los libros que tienen tramas o historias muy similares, a películas con semejanzas extremas en la forma de expresión de las ideas, a un invento muy similar a uno patentado, a una obra de arte similar o con alguna pieza del original, marcas; incluyendo logotipos, colores, formas, frases, entre otros distintivos de algún producto, o simplemente a ideas.	
1.00	La denominada propiedad intelectual es una colección de marcos jurídicos diferentes que protegen los intereses de autores e inventores en relación a obras creativas, ya sean estas, expresiones de ideas como en el caso del derecho de autor o aplicaciones prácticas e industriales de ideas como en el caso de las patentes.	
1.00	En el caso de documentos escritos, por ejemplo, se comete plagio al no citar la fuente original de la información incluyendo la idea, párrafo o frase dentro del documento sin comillas o sin indicar explícitamente su origen.	
1.00	Según la legislación de cada país, el castigo por este tipo de infracción puede ser una sanción penal o una sanción económica y la obligación de indemnizar los daños y perjuicios.	
0.94	El plagio es definido por el Diccionario de la lengua española de la Real Academia Española como la acción de ≪copiar en lo sustancial obras ajenas, dándolas como propias≫.	El plagio es definido por el Diccionario de la Real Academia Española como la acción de "copiar en lo sustancial obras ajenas, dándolas como propias".
0.40	En cualquier caso, la mera repetición de cadenas de palabras no es una prueba concluyente de deshonestidad intelectual; una gran parte del discurso científico es repetición de conocimientos (fórmulas, datos, etc.) e hipótesis compartidas por el conjunto de la comunicad científica, por lo que se deberían evitar los pronunciamientos apresurados sin un examen detallado de las posibles violaciones o suplantaciones de la autoría intelectual.	La denominada propiedad intelectual es una colección de marcos jurídicos diferentes que protegen los intereses de autores e inventores en relación a obras creativas, ya sean éstas, expresiones de ideas como en el caso del derecho de autor o aplicaciones prácticas e industriales de ideas como en el caso de las patentes.

The first case we discuss has nothing to do with text, but music. It concerns the 1970s *My Sweet Lord*, by *George Harrison*, released in 1971 within the *All Things Must Pass* album. The Britain number 1 hit resembled another number 1, in the US, released nine years before by *The Chiffons*: *He's So Fine*. The lyrics of both songs are unrelated. The similarity level becomes completely different when listening to the music, though. The melody, perhaps with a slightly different cadence, is practically the same for both songs.[44] The legal process found Harrison culpable, forcing him to pay a $587,000 fine (Wikipedia, 2011k,o). The most interesting fact from the point of view of plagiarism analysis is that Harrison claims he was not conscious of the re-use; he did not remember having listened to the other song. The similarities are so evident, that this seems to be a possible case

[44]We invite the reader to listening both songs and judge. If no record of G. Harrison or The Chiffons is at hand, the songs can be easily found on *Youtube*: `http://bit.ly/youtube_harrison` and `http://bit.ly/youtube_chiffons` (last visited 27/Jan/2012). Curiously, during a tutorial on text re-use and plagiarism detection we offered in India by the end of 2010 (`http://www.icon2010.in/tutorial3.php`), we reproduced the hits to the audience, which considered the songs were not that similar.

of unintended plagiarism; particularly cryptomnesia (cf. page 15).

Cases occur all around the world, implying lyrics as well and, in general, text. *Enrique Bunbury*, Spanish rock star, was qualified by the media as plagiarist after his album *Hellville de Luxe* (Alsedo, 2008; El País, 2008; Público, 2008). In particular for its first single: *El hombre delgado que no flaqueará jamás*[45]. The lyrics of the song include a few verses inspired by the late poet *Pedro Casariego* (El País, 2008). As Bunbury wrote, "through the story of popular music, great and unknown song writers have done similar practices, taking phrases from traditional songs, coming out with new, very different, creations" (El País, 2008). The idea seems nothing but reasonable. For example, he mentions the cases of Bob Dylan or John Lennon in music and Edgar Allan Poe or Shakespeare in literature. Bunbury affirms that "two phrases are not plagiarism" (Público, 2008), which could be questionable.

Staying in the Spanish setting, *Ana Rosa Quintana* is a Spanish TV host that aimed at entering the literary world. In 2000, she published the novel *Sabor a hiel*[46], with Grupo Planeta (once again, the publisher related to the Cela case) (Wikipedia, 2011a). More than 100,000 copies of this book about abused women was sold. However, it was found that the top seller contained various paragraphs from *Danielle Steel*'s *Family Album* and *Ángeles Mastretta*'s *Mujeres de ojos grandes*[47]. As a result, the publisher opted for retiring all the copies from the booksellers and cancelling the second edition (Rodríguez, 2000b). Once again, the plagiarist claimed a computer error was the cause of the fault, but according to Rodríguez (2000b), the paragraphs are reproduced from the sources changing the character's names. Afterwards she went even further by blaming a trustworthy collaborator for the fault (Mora, 2000; Rodríguez, 2000a). Quintana feels "not culpable, but responsible and victim" of the facts and apologised for the fraud to her readers.

The last case corresponds to the films world. A recent case occurred where *Arturo Pérez Reverte* was accused of plagiarising the script of the film *Corazones púrpura*[48], by *González-Vigil*. The audience determined that Corazones púrpura had been incorporated into Pérez Reverte's *Gitano*, with further modifications (ABC, 2011). According to Barrio (2011), as in Cela's case (cf. Section 2.4.1.4), both scripts, had been managed by the same producer.

2.4.1.7 Discussion on the Explosion of Plagiarism Cases

We have overviewed many cases of plagiarism in academia, research, journalism, literature, politics and show business. Plagiarism is not a phenomenon enclosed in a class room; it is diversified and exist across many disciplines and environments. A factor worth analysing is the impact the uncovered fault has had on the perpetrator. Whereas some people —zu Guttenberg or Shalit— lost their academic achievements or jobs, some others —e.g. Mejuto— have not suffered any consequence. Some others had to pay a fine, even when they seemed not to be conscious of the fault, as in the case of Harrison.

[45]The thin man that will never waver.
[46]Gall Taste.
[47]Big eyes women
[48]Purple Hearts.

As stated already, the final decision, as well as the sentence, is taken by experts, although not always seems to be the same one across the world. For instance, with respect to some of the previously mentioned plagiarism cases (in research and in academia), could the German and Spanish reactions be related to the cultural differences between the two countries?

One thing is clear: in many cases the plagiarist is conscious. She expects never to be uncovered and trusts that the link between her texts and the source will remain broken.

2.5 Surveying Plagiarism in Academia

As Martin (1994) points out, most intellectuals consider plagiarism to be a serious offence that should be discouraged among people. It is thought to be rare among scholars; but not so among students. However, plagiarism is more common among both students and scholars than recognised, as seen in Section 2.4.1. Different surveys have been carried out, trying to analyse plagiarism and cheating perspectives among students around the world.[49] We reviewed eighteen surveys that assess students and professors attitudes and experiences with plagiarism and cheating. Additionally, we ran a new survey among Mexican students, confirming and refusing some previous findings and rising some new insights.

2.5.1 Overview of Surveys on Plagiarism

One of the first studies we have track of is the one of Haines *et al.* (1986). The twenty five years old survey showed that more than 30% of the 380 considered students admitted cheating in their assignments at least once per academic annum. Fifteen years later, with the advent of the Web, plagiarism is on the rise (Baty, 2000), even generating new terms, such as cyberplagiarism (Anderson, 1999) (where the considered source is on the Web). Recent studies claim that nowadays Internet is the main source for plagiarism. Professors estimate that around 28% of their pupils' reports include plagiarism (Association of Teachers and Lecturers, 2008). The source of the borrowed material is diverse, but some people identify Wikipedia as a preferred one (Head, 2010; Martínez, 2009). Some surveys conclude that plagiarism is committed because of a lack of knowledge; students ignore what plagiarism is, what it represents, and how to avoid it. Many of them ignore the consequences of this misconduct or simply commit it because available technology makes it an easy short-cut (Pupovac *et al.*, 2008). Even political factors are identified as having strong influence. The same Pupovac *et al.* (2008) consider that post-communist and, in particular, countries with a high rate of corruption develop a high level of tolerance

[49] Plagiarism represents only one kind of cheating in academic environments, together with cheating in exams and quizzes. Haines, Diekhoff, LaBeff, and Clarck (1986) identified some factors that increase the probability for a student to cheat: (*i*) younger students, (*ii*) single students, (*iii*) those with lower grade-point averages, (*iv*) those whose parents pay for their tuition, and (*v*) those that consider that other students cheat as well. The survey they ran shows that while 25% of students admitted cheating on quizzes and exams, 33% of them admitted cheating on assignments. Only 1.3% of students reported to have been caught.

Table 2.5: Attitudes regarding exam cheating in seven surveys applied to undergraduate students. Values stand for percentage of affirmative answers. The corresponding results were published in: (*i*) Spain (Blanch-Mur *et al.*, 2006); (*ii*) UK, Bulgaria, and Croatia (Pupovac *et al.*, 2008); (*iii*) USA (Park, 2003); (*iv*) USA' and Hong Kong (Chapman and Lupton, 2004).

Question	Spain	UK	Bulgaria	Croatia	USA	USA'	Hong Kong
Have you ever cheated at school?	68				63-87	55.4	30.2
Have you copied during an exam?	23						
Do you consider cheating on exams acceptable?	25	7	18	20		2.3	2.2
Invalidating an exam is punishment enough for cheating?	77						

toward academic cheating.

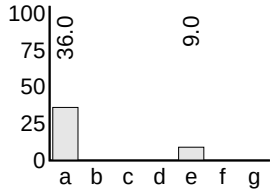
When analysing the broad topic of academic cheating, the interest has been focussed at whether students copy in exams or how serious this misconduct is considered. A summary of the results obtained in seven surveys is presented in Table 2.5. In three out of four surveys, more than half of students admitted cheating during their academic trajectory at least once. When comparing this high percentage to the number of people that consider cheating on exams acceptable, the difference is revealing. In the Spanish survey, 43% of students have cheated but, we could suppose, not during an exam (only 25% consider this fault acceptable). Presumably these students have cheated in a different way, probably by presenting a work they did not write themselves; a work they copied or bought.

We now centre our attention on plagiarism. We consider six surveys recently conducted in Spain and discuss them with respect to seven other surveys conducted around the world. Among the questions of the surveys, we found eight particularly interesting and revealing.[50] Table 2.6 presents the eight questions together with the results obtained by the six studies. The answers are sparse: at most five out of eight questions were included in a single survey. Nevertheless, the histogram representation is displayed for easy direct comparison.

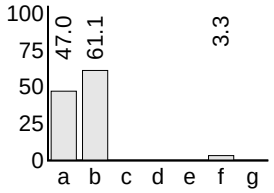
Question (a) regards at plagiarism from books, magazines, partners documents, and any other printed text. In this case the student is supposed to transcribe the contents he aims at borrowing. An important fact: if no computer accessible version of the text is available, the case could go unnoticed. Questions (b) to (d) try to reflect how students look at Internet as an optimal source for copying. They differentiate between copying excerpts (which could imply some kind of editing) and simple cut & pasting. The numbers obtained by Six Degrés *et al.* (2008) for questions (b) and (c) are particularly interesting. More than 90% of students admitted copying from the Web, but only 25% declared cut & pasting from this resource. As a result, presumably nearly 70% committed paraphrase plagiarism from the Web, rephrasing what was copied. Question (e) reflects how common the practice of self-plagiarism is. It includes presenting the same entire

[50]The questions are not exactly the same in every survey (and none of the surveys includes all the questions). In most cases they are simple paraphrases of each other, while in some others further interpretation was necessary to match them.

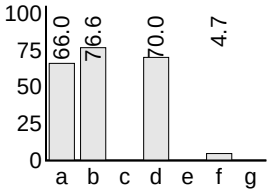
Table 2.6: Attitudes regarding plagiarism in six surveys conducted in Spain. The eight considered questions are at the bottom. The histograms represent the percentage of affirmative answers per question. Histograms legends stand for [*survey year; institution; students surveyed; source*] (UIB=U. de les Illes Balears, URL= U. Ramon Llull; UA = U. d'Alacant; "?" if no data is available).



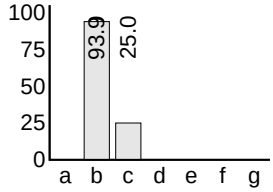
2004; URL; 468; Blanch-Mut et al. Blanch-Mur *et al.* (2006)



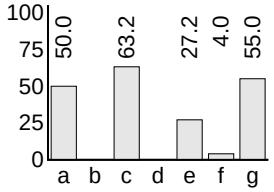
2007; multiple; 560; Comas and Sureda (2008b) Comas-Forgas et al. (2010)



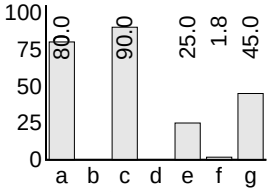
2007; UIB; 727; Comas and Sureda (2008b)



2008; ?; 299; Six Degrés et al. (2008)



2008; ?; 540; Sureda and Comas (2008)



2011; UA; 110; Beléndez Vázquez et al. (2011)

a) Have you ever plagiarised from non-digital sources?
b) Have you copied excerpts from the Web at least once?
c) Have you cut & pasted from the Web at least once?
d) Do you consider Internet is the main source for plagiarism?
e) Have you committed self-plagiarism at least once?
f) Have you ever bought works?
g) Have you fabricated references?

document to different instructors or just re-using some previously produced fragment when writing a new document. Depending on the case, these faults could be caused by lack of knowledge about plagiarism and/or bad investigation practices. Events analysed in questions (f) and (g) can hardly be justified in any way. The former one includes buying works from some physical source or appealing Web sources, such as paper mills. The latter one regards at whether students include references into their documents without actually consulting them, simply to give the impression of a well worked text.

The surveys are chronologically ordered from 2004 to 2011.[51] The aspect to note is that the number of plagiarists seems to be increasing over the years. The percentage of students admitting this borrowing from printed text starts at 36% in 2004 and goes to 80% in 2011. The impact of Internet and the *copy-paste syndrome* (Weber, 2007)[52]

[51]Four of them were developed by the *Grup de Recerca Educació i Ciutadania* (Education and Citizenship Research Group), *Universitat de les Illes Balears*. This group has paid special attention to studying academic plagiarism during the last five years in Spain and the European Union. (cf. `http://ciberplagio.es`).

[52]Indeed, Weber (2007) calls it "google copy-paste syndrome", but as we do not consider that Google is the responsible of this behaviour, we omit it.

seems not to be very influential. However, Internet and the easy access to information it provides has obviously increased the number of cases. While approximately 61% of students accepted borrowing contents from the Web by 2007, more than 90% did in 2008. This increase is in agreement with the number of students cut & pasting directly from their browsers. In 2008, 25% and 66% of the students declared doing so. The figure rose dramatically to 90% in 2011. The number of students re-using their own texts over and over again is on the rise as well. Levels are still far from previous figures though, with a maximum around 26%. Students seem to be under pressure. In order to make their works to seem better researched, they deliberately increase the number of references they consulted, even if they did not. Roughly, 50% of students are committing this kind of plagiarism. 54% of students surveyed in UK seemed to have similar practices back in 1995 (Franklyn-Stokes and Newstead, 1995). The number of people paying for works is close to zero.

These figures do not disagree with others obtained in different regions. According to Comas, Sureda, Nava, and Serrano (2010), nearly 40% of the students in a private Mexican university admitted plagiarising from printed documents. Almost 45% admitted borrowing texts from the Web. The figure is more dramatic when looking at Croatia and Hong Kong. The results presented by Pupovac *et al.* (2008) reveal that 82% of Croatian students copy from the Web; the level raises to 92% in Hong Kong (Chapman and Lupton, 2004). Respect to self-plagiarism, rates are at similar levels in two European countries, UK and Bulgaria: around 40% (Pupovac *et al.*, 2008). Yet another Spanish study shows an increase, by 2010, to 67% (Comas Forgas, Sureda Negre, and Oliver Trobat, 2011).

Finally, a question not considered before: whether students believe that instructors are able or not to detect their misbehaviour. British students seem to be more worried than Bulgarian. While only 47% of the former felt confident, 85% of the latter considered plagiarism was not uncovered by reviewers (Pupovac *et al.*, 2008). The reason could be UK's broad use of computer programs for automatic plagiarism detection together with well established conduct regulations.

These surveys contemplate the students' point of view. Nevertheless, tutors point of view has to be listened as well. For instance, let us consider a survey conducted in UK in 2001 (Bull *et al.*, 2001). 72% of teachers considered that, rather than having the source of a specific re-used fragment, a change in writing style through a document is enough signal of plagiarism.[53] The extent at which plagiarism has infiltrated the scholar environment is a reason for alarm. Chapman and Lupton (2004) found that 80% of the fifty-three surveyed professors declared having found plagiarism in their students works. Even worst, 87% of the people surveyed by Morgan and Foster (1992), nearly twenty years ago, consider that people who routinely cheat in education repeat in workplace.

From the point of view of text re-use and plagiarism detection, what these surveys show is interesting. Firstly, even if a perfect system for plagiarism detection could exist, many cases will remain undetected. Two facts back this claim: (*i*) there are many cases of plagiarism for which the source text is just no there —not in the Web, not in any (public) electronic repository and (*ii*) it is hard (nearly impossible) to automatically figure out

[53]This changes can be often detected by means of intrinsic plagiarism detection (Section 5.1.1).

whether a person actually consulted a properly cited material (models could be generated that detect dangling references, though.) Secondly, Internet is certainly having a negative impact on plagiarism. More and more cases of borrowing from electronic documents occur. However, the numbers indicate that, with or without Internet, plagiarism is being committed more often every day. These are the cases that could be detected by means of a computational model.

Paraphrasing Boisvert and Irwin (2006), plagiarism, and not only because of the Internet, is certainly on the rise.

2.5.2 A New Survey on Plagiarism Attitudes

Whereas the aforementioned surveys show clear numbers and attitudes respect to plagiarism in academia, they lack of providing insights on one particular case of misbehaviour: cross-language plagiarism. In order to analyse students' attitudes to this kind of text re-use, we developed a new survey on "Investigation strategies, text re-use, and plagiarism", addressing specially the cross-language plagiarism issue.[54] The survey consisted of 40 questions, divided in four blocks: (*i*) general information; (*ii*) scholar practices; (*iii*) attitudes respect to plagiarism; and (*iv*) final opinions. Answering every question was mandatory.

The survey starts with general questions on how students investigate and write. It gradually gets into more specific aspects of plagiarism and academic cheating. In order to avoid considering posterior facts when answering to previous questions, we asked the respondents not to get back to modify any answer once they had passed it.

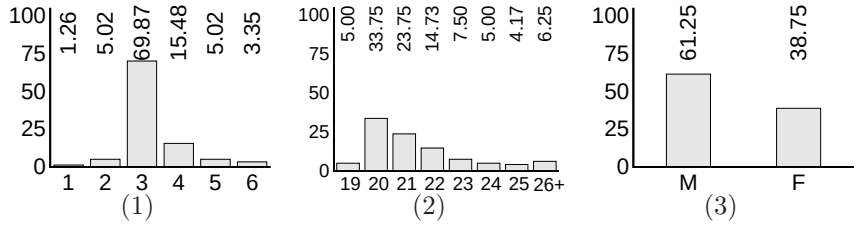
The survey was applied to more than 250 Mexican students from four different states. Three from the centre: *Universidad Tecnológica del Valle de Toluca*, *Instituto Tecnológico de León*, and *Instituto Tecnológico de La Piedad*; and one from the north: *Universidad Autónoma de Sinaloa*. Bachelor, pre-grade and post-graduate students took part in the survey. For this analysis, we only considered the 240 pre-grade students that completed the survey. Following, we describe the four sections and discuss the obtained results.

2.5.2.1 General information

In the first section we required some socio-demographic information. We aimed at preserving anonymity, and avoided requesting information that might uncover volunteers' identity. This decision was taken in order to encourage honest answering. As a result, not even the name of the participants was requested. The three questions and a summary of the obtained answers are included in Table 2.7. Most of the surveyed students are in their third year of studies. One third of them are 20 years old and just a few are older than 26 (the oldest one is 33 years old). The mean age is 21.72 ± 2.94. The sample is slightly skewed through male participants, which seems to reflect students' population

[54] The survey was conducted in several Mexican universities, as Adelina Escobar Acevedo, a researcher from that country interested in the topic, contacted us after knowing the research topic of this PhD.

Table 2.7: Survey general information. Students were requested to provide some socio-demographic information regarding their age, gender, and year of studies. The numbers on top of the bars stand for the percentage of participants in the corresponding group. The meaning of the bottom numbers is located next to the corresponding question.



(1)	Year of studies	$[1, 2, \ldots 6]$ years
(2)	Age	$[19, 20, \ldots 33]$ years old
(3)	Gender	$[Male, Female]$

in the considered disciplines (most of them related to computer science and engineering). Question (4) requested to select the institution the student belonged to. We omit it in this study.

2.5.2.2 Scholar Practices

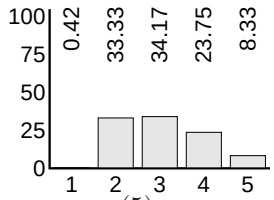
We aimed at analysing students' strategies when developing some assignment. For 13 out of 14 questions a *Likert scale* was used with five possible values: *Never*, *Sometimes*, *Often*, *Very often*, and *Always*. One question had to be answered either Yes or No. The questions and the obtained answers are included in Table 2.8.

Questions (5) and (6) reflect very well the sources students use to consider when facing an assignment. The most recurred resource is the Web; for 60% of assignments at least one Web page is consulted. For the case of printed material it is the other way around. According to question (7) a search engine is used to obtain the information in most cases, whereas Wikipedia is often a considered source (question (8)). The encyclopedia is identified as a favourite source for academic plagiarism (Head, 2010; Martínez, 2009). Therefore, we carried out research on the analysis of text re-use inside of Wikipedia and started analysing cases of borrowing from it (cf. Chapter 9).

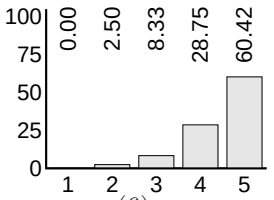
The following questions are of particular relevance for our research. These are questions never asked before in a survey of this nature and regard at analysing how likely it is that a student will re-use text from a different language than her native one. The numbers obtained from question (9) are clear: it is quite likely that a student will consult material in a language other than Spanish (the native language of the respondents). In agreement with these figures, question (10) shows that automatic translation systems are commonly used.

If we compare the outcomes of questions (11) and (12), according to question (11) almost two thirds of the population include text fragments from the Web in their reports often or very often (in contrast, only 5% declares never doing so). The numbers are not very different when the included text is translated from a foreign language (ques-

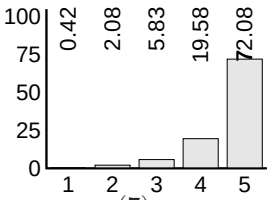
Table 2.8: Survey scholar practices. Students were questioned about their investigation strategies when approaching an assignment. The numbers at the bottom of the histograms stand for: (1) Never, (2) Sometimes, (3) Often, (4) Very often, and (5) Always. The numbers on top stand for the percentage of participants that selected a given option. Bar charts and questions related to cross-language issues and Wikipedia are highlighted.



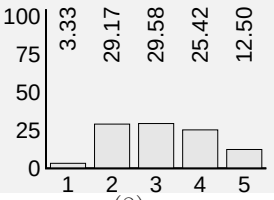
(5)



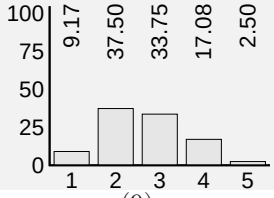
(6)



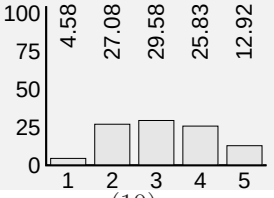
(7)



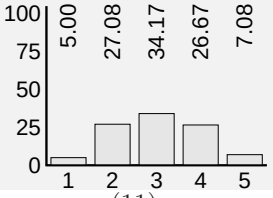
(8)



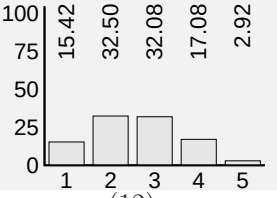
(9)



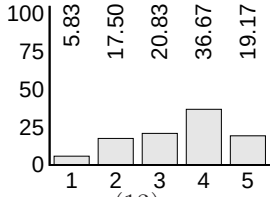
(10)



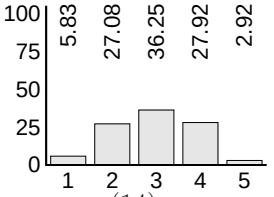
(11)



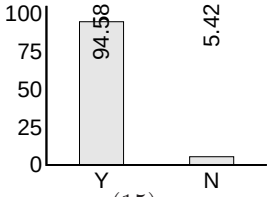
(12)



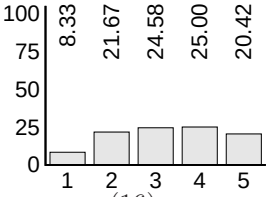
(13)



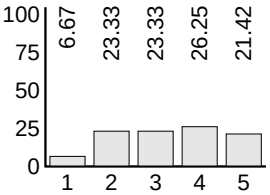
(14)



(15)



(16)



(17)

(5) When doing a report, I consult printed material (books, journals, etc.).

(6) When doing a report, I consult electronic material (websites, e-journals, etc.).

(7) I use search engines as Google, Yahoo! and Bing.

(8) **I consider Wikipedia articles as an information source.**

(9) **I consult sources written in a language other than Spanish (English, French, etc.).**

(10) **I use automatic translators in order to translate contents in a foreign language relevant to the topic I am approaching.**

(11) I include fragments from Web pages in my reports.

(12) **I include fragments translated from Web pages in my reports.**

(13) I use images from Web pages in order to illustrate my reports.

14) I copy text fragments from printed sources to include them into my reports.

15) I have cut & pasted a text fragment from a Web page and included it into a report at least once.

16) When I consider printed material (books, journals, magazines, newspapers, etc.) in my report, I include the corresponding bibliographic reference.

17) When I consider electronic material (from Web pages, electronic journals, etc.) in my report, I include the corresponding bibliographic reference.

tion (12)). The correlation with respect to question (11) is 0.58 with $p = 0.00$. Almost 50% of students translate contents for inclusion in their documents often or very often! Therefore, text re-use across languages is indeed a common phenomenon. We consider that the numbers would raise even more in more multilingual scenarios. As we identified a gap in current research on text re-use and plagiarism detection regarding cross-language cases, we worked extensively on this problem (cf. Chapter 6 as well as Sections 7.5, 9.4, and 9.5).

As aforementioned, re-use and plagiarism are not limited to text. Question (13) tries to analyse how often the students use images downloaded from websites for illustrating purposes: 55% of them make it very often or nearly always. When comparing these figures to those of question (11), it becomes evident that images are more commonly re-used than text. This is certainly a gap that should be filled. Indeed, other IR communities such as that of ImageCLEF, which aims at providing "an evaluation forum for the cross-language annotation and retrieval of images"[55], could be interested (and capable) in looking at this problem.

Questions (14) to (17) try to contrast the most common re-use practices when the source of the borrowed text is printed or electronic. According to the first question of this block, more than one third of the students copy excerpts from printed material often. Almost another third accepts following this practice very often. Question (15) shows that nearly the entire surveyed population has cut & pasted at least once from the Web. These percentage is higher than those obtained by Six Degrés *et al.* (2008), Sureda and Comas (2008), and Beléndez Vázquez *et al.* (2011), and only comparable to the most recent one (cf. Table 2.6).

The following queries investigate whether these borrowings include the corresponding citation. Questions (16) and (17) analyse this issue when considering printed and electronic material respectively. Interestingly, only around 7% of students admits not including any reference, despite the source is electronic or printed. The distribution of answers for the rest of the options is very flat. Around 20% of the surveyed population is included in every group; those that include the corresponding citation sometimes, often, very often, or always. There is no significant difference in the behaviour when considering electronic or printed material.

After analysing this set of questions it is clear that texts from printed and electronic sources (and even images) are being re-used very often indeed. Nevertheless, the corresponding credit is not included so frequently. It is claimed that the Web has increased the amount of plagiarism cases, and it seems to be truth (we go deeper into this issue in the next set of questions). However, from our point of view, the documentation and citation practices seem to be the problem. This is reflected by the outcome of the two last questions. More and better instruction on citation practices is necessary.

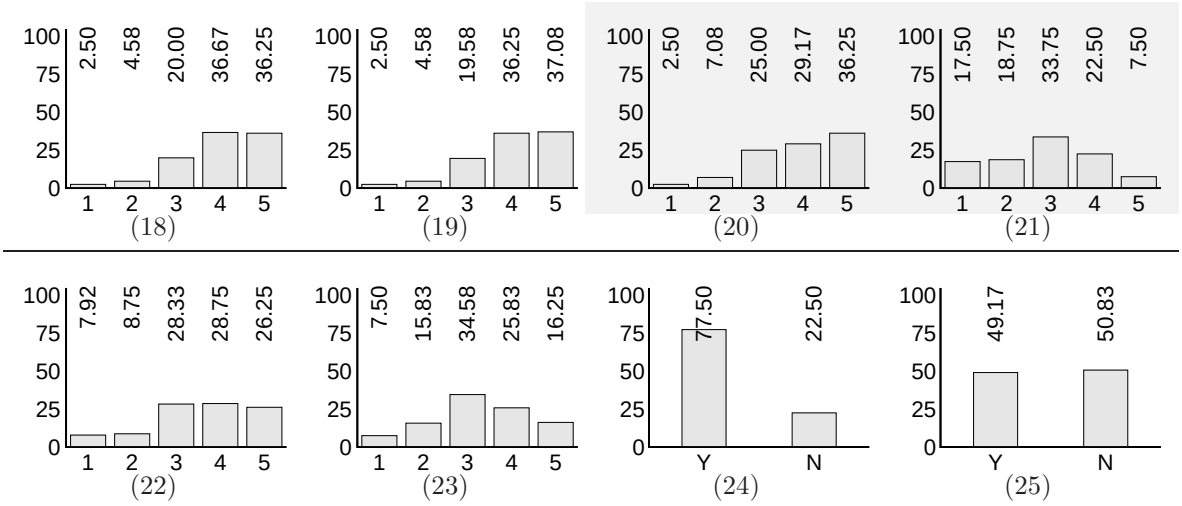
2.5.2.3 Attitudes Respect to Plagiarism

We aimed at realising how students think of plagiarism. In 7 out of 16 questions we used a Likert scale with five possible values: *Strongly disagree*, *Disagree*, *Neutral*, *Agree*,

[55] http://www.imageclef.org

and *Strongly Agree*. The rest 9 questions can be answered either with Yes or No. The questions and obtained results are included in Tables 2.9 and 2.10.

Table 2.9: Survey attitudes respect to plagiarism (1 of 2). Students were asked about their knowledge and attitude respect to plagiarism. The numbers at the bottom of the histograms stand for: (1) Strongly disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree. Bar charts and questions related to cross-language and paraphrase plagiarism issues are highlighted.



(18) To plagiarise means copying a text and including it into a document without modifying it, and without mentioning the source.

(19) Re-using a text fragment is not plagiarising, as far as the source document is cited.

(20) **If I express with my own words an idea I have read, I do not need to include any citation and I am not plagiarising.**

(21) **Translating a text fragment and including it into my report is not plagiarism, as the words are different.**

22) I have made a report and presented it to a class. If there is a different class in which I can present it again, I can do it without problem, as I wrote it myself.

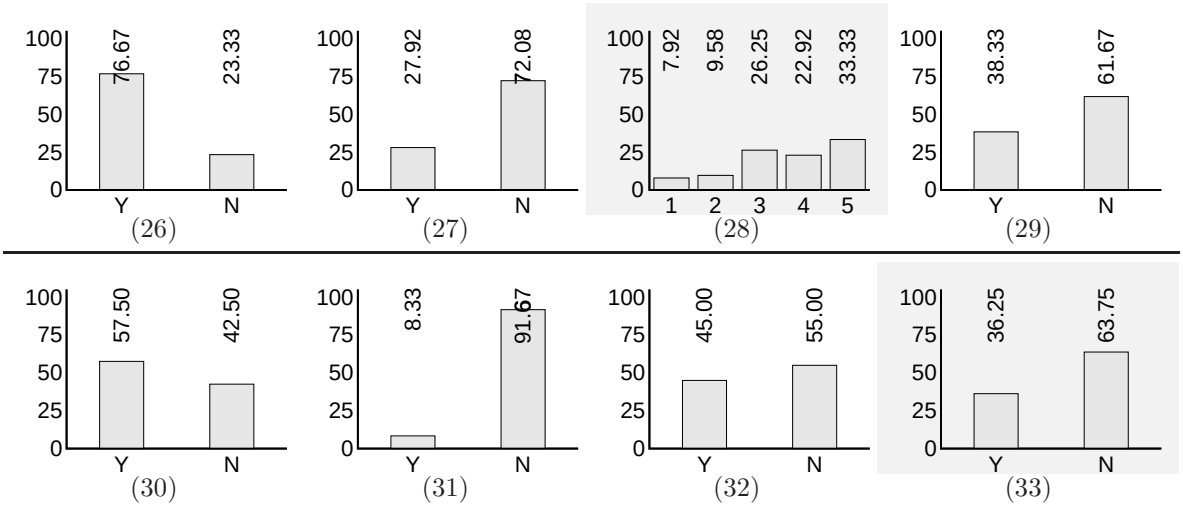
23) Images downloaded from Internet can be used without including any reference.

24) I have plagiarised at least once.

25) I have plagiarised at least once from printed sources.

Question (18) aimed at realising whether students agreed with a "common" definition of plagiarism: copying text without citing. More than two thirds of the population agreed or strongly agreed. On the other side, 7% of students strongly disagreed. Following this trend, we asked in question (19) whether citing was enough for a re-used text to be not considered plagiarism any more. The distribution is practically the same than for the previous question. These two questions have to do specifically with text borrowing. In question (20) we went further and asked whether re-using an idea (independently of words) required citation. Once again the numbers are extremely similar. This behaviour clearly reflects that in general students accept that plagiarism occurs when text is re-used without citation, but they lack of the principle of citing an idea they read about but paraphrase. It seems clear that for them the concept of plagiarism covers verbatim copy

Table 2.10: Survey attitudes respect to plagiarism (2 of 2). Students were asked about their knowledge and attitude respect to plagiarism. The numbers at the bottom of the histograms stand for: (1) Strongly disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree. Bar charts and questions related to cross-language issues and Wikipedia are highlighted.



(26) I have plagiarised at least once from electronic sources.

(27) I have presented at least once (practically) the same work to different classes or professors.

(28) **Among all of the websites, Wikipedia is one of the most frequently used for plagiarising.**

(29) I have included, at least once, references in my reports in order to make it appear better researched, but without actually consulting them.

(30) I have copied from my fellows works at least once.

(31) I have paid a fellow for doing a work for me at least once.

(32) I have presented reports downloaded from specialised websites on the Internet at least once.

(33) **I have plagiarised from sources in a different language than mine at least once.**

only, something far to be truth; paraphrase plagiarism may occur very often as well. In Chapter 8 we illustrate the insights of the research done on this kind of re-use.

The trend of similarity stops in question (21). Note that this question is quite similar to question (20). The only difference is that in this case the text is translated before re-using it. One third of the students show not to be sure about whether translation implies plagiarism. Another third considers that a translated passage does not require any reference: cross-language plagiarism is very likely to happen.

The following questions try to analyse specific plagiarism-related misconducts. Question (22) regards at determining how welcome practices related to self-plagiarism are among the students. More than half of them agree or strongly agree that, given that they wrote some report, they are free to use it for different lectures once and again. Only 17% of them disagree with this idea.

Question (23) is highly related to question (13) in the previous set. We already noted that students use images from Internet very often. Here we can see that more than one third of them are not sure whether these images should be referenced. The bad news is

that more than 40% consider that images can be used citation-less.

The following three questions have to do with whether students admit having plagiarised at least once in their academic life. The first of them, question (24), shows that more than 75% of the people admit plagiarising at least once. Questions (25) and (26) are more specific and request for declaring if students have plagiarised at least once from printed or electronic material, respectively. Only 50% admits plagiarising from printed, respect to 77% that do so from electronic material. For the case of printed material, the numbers are comparable to those of Table 2.6 (question a), except for the result obtained by Beléndez Vázquez *et al.* (2011) (80% of the students had admitted plagiarising from non-digital material at least once. The idea that Internet has caused an increase in the cases of plagiarism is somehow supported by these figures.

Question (27) is very related to question *(e)* in Table 2.6; whether students have presented roughly the same work to different classes. In our case, 27% of them admitted so. Practically the same numbers were obtained by Sureda and Comas (2008) and Beléndez Vázquez *et al.* (2011).

As aforementioned, Martínez (2009) identified Wikipedia as one of the plagiarist's favourite sources. Answers to question (28) seem to support this claim. More than 55% of the students agree or strongly agree with this affirmation.

The next four questions are definitively conscious kinds of misconduct regarding plagiarism. Question (29) has to do with the deliberate inclusion of non-consulted references. As shown by the distribution of the answers, nearly 40% of students have done so, just to make their document seem better worked. One more time, we can compare these numbers to those of Table 2.6 (question g). The figure in the surveys of Sureda and Comas (2008) and Beléndez Vázquez *et al.* (2011), are around 50%. Question (30) refers to whether students have copied a work from their peers. More than half have committed this misconduct. According to question (31), only 8% have paid their fellows for doing the work for them. The number increased when question (32) asked whether they had presented reports from Internet specialised websites, such as paper mills: up to 45%.

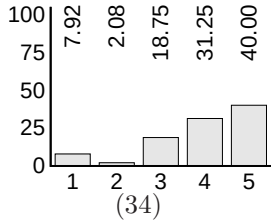
The last question in this set just remarks our findings on attitudes respect to cross-language plagiarism. More than 36% of the students answered affirmatively to question (33); whether they had plagiarised from a source in a foreign language. This is certainly an amount of potential cases of plagiarism that requests for attention.

2.5.2.4 Final Opinions

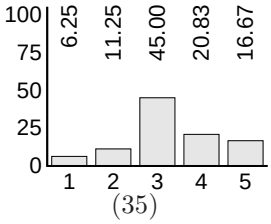
In the last set of questions, we aimed at knowing further students' thoughts regarding plagiarism. We try to analyse the reasons behind academic plagiarism. Once again, we used a Likert scale with five possible values: *Strongly disagree*, *Disagree*, *Neutral*, *Agree*, and *Strongly Agree*. The questions and obtained results are included in Table 2.11.

Question (34) regarded at knowing the students own perception of Internet as the main source of the plagiarism cases. More than half of them agreed or strongly agreed about this assertion. Pupovac *et al.* (2008) consider that countries with high rates of

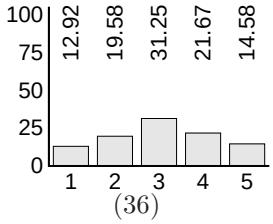
Table 2.11: Survey final opinions. Students were asked about the reasons and attitudes behind plagiarism. The numbers at the bottom of the histograms stand for: (1) Strongly disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree.



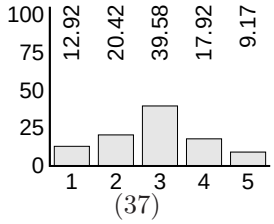
(34)



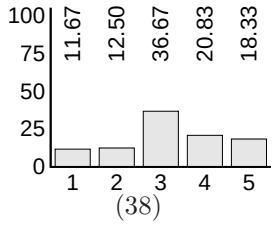
(35)



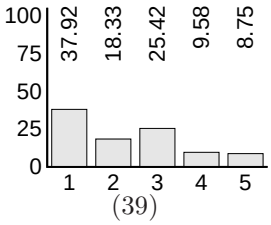
(36)



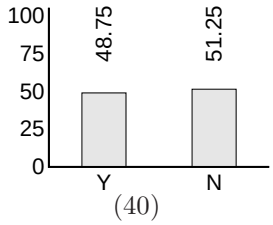
(37)



(38)



(39)



(40)

(34) Internet is the main source for cases of plagiarism.

(35) I would not accuse a fellow if I knew he committed plagiarism.

(36) I am requested to much work, causing me to plagiarise in order to comply with all of them.

(37) Professors do not detect cases of plagiarism.

(38) Professors plagiarise as well, so there is no matter if I do.

(39) In this country everybody cheats, so I do not see why I would not.

(40) If you knew that a report is going to be automatically analysed by a computer program, looking for cases of plagiarism, would you plagiarise anyway?

corruption develop a high level of tolerance toward academic cheating. Being Mexico considered as such a country, we aimed at studying this fact through question (35). The obtained results are not very clear. Almost half of the students seem not to be very sure about how they would act in case they discover a peer fault. Whereas 17% let see that they would accuse a fellow, more than one third says the contrary. The level of tolerance does not seem to be so high.

As aforementioned, some scholars justify the increase in cases of plagiarism by the fact that students are required more work than they manage to comply with. This was precisely what question (36) was about. The answers are completely divided: one third is completely neutral, another disagrees and the last one agrees. Evidently, the load of work is not the only trigger.

Questions (37) and (38) regarded at students perceptions about their professors. Firstly, we asked how students felt about the risk of being caught "red-handed". If they thought that professors are able to detect cases of plagiarism. More than one third decided to stay neutral, avoiding to give their opinion. Nearly 30% considers that professors do not detect plagiarism. Secondly, we asked students if they thought that professors were plagiarist as well, somehow decreasing the gravity of their own cheating. Almost 40% of them declared that they believed so, and 25% disagreed. Question (39) is related to the assertion of Pupovac *et al.* (2008), whether the perception of the students

was that their misconduct is somehow justified by an environment full of cheaters. More than 55% disagreed and 25% decided to stay neutral on this issue.

Question (40) is much related to the main topic of our research: whether students would plagiarise even if their reports would be analysed by an automatic plagiarism detector. More than 50% admitted they would avoid committing plagiarism in this case. Beside the quality of the results obtained by means of the software, this seems to be already a success.

At the end of the survey we invited the respondents to write an optional final comment. Some of them were extremely interesting. One of the students considers that plagiarism is hard to eradicate because of the low academic level and the low interest it causes in students. She considers that better lectures would cause more interest and less temptation to misbehave. Many participants consider that the information on the Web is there with the aim of sharing and, therefore, re-using it is not plagiarism. Let us recall the quote that kicks off this document: "Many students seem to almost reflexively embrace a philosophy rooted in the subculture of computer hackers: that all information is, or should be, free for the taking", by Mark Fritz.[56] Participants with "better intentions" mention that some websites do not include the author of the information they contain. As a result, they ignore how to cite it. Still others consider the existence of a circle where the student commits plagiarism, the professor does not uncover it and, as a consequence, he does not take any corrective action, letting the student to cheat again. A few students mentioned that they had never looked at plagiarism from the point of view presented in the survey. They consider that their attitude would change. On the one hand, many of the answers and final comments show the lack of citation, documentation and plagiarism culture that students have in general. On the other hand, running the survey was useful to analyse students perspectives but also to change some of them.

As observed throughout the queries of the survey: (*a*) paraphrasing is perceived as a mechanism of re-use that does not require citation because it allows to change a text and make it "more yours", (*b*) cross-language text re-use and plagiarism are common practice nowadays, and (*c*) Wikipedia is observed as a likely source for re-use. These are the three axes of this research: detection of paraphrase plagiarism, detection of translated plagiarism, and detection of text re-use in Wikipedia.

2.6 Automatic Text Re-Use and Plagiarism Detection

The difference between analysing plagiarism or other kinds of text re-use is not clear. It should be noted that, as Clough *et al.* (2002, p. 1680) state, from a text processing point of view the tasks of distinguishing plagiarism and journalistic text re-use (cf. Section 4.2.1) are similar. Therefore, we treat them here as equivalent.

Detecting text re-use can be considered as a "generic attribution technology" (Wilks,

[56] As seen in (Mallon, 2001, p. 245).

Table 2.12: Taxonomy of text attribution technology as proposed by Wilks (2004, p. 119).

(*i*)	Of these k texts or sets of texts, to which is text d most similar?
(*ii*)	Has this set of k texts one or more subsets whose members are improbably similar to each other?
(*iii*)	Is there a subset of texts (out there in a wide set like the Web) similar to text d?
(*iv*)	Is text d_1 improbably similar to text d_2?

2004, p. 118). Wilks himself offers a taxonomy of such attribution technologies, which is reproduced in Table 2.12. Task (*i*) is identified to be located between plagiarism detection and authorship identification. In the former case, the claimed sources are known, whereas in the latter the possible author candidates are known. Task (*ii*) resembles detecting re-use within a closed set of documents; for instance, cheating on student exercises, or self-plagiarism. Task (*iii*) regards at searching for cases of re-use from the Web. Finally, task (*iv*) is more specific than tasks (*ii*) and (*iii*) as it looks at two specific documents rather than a comprehensive set. In this dissertation we investigate all the tasks except for the one related to authorship identification.

Before discussing automatic approaches, let us boarding the "traditional" (manual) plagiarism detection process. Plagiarism detection...

> *involves finding similarities which are more than just coincidence and more likely to be the result of copying or collaboration between multiple authors. In some cases, a single text is first read and certain characteristics found which suggest plagiarism. The second stage is to then find possible source texts using tools such as Web search engines for unknown on-line sources, or manually finding non-digital material for known sources* (Clough, 2003, p. 4).

Human beings are the best in uncovering plagiarism, but they are unable to keep track of all the information they could require. Therefore, a necessity for automatising the process arise. The models for automation are known as automatic plagiarism detectors. As expected, the task of automatic text re-use and plagiarism detection is not isolated. We locate it inside of three main axes: forensic linguistics, natural language processing, and information retrieval.

In the case of FL, typical problems related to written language are investigating whether a suicide note is real and determining who wrote a threat letter, among others. The aim is one of authorship attribution: determining whether a text was actually written by whom it is supposed to. Approaching these problems resemble the analysis of textual evidence (cf. Section 2.3).

According to Jackson and Moulinier (2002, p. 3), NLP aims at "analysing or synthesising spoken or written [natural] language". In NLP, text re-use analysis is highly related to the analysis of paraphrases, as paraphrasing occurs in many plagiarism instances (cf. Chapter 8). The core of one of the two main approaches to text re-use detection is measuring similarity. This is the core of other tasks as well, such as summarisation, in particular of multiple documents (Goldstein, Mittal, Carbonell, and Kantrowitz, 2000).

The third axis is IR. IR is defined by Jackson and Moulinier (2002, p. 26) as the "application of computer technology to the acquisition, organisation, storage, retrieval, and distribution of information". Tasks related to text re-use and plagiarism detection have to do with near-duplicate detection (Potthast and Stein, 2008) and text categorisation and clustering (Pinto, Benedí, and Rosso, 2007), among others. In the former case, nearly exact copies of a document are searched for (which indeed, are a kind of re-use). If website d_1 is relevant to the user and it is similar to d_2, it is very likely that d_2 is relevant as well. Nevertheless, if both documents are too similar, they could be a case of near-duplication, causing one of them to become practically irrelevant.

Automatic plagiarism detectors have been proposed during the last 40 years. One of the first models we are aware of is the one proposed by Ottenstein (1976), that aimed at detecting plagiarism in students' source code. Analysis of programming languages is considered less complex than analysis of natural languages, as they have a more rigid syntactic structure. This causes the detection of re-used code to be more affordable. However, when facing documents in natural language, known as free text, the task is more complicated.

The methods approaching detection in natural language can be grouped into two main approaches: intrinsic and external. In brief, in intrinsic detection the aim is determining whether a text fragment within a document has been borrowed from another one. Stylometric and complexity features are considered. In external detection the aim is identifying a borrowed text fragment in a document together with its source.[57] Therefore, IR techniques are necessary to retrieve documents similar to the one we are analysing; NLP techniques are required to analyse the style and complexity of a text and to determine how similar (even at semantic level) two documents are; and FL can be benefited by the automation of part of its analysis process.

As it is going to be shown throughout the document, the higher the modification applied to the source text when re-using, the harder to detect it. The most simple setting is that of cut & paste re-use (verbatim copy), but the difficulty increases with paraphrasing: from a simple word insertion, deletion or substitution, to an entire process of reformulation up to translation from one language into another. From our point of view, one of the most interesting, and hard to detect, kind of re-use implies a translation process from one language into another.

2.7 Commercial Plagiarism Detection Tools

The explosion in number of cases of plagiarism is a business opportunity as well. Plenty of commercial —and non-profit— services have been created during the last twenty years. Their objective: detecting and preventing cases of plagiarism in free text. Whereas the most of them are particularly focussed to academic environments, some others have a broader scope, mainly upon the Web.[58]

[57] The two approaches are discussed in depth in Chapter 5.

[58] A particular case is that of eTBLAST. It is not a system for plagiarism detection, but a "text comparison engine" that searches for the keywords in the input over a set of (public) databases, such as Pubmed, Medline or ArXiv (eTBLAST, 2011).

Table 2.13 offers a brief comparison of some of these products.[59] No technical information is provided as practically all of them keep the technology behind their detection engine hidden. Following we discuss the particularities found in some of the systems. Only a bunch of systems offer reports on how long analysing a document takes. One of them is *Academic Plagiarism Checker* (2011), which estimates that scanning a document takes up to five minutes (they consider that analysing a thesis dissertation could take up to four hours.) The workload for analysing a document is reflected by this system. Its premium edition lets for scanning only five documents a day, while the free edition allows for one every three days. This limitation could be also caused by the system's Web search module dependency on search engines, such as *Google*, *Yahoo!*, or *Bing*, which often limit the amount of queries they serve. To avoid these limitations, other systems offering the capability of Web search, such as Turnitin (iParadigms, 2010), maintain their own Web crawlers and carry out internal Web search.

Turnitin is probably one of the most popular systems: the BBC (2011) reports that 98% of UK universities use it. Their reported 155 million student papers, 110 million documents and 14 billion Web pages, together with their wide range of writing assisting modules compose Turnitin's strengths. According to Suri (2007) their core technology is based on finding matches of strings of eight to ten words.[60] Therefore, some scholars consider it is only good for detecting exact copies, and probably not so good to detect paraphrase plagiarism. Their analysis also reveals that it does not differentiate between actual cases of plagiarism and properly quoted material. This seems to be an open research issue: no system (or research development) seems to have approached this problem yet. As far as we know, Turnitin is the only system with capabilities for detecting cross-language plagiarism. It offers to detect cases of plagiarism that were generated from English to other languages, and presumably do so on the basis of automatic machine translation (Turnitin, 2010). The content in the suspicious document is translated into English and the rest of the process is done at monolingual level.[61] As Culwin (2008, p. 190) points out, Turnitin seems to be "capricious in its operation"; in his experience, submitting the same document once and again can result in completely different outputs. This makes us suspect that this system randomly (or at least heuristically) selects a few fragments of a suspicious document to further analyse them.

Together with Turnitin, *Grammarly* (2011) is another service that provides a broad offer of writing assistance technology. It checks documents for grammar, spell check and, of course, plagiarism. An interesting characteristic of this service is its post-detection process. Instead of simply reporting a case of plagiarism and its potential source, it proposes an appropriate citation for the borrowed text fragment. The focus of Grammarly is clearly on plagiarism prevention. This was also the case of *PlagiarismDetect* (2011) that aimed at assisting writers (however, since August 2011, the service is closed).

An important issue that has been briefly commented already is how the systems query the Web or maintain their own text databases. Many online systems exist, such as *Duplichecker* (2011), which are simple interfaces for quoted queries to search engines.

[59] We are aware of some of them thanks to the surveys published by Maurer *et al.* (2006) and Sureda, Comas, and Morey (2008), together with a thoroughly Web search.

[60] As seen in (Jones *et al.*, 2008).

[61] This is a common approach to cross-language plagiarism and we further review it Chapter 6.

Table 2.13: Overview of "commercial" plagiarism detectors. It indicates whether the system: (*i*) compares documents against Internet through a Web search; (*ii*) compares documents against a *private database*; (*iii*) performs a *pairwise comparison* among submitted documents; (*iv*) performs *cross-language* comparison; (*v*) includes a *student service* to prevent plagiarism; (*vi*) it is *opensource*; and (*vii*) it is *free*. A white square stands for services with both free (limited) and paid versions. Information as published in the systems websites. There could be features missed in the table that systems accomplish, but were not found on the organisations' public website.

Name	Web search	Private database	Pairwise comparison	Cross-language	Student service	Opensource	Free
Academicplagiarism	■	■			■		□
Chimpsky	■		■				■
Compilatio.net	■	■	■		■		
Copionic	■		■			■	
Copycatch			■				
Copy Tracker	■	■	■			■	■
Crot	■	■	■			■	■
DOC Cop	■		■				■
Docode	■						□
Docol©c	■						
Dupli Checker	■						■
Ephorus	■	■					
eTBLAST 3.0		■	■				■
Eve2	■						
Grammarly	■				■		
PlagiarismDetect.com	■				■		
Plagiarism-detector	■		■				
PlagiarismScanner.com	■				■		
Plagium	■						□
PlagScan	■	■					
Safe assign	■	■			■		
Sherlock			■			■	■
Turnitin		■		■	■		
Urkund	■	■			■		
VeriGuide					■		
Viper	■	■	■		■		□
WCopyfind			■				■
Yap3			■			■	■

Others, such as the aforementioned Academic Plagiarism Checker and Turnitin, create indoor databases to perform different pre-processing actions and offering a more controlled search. Many of them crawl the Web to retrieve such a database, but there are other, personalised, approaches.

Ephorus (2011), *Plagscan* (2011) and *Copytracker* (2011) allow users to upload their documents into the server in order to compose their own private database. Others, such as *Safe Assign* (2011) compose the institutions database of their submitted documents. Ethical issues come out when the companies must decide whether the documents submitted by the users for analysis should be integrated into the systems database, making them available for the rest of users. *Urkund* (2011) and others do that.[62]

More pragmatic systems assume that it is very likely that plagiarised texts have been acquired from paper mills and compose their database of their contents. As reported by the *PlagioStop project*[63], this was the philosophy behind the Spanish *Educared* plagiarism detection system[64], that exploited the paper mill known as *El rincón del vago* (The corner of the vague)[65]. Unfortunately, this application does not seem to be available any more.

An issue has to be noted when considering all of the possibilities: if the source of a case of plagiarism is not available either in the systems own database or on the Web, the system will fail, and claim that the text is original. *Copycatch* (2011) tries to avoid this problem by analysing the suspicious document contents in isolation. It looks for unexpected changes in style and complexity based on the known as intrinsic plagiarism analysis (cf. Section 5.1.1 for an overview of the approaches to this kind of detection.)

A question that remains open is what systems perform best. As most of them are closed and do not report any evaluation on their models, this question is practically impossible to answer. The only information at hand can be found by considering the people behind two of these systems, whom participated to the International Competition on Plagiarism Detection (cf. Chapter 7). It is the case of *Crot* (2011) and PlagiarismDetector (2011), which got the sixth and fourth position in the 2009 edition, respectively (Palkovskii, 2009; Scherbinin and Butakov, 2009)

As already discussed, diverse commercial plagiarism detectors exist. Some of them provide good results, particularly in verbatim copy cases, where the re-use implies no further paraphrasing. Indeed, exploratory experiments have shown that a few word substitutions can prevent these systems to properly detect cases of plagiarism (Gillam, Marinuzzi, and Ioannou, 2010). Up to 2010, no commercial plagiarism detector existed that was capable to detect cases of translated plagiarism. In their study, Maurer *et al.* (2006) mention that neither Turnitin, Mydropbox, nor Docol©c supported cross-language analysis. Interestingly, Turnitin is starting to look at the problem of cross-language plagiarism, demonstrating, once again, that this was a gap in plagiarism detection that needed to be filled.

Sureda *et al.* (2008) offer an interesting overview of by-products of the worldwide

[62]This issue has caused many conflicts as some people claim that their texts should not be kept in that kind of databases without their consent.

[63]`http://plagiostop.wordpress.com`

[64]`http://www.educared.org`

[65]`http://www.rincondelvago.com`

study of plagiarism. It includes projects, centres for plagiarism research, specialised journals and conferences, overviews of field surveys, Websites about plagiarism and tutorials for prevention. See Lukashenko, Graudina, and Grundspenkis (2007) for further comparisons of commercial systems for plagiarism detection.

2.8 Chapter Summary

In this chapter the concepts of text-reuse and plagiarism are introduced. The particularities that make plagiarism one of the most famous kinds of the text re-use are discussed. Special attention is paid to the mechanisms behind the text re-use process, particularly from a paraphrases and cross-language point of view, and the circumstances that can become a case of re-use into plagiarism.

To better understand plagiarism and, in particular the huge amount of cases we are witnessing nowadays, a brief history of plagiarism is reviewed. The history goes from the emergence of the term and its development through the centuries, up to its formalisation as a fault. Afterwards we focussed on the academic environment to analyse why plagiarism is committed and how to deter it. Plagiarism detection is just a piece in the process of deterring plagiarism. Plagiarism prevention, surveillance, and, more briefly, reaction were discussed as complementary countermeasures.

As plagiarism is not limited to occur in academia, we looked at it as a problem in Forensic Linguistics. After a brief introduction to this topic, relationships between Computational and Forensic Linguistics were stressed when approaching common problems of plagiarism detection and authorship attribution. We were able to observe that forensic linguists are looking eagerly to get better tools for detecting cases of plagiarism, even after paraphrasing and translation.[66] In order to figure out how serious the problem of plagiarism is nowadays, an overview was offered with recent cases of plagiarism. Cases in academia, research, journalism, literature, politics and show business were reviewed. In cases where the evidence texts were available, a few automatic analyses were carried out in order to show what we can expect from some of the automatic models described in the chapters to come.

Heading back to academia, an analysis of surveys on academic cheating was included, paying special attention to paraphrase and translation in plagiarism. We ran a new survey on plagiarism attitudes in order to especially investigate the cross-language plagiarism behaviour. The results showed that the cross-language practice is almost as common as its monolingual counterpart.

Following, the necessity of developing automatic models for the detection of text reuse and plagiarism was stressed. The problem of text re-use detection was analysed from three intersecting points of view: natural language processing, information retrieval, and forensic linguistics. Finishing the chapter, an overview of available commercial plagiarism detection tools was given. Their advantages and disadvantages were discussed, paying special attention to the kinds of re-use they still not uncover so well: paraphrase and cross-language cases.

[66] `http://bit.ly/pan_11_turell-coulthard`

Chapter 3

Text Analysis for Text Re-use and Plagiarism Detection

> Similarity of words is the easiest net in which to catch the plagiarist.
>
> Philip Wittenberg

When looking for a potential case of text re-use, digging into a set of texts is necessary, looking for some element that triggers suspicion. This task requires a combination of NLP and IR techniques.In this chapter we describe different models to represent texts for analysis and comparison. Section 3.1 describes different models for text representation. The terms representing a document may be more or less relevant or descriptive, depending on different factors. An overview of relevance estimation models is presented in Section 3.2. Models for measuring similarity between texts are described in Section 3.3 and measures for analysing style and complexity within a document are discussed in Section 3.4. Experienced readers on these topics can skip this chapter.

Key contribution A novel model for assessing cross-language similarity between texts is introduced (Section 3.3.2.2).

3.1 Text Representation

A key factor when processing a document is how to represent it; what its relevant features are.[1] As Grefenstette and Tapanainen (1994) mention, the input text for a number of applications is simply a sequence of characters from which words (and sentences) must be identified. In agreement with Jurafsky and Martin (2000, p. 647), we consider a term t to be a lexical item that occurs in a document or collection, "but may also include phrases". Therefore, t could be either a string of characters, a word, a sequence of

[1] We address the problem of representing documents in terms of their contents: their text. Other methods, non considered, perform analyses on the basis of structural features (Buttler, 2004) (e.g. Si, Leong, and Lau (1997) base the comparison of two documents on their LaTeX structural features).

words, a sentence, etc. Such units are the representatives of a document d, the units to analyse and/or compare in order to determine the similarity among a set of documents; to determine whether a text fragment has been generated by re-use from another one. In this section we describe the most common pre-processing techniques applied when looking for re-used text and describe the most common text representation models for a given (fragment of a) document.

3.1.1 Pre-Processing

Many existing pre-processing techniques may be used before actually looking for cases of text re-use. Following, we enumerate some of the most well-known.

3.1.1.1 Character Normalisation

At character level, a few operations can be performed aiming at discarding irrelevant features of a text. The fist aspect to consider is capitalisation. *Case folding* is a common operation in IR for text comparison. Failing to determine that `Example`, `EXAMPLE`, and `example` are the same word because of capitalisation is unacceptable. A risk exists though: considering two different words as the same; e.g. `Valencia` (the place) and `valencia` (the noun)[2]. This risk exists with other pre-processing operations such as stemming, later described. Nevertheless, as analysing a document for re-use implies considering words within their context, we can safely normalise the vocabulary by case folding.

Another character level normalisation operation is *diacritics* elimination. This operation is even more relevant when dealing with cross-language text re-use detection (for instance by considering texts in English and French). As discussed in Chapter 6, some cross-language models exist based on direct comparison of text strings, such as *cross-language character n-grams* and *cognateness*. The vocabularies of different languages, in particular words roots, may be shared across them, but with slight changes, such as diacritics. Eliminating them makes sense just in the same way as case folding does.

3.1.1.2 Tokenisation

In agreement with Grefenstette and Tapanainen (1994), we consider *tokens* to be numbers, punctuations, dates an words (Grefenstette and Tapanainen identify them as "units which will undergo a morphological analysis"). Tokenisation consists of splitting a text into these tokens. The biggest difficulty of this task is considering what punctuation marks should be considered as part of a word and which not. For instance, consider the previous string "not." with respect to "Mr.".[3]

[2]Paraphrasing Jurafsky and Martin (2000, p. 195), it is not always clear whether the Spanish words *Valencia* and *valencia* should be treated as the same. *Valencia*, if not at the beginning of a sentence, may well refer to the city, whereas *valencia* may refer to the chemical concept (both are `[V|v]alence` in English).

[3]In this research we work with languages for which spaces are used to separate words. More complicated languages in this aspect, such as Chinese, require other considerations.

3.1.1.3 Stemming and Lemmatisation

For many tasks, most of them related to IR, considering a reduced version of a word, either in the form of a *stem* or a *lemma*, is better than considering its entire form. The reason is that IR systems use to be based upon morphological information, for which suffixes are irrelevant (Jurafsky and Martin, 2000, p. 83).[4] According to Baeza-Yates and Ribeiro-Neto (1999, p. 168), a stem is "the portion of a word which is left after the removal of its affixes. Therefore, *stemming* is the process that "strips off affixes and leaves a stem", removing the inflectional endings from words (Manning and Schütze, 2002, p. 132, 194)[5]. The stem of two different words could be the same, though, inserting noise to the characterisation.

When *lemmatising*, the problem is finding the lemma of an inflected word form (e.g. `went`→ `go`). As a result, lemmatisation is a more complex process, as disambiguation is required for determining what the actual lemma of a word is. For simplicity, stemming is often preferred over lemmatisation.

Different researches have analysed the impact of stemming in IR@. For instance, Krovetz (1993)[6] found that stemming improves the results when handling small documents only. The reason is simple: the larger a document, the more likely it will contain a *wordform*. Baeza-Yates and Ribeiro-Neto (1999, p. 168) agrees with the controversy, as different studies "lead to rather conflicting conclusions". As a result, some IR engines ignore any stemming process.[7]

3.1.1.4 Sentence Identification

A question causing strong discussion when dealing with text re-use and, in particular, plagiarism, is how long a borrowed text should be to actually be considered as plagiarised. In order to take advantage of documents' structure, we consider the sentence as a relevant unit for plagiarism detection. Therefore, sentence identifiers are often required as one of the stages of the detection process.

3.1.1.5 Punctuation Removal

A common pre-processing operation in IR and NLP is punctuation removal. However, a few tasks exist where punctuation is relevant; one of which is precisely related to text re-use detection. Punctuation could be used as a feature for author-identification, when looking for patterns that identify an author's style (Jurafsky and Martin, 2000, p. 194)

[4] In other tasks, such as *part-of-speech tagging* or translation, the *wordform* —the inflected form of a word (Jurafsky and Martin, 2000, p. 195)— is relevant for the analysis.

[5] According to Jurafsky and Martin (2000, p. 59), the stem is the "main morpheme of the word, supplying the main meaning".

[6] As seen in (Jurafsky and Martin, 2000, p. 83).

[7] In Chapter 5 models that either apply stemming or not will be discussed. This is an interesting parameter to consider when dealing with plagiarism detection. It has been particularly interesting when applied —or not— at the different editions of the International Competition on Plagiarism Detection (cf. Section 7).

(cf. the "common punctuation" discriminator in page 110).

3.1.1.6 Words Filtering

A *stop list* is "a list of high frequency words that are eliminated from the representation of documents" (Jurafsky and Martin, 2000, p. 655). The words in this kind of list, also known as *stopwords*, include *function words* (grammatical words such as *of*, *it*, *and* (Jurafsky and Martin, 2000, p. 289)), and in general, *closed-class terms* (additionally to function words, they include prepositions, articles, etc.).

In IR, the reason behind stopword removal is twofold: (*i*) they have little semantic weight, causing them to be practically useless for the task (Jurafsky and Martin, 2000, p. 655); and (*ii*) they let for saving considerable space when representing documents, as the most frequent words are stopwords, hence letting for discarding nearly half of the words in a document (Baeza-Yates and Ribeiro-Neto, 1999, p. 146), even reducing its size to up to one third (Hariharan and Srinivasan, 2008).

3.1.2 Bag of Words Representation

After pre-processing, the documents' contents have to be characterised. The bag of words (BoW) is probably one of the most frequently used representation models in IR. It assumes that a document is a bag containing all of the words in it. As a result, "the presence of a word is independent of another" (Manning and Schütze, 2002, p. 237). The order of the words, how they compose the phrases and sentences, is completely neglected, and no syntactic information is considered (Jurafsky and Martin, 2000, p. 647). In the BoW model, a document is represented by a tokens vector, either Boolean or weighted. Despite its simplicity, this characterisation offers good results.

In the detection of text re-use, the BoW model seems reasonable. However, when dealing with re-use, the syntactic information does matter. Remind we are looking for text fragments borrowed from an external source, and not only topic-level similarity (as in general IR). As a result, higher levels of representation are necessary, such as the word n-grams

3.1.3 n-Grams

n-grams can be composed of characters, words, phonemes, etc. An n-gram is a sequence of overlapping units of length n over a given sample. The overlapping may be defined from 1 (i.e., the last element in an n-gram is the first one in the following one) up to $n-1$ (i.e., the last $n-1$ elements in an n-gram are the first in the following one). The $n-1$ overlapping is the most frequently used.

Low level n-grams, $n = \{1, 2, 3\}$ are often known as *unigram*, *bigram*, and *trigram*, respectively. For $n \geq 4$ their are known as *four-gram*, *five-gram*, etc. (Manning and Schütze, 2002, p. 193). In different prediction tasks (for instance optical character or speech recognition) n-gram models usually consider $n = \{2, 3, 4\}$. The word n-gram

n	**word n-grams**						
1	I	picked	something	up	like	an	ornament
2	I picked up like		picked something like an		something up an ornament		
3	I picked something up like an		picked something up like an ornament		something up like		
4	I picked something up something up like an			picked something up like up like an ornament			
5	I picked something up like something up like an ornament				picked something up like an		
6	I picked something up like an				picked something up like an ornament		
7	I picked something up like an ornament						

Figure 3.1: Word n-grams example. We consider the word n-grams for $n = [1, \ldots, 7]$ of phrase Forensic$_2$ from page 22: "I picked something up like an ornament".

n	**character n-grams**									n	**character n-grams**						
1	o	r	n	a	m	e	n	t		2	or	rn	na	am	me	en	nt
3	orn		rna		nam		ame		men	4	orna		rnam		name		amen
5	ornam			rname			namen		ament	6	orname			rnamen			nament
7	ornamen				rnament					8	ornament						

Figure 3.2: Character n-grams example. We consider the character n-grams for $n = [1, \ldots, 9]$ of the word "ornament".

model is well known for performing word prediction: using "the previous $n - 1$ words to predict the next one" (Jurafsky and Martin, 2000, p. 194). In theory, the higher the n, the better the obtained results. However, in practice high values of n are hard to consider due to the lack of linguistic resources.

When analysing text re-use, text n-grams can be considered at word or character, and even at POS, level. In the case of word n-grams, the units are the tokens in the text (optionally considering punctuation marks). An example of word n-grams is included in Fig. 3.1. In the case of character n-grams, the characters are considered the units to combine (optionally considering spaces and punctuation marks). An example of character n-grams is included in Fig. 3.2. POS n-grams are just as word n-grams, but the morphosyntactic categories are used rather than the actual words. An example of POS n-grams is shown in Fig. 3.3.

In the chapters to come, we discuss what levels of n better fit for text re-use analyses, but we can anticipate some of the discussion. The most simple case, considering $n = 1$, is equivalent to the BoW model (indeed, the BoW model is just an instance of the n-gram model with $n = 1$). It shows good results for topic similarity estimation, but not so for text re-use detection. On the one side, a low level word n-gram, for instance $n = 2$, causes the comparison to be highly flexible and some syntactic information is already

n	**POS n-grams**
1	PP VBD NN RP IN DT NN
2	PP VBD VBD NN NN RP RP IN IN DT DT NN
3	PP VBD NN VBD NN RP NN RP IN RP IN DT IN DT NN
4	PP VBD NN RP VBD NN RP IN NN RP IN DT RP IN DT NN
5	PP VBD NN RP IN VBD NN RP IN DT NN RP IN DT NN
6	PP VBD NN RP IN DT VBD NN RP IN DT NN
7	PP VBD NN RP IN DT NN

Figure 3.3: POS n-grams example. We consider the POS n-grams for $n = [1, \ldots, 7]$ of the part-of-speech of phrase Forensic_2 from page 22: "PP VBD NN RP IN DT NN".

captured. As a result, it is possible to detect cases of re-use with high rewriting levels, but at the cost of retrieving too many false negatives. On the other side, high levels, for instance $n = 7$, are too strict, making them ideal for detecting cases of verbatim copy. The cost is that they are very sensitive, missing cases with very slight modifications (cf. Fig. 2.2 at page 22 for a graphical representation sustaining this idea).

Some of the successful approaches to text re-use detection based upon word level n-grams are discussed in Section 5.1.2. Approaches based on character level n-grams, in particular for cross-language detection, are discussed in Section 6.2.2.1. The exploitation of POS n-grams, useful in intrinsic plagiarism detection, is discussed in Section 5.1.1.

3.1.4 Cognates

Cognates are defined as "words that are similar across languages" (Manning and Schütze, 2002, p. 475). The models based upon this kind of representation (including character n-grams), can be used for monolingual comparison, but are particularly useful in cross-language settings. The reason is that they take advantage of syntactic similarities between languages. Such similarities are particularly significant between languages that belong to common families (such as English-French or Italian-Spanish) or, to a less extent, with strong influence with respect to each other (such as English-German or Spanish-Basque) (Mcnamee and Mayfield, 2004; Simard, Foster, and Isabelle, 1992).

A way of representing a text by a collection of good potential cognates is through *cognateness*. This concept was created by Simard *et al.* (1992) in order to identify parallel sentences. Under this model, a word w is a candidate to share a cognateness relationship if:

(a) w contains at least one digit,
(b) w is exclusively composed of letters and $|w| \geq 4$, or
(c) w is a single punctuation mark.

Words w and w' are pseudo-cognates if both belong to (a) or (c) and are identical, or belong to (b) and share exactly the same four first characters. A document can be easily

characterised for comparison in agreement with this model: if w accomplishes (a) or (c), it is maintained verbatim, if it accomplishes (b) it is cut down to its first four characters.

In cross-language settings this kind of models may be erroneously considered weak due to the assumption that languages considered have to be related, but this is far to be truth. As Church (1993, p. 3) mentions, this model can be exploited when handling languages with many cognates and "almost any language using the Roman alphabet, since there are usually many proper names and numbers present". As seen in Chapter 9, it can even work between languages in Roman and Cyrillic alphabet, as far as one of them is transliterated.

3.1.5 Hash Model

Comparison between strings is computationally and spatially expensive. As a result, models have been designed to represent text contents that require low amounts of space and let for an efficient comparison. This is precisely the case of the family of hash models. The purpose of a *hash function* $\mathcal{H}$ is mapping a string (for instance, a word, a sentence, or an entire document), into a numerical value. The resulting numbers can be saved into the so called *hash table*, which allows for efficiently "see whether a [text string] has been seen before" (Manning and Schütze, 2002, p. 122).

Indeed, the aforementioned representations can be plugged into a hash model in order to speed up the comparison process among documents; n-grams, sentences or fixed length text fragments from a document collection D can be hashed and inserted into a hash table. When analysing a suspicious document d_q, it can be hashed by means of the same function and queried against the table. If a match occurs, two exact text fragments have been found (cf. Section 5.1.2.1).

When looking for text re-use cases, the main advantages of the hash functions are the following: (a) the resulting hash value is a compact representation of the input string, saving space; and (b) *collisions* are extremely unlikely.[8] As the probability of such event is extremely low in different hashing models, hash values can be confidently exploited to represent documents, aiming at comparing them.

Examples of hash function are the well known *md5sum* (Rivest, 1992) and *randomised Rabin-Karp* (Rabin, 1981). The md5sum is a hexadecimal function that results in long hash values, which could not be the best option when aiming at hashing words or short n-grams, but can be used when searching for exact duplicates (indeed, md5sum is a popular mechanism to ensure that a file has not changed). Karp-Rabin is a more compact, decimal model, more oriented to text strings, which still guarantees a low collision probability. As a result, it is popular for documents processing techniques.

An example of the results of applying md5sum and Karp-Rabin is shown in Fig. 3.4. It is worth noting the difference between the values obtained for the two analysed text fragments with the two models. Whereas the difference between f' and f'' is just the last character ($i \rightarrow 1$), the resulting hashes are significantly different. As observed in Section 5.1.2, this is one of the reasons why this kind of representation is suitable only

[8] In this case, a collision implies obtaining the same hash value from two different text strings.

f	=	'Star Wars is an epic space opera franchise initially concei'
f'	=	'starwarsisanepicspaceoperafranchiseinitiallyconcei'
f''	=	'starwarsisanepicspaceoperafranchiseinitiallyconce1'
$md5sum(f')$	=	1228fcad06a3dfb706074abc32596eb2
$md5sum(f'')$	=	11628ef8a6d263f8438c38b889796319
$karp\text{-}rabin(f')$	=	-9156124894797263357
$karp\text{-}rabin(f'')$	=	-9156124894797263269

Figure 3.4: Hashing functions example. Pre-processing operations: case folding, diacritics elimination, and space normalisation, generating f'. Both *md5sum* and *randomised Karp-Rabin* functions are applied to f' and f''. The used md5sum comes from the GNU coreutils 8.9. The Karp-Rabin implementation was kindly provided by the Webis group of the Bauhaus-Universität Weimar. String borrowed from the leading 60 characters in the Wikipedia article about Star Wars by mid-2010.

if the purpose is looking for exact copies.

3.2 Weighting

As described in Section 3.3.1, characterising a document for comparison in general implies representing d on the basis of its contents. This implies, in most cases, coming up with a vector representation where each dimension represents an element in the document. The relevance of the dimension, a term, has to be computed beforehand in most cases. This relevance value is the known as weight of the term.

In this section we discuss a few weighting models that can be considered when assessing similarity between documents. The kinds of weighting models are strongly related to the similarity models considered in Section 3.3.

3.2.1 Boolean Weighting

This is the most simple of the weighting models. The weight of the i-th term in document d is defined as:

$$w_{i,d} = \begin{cases} 1 \text{ if } t_i \text{ exists in } d \\ 0 \text{ otherwise} \end{cases} \tag{3.1}$$

Evidently, this model makes a naïve assumption: that every term in a document has the exact same relevance. At first sight such assumption might not be the best idea. However, when considering text representations resulting in a flat distribution (i.e., most of the terms occur only once in a document) it becomes sensitive. Consider for instance representations based on text sentences or high level word n-grams.[9]

[9]Refer once again to the example presented in Fig. 2.2, at page 22.

3.2.2 Real Valued Weighting

When opting for a more varied distribution, Boolean weighting is no longer an option. In real valued weighting the relevance of a term in a document takes a real value between 0 and 1. According to Manning and Schütze (2002, p. 542), this schema can be based upon three basic notions: *term frequency*, *document frequency*, and *collection frequency*. Here we include the first two only. The third one is related to the occurrences of a token in a given collection of documents, and is not so useful from a text re-use and plagiarism detection perspective.

3.2.2.1 Term Frequency

The basic assumption in this model is that the more frequently a term appears in a document, the more relevant it is. *Term frequency* (tf) is "the raw frequency of a term within a document" (Jurafsky and Martin, 2000, p. 651). The tf of the i-th term in document d is defined as

$$tf_{i,d} = \frac{n_{i,d}}{\sum_k n_{k,d}} \ , \tag{3.2}$$

where $n_{i,j}$ is the frequency of t_i in d. The normalisation is based on the overall frequency of all the terms $t \in d$. As Jurafsky and Martin (2000, p. 200) point out, the use of relative frequencies as a way to estimate probabilities is one example of the known as *Maximum Likelihood Estimation* (MLE), because the resulting parameter set is one in which the likelihood of the training set T given the model M, i.e., $P(T \mid M)$, is maximised.

As aforementioned, the most frequent terms in a document are closed-class terms (cf. page 56). If considering the BOW model, the most relevant terms in a document would be no other than prepositions and articles. In order to fix this issue, a stop list can be used to discard those terms that are known to be irrelevant beforehand.

Another option to weight terms considering a document in isolation is the known as *transition point*. The idea is that those terms appearing more than once, but not extremely often (such as prepositions), are the most relevant (Pinto, Jiménez-Salazar, and Rosso, 2006; Pinto, Rosso, and Jiménez-Salazar, 2011). The transition point tp^* is defined as:

$$tp^* = \frac{\sqrt{8 \cdot I_1 + 1} - 1}{2} \ , \tag{3.3}$$

where I_1 represents the number of *hapax legomena* in d. Aiming at assigning higher relevance to those terms around tp^*, the term weights are calculated as:

$$tp_{i,d} = (\langle tp^* - f(t_i, d) \rangle + 1)^{-1} \ , \tag{3.4}$$

where, $f(t_i, d)$ represents the absolute frequency of $t_i \in d$. In order to guarantee positive values, $\langle \cdot \rangle$ represents the absolute value function. This approach, as many other document-level weighting schemas, seems to be rooted in the ideas of Luhn (1958) for measuring words significance.

If the required process implies handling an entire collection of documents D, document frequency can be considered for a better relevance estimation.

3.2.2.2 Document Frequency

The main idea behind this model is that the fewer documents a term appears in, the more relevant it is for them (Jurafsky and Martin, 2000, p. 653). In order to estimate such relevance, *inverse document frequency* (*idf*) (Spärck Jones, 1972)[10] could be used. It is computed as:

$$idf_i = \log \frac{|D|}{|D_i|} \ , \tag{3.5}$$

where $|D|$ stands for the size of the documents collection and $|D_i|$ is the number of documents in D containing t_i. The obtained value behaves precisely as required: it is higher if less documents contain t_i. *idf* is defined as a measure of how rare a term in a collection of documents is.[11] By combining Eqs. (3.2) and (3.5) we obtain one of the best known real valued weighting models, *tf-idf*:

$$tf{\cdot}idf_{i,d} = tf_{i,d} \cdot idf_i = \frac{n_{i,d}}{\sum_k n_{k,d}} \cdot \log \frac{|D|}{|D_i|} \ . \tag{3.6}$$

3.3 Text Similarity

Measuring how similar two pieces of text are is relevant for other tasks than detection of text re-use and plagiarism. Plenty of other tasks take advantage of these techniques as well, including documents clustering and categorisation (Bigi, 2003), multi-document summarisation (Goldstein *et al.*, 2000), and version control (Hoad and Zobel, 2003). For the particular case of text re-use analysis, measuring text similarity is useful for the detection of co-derivatives, text re-use, and plagiarism detection (Maurer *et al.*, 2006), and information flow tracking (Metzler, Bernstein, Croft, Moffat, and Zobel, 2005), among others.

The similarity between two documents can be measured from three different perspectives, namely (Hariharan and Srinivasan, 2008):

Size and structure. Documents are considered to be similar on the basis of their length (in paragraphs, sentences, words, or characters), their sectioning, format, or distribution.

[10] As seen in (Jurafsky and Martin, 2000, p. 653).

[11] This sentence has been borrowed from (Hariharan and Srinivasan, 2008), but the definition of *idf* is well-known, part of the general knowledge in NLP and IR. Is a reference to Hariharan and Srinivasan (2008) necessary? This is an illustrative example that depicts how subjective and fuzzy the line between plagiarism and fair text re-use is.

Contents. Two documents are considered as similar as the amount of common terms they contain.[12]

Style. Similarity between documents is estimated on the basis of stylistic features, such as the grammatical person used, the complexity of their contents, etc.[13]

In order to analyse a document for plagiarism, the most relevant similarity measures are those based on contents and style. Refer to Buttler (2004) for an overview of models for structural similarity measurement.

The most of the models for similarity measurement are applied to the entire document's content (often sub-sampling). However, some scholars perform a selective comparison, by considering key sections of a document, such as the title, abstract, and index terms (Bani-Ahmad, Cakmak, Ozsoyoglu, and Hamdani, 2005)[14]. Obviously, these sections are only available in some kinds of documents, such as scientific papers.

Two characteristics are considered beneficial for a similarity measure: (*i*) its range is well defined, i.e., $sim(d, d_q) \in= [0, 1]$ (1 means that the texts are identical, 0 that they are not similar at all); and (*ii*) it is symmetrical, i.e., $sim(d_q, d) = sim(d, d_q)$ (Markov and Larose, 2007, p. 39). Nevertheless, there are similarity measures without defined range (cf. Section 3.3.2.2) or asymmetrical (cf. Sections 3.3.1.1 and 3.3.2.1), but still useful.

In this section we describe two kinds of text similarity models: standard Boolean and real weighted (Section 3.3.1) and probabilistic, including information theory and machine translation-based models (Section 3.3.2). We selected these two families in order to separate the different characterisation and comparison philosophies. However, no strict border among them exists. Indeed, most of them can be considered an extension or be directly mapped to the vector space model. Moreover, many of them can be combined to come up with a more robust similarity estimation model.

3.3.1 Vector Space Models

In the *vector space model* (VSM), both documents and queries[15] are represented as feature vectors (Jurafsky and Martin, 2000, p. 647) (indeed, as observed by Maurer *et al.* (2006), this approach was borrowed from pattern recognition). The considered features are the terms within the documents collection; either words, character n-grams, word n-grams or more complex representations (cf. Section 3.1).

Two kinds of vector representation schemas exist: *Boolean* (also known as binary), in which the existence/non-existence of a term is indicated with a value in $\{0, 1\}$ and

[12] Hariharan and Srinivasan (2008) differentiate between contents and vocabulary similarity. However, we consider them as a single one.

[13] Again, Hariharan and Srinivasan (2008) consider that the average number of characters per word is a structural feature. We consider that, as it will be described in Section 3.4, it fits better as a stylistic one.

[14] As originally seen at (Hariharan and Srinivasan, 2008).

[15] For the case of text re-use detection, the queries are indeed other documents or, at least text fragments.

real valued in which every term is weighted with a value in the range $[0, 1]$. In agreement with the notation used in Jurafsky and Martin (2000) and Manning and Schütze (2002), we represent the characteristic vector of a document d as:

$$\vec{d_k} = (t_{1,k}, t_{2,k}, t_{3,k}, \cdots, t_{N,k},)$$

where $t_{1,k}$ represents the weight of the i-th term in document k.

3.3.1.1 Boolean Models

As seen in Section 3.2.1, the only relevance factor in Boolean models is the presence or absence of a term. The following is a simple similarity metric for Boolean representations (Jurafsky and Martin, 2000, p. 647–648):

$$sim(\vec{d}, \vec{d_q}) = \sum_{i=1}^{N} t_{i,d} \cdot t_{i,d_q} \ , \tag{3.7}$$

i.e., the similarity between d_q and d is computed by considering the number of terms they share. Indeed, this is a simple set operation. Note that $t_{i,d_q} \cdot t_{i,d} = 1$ iff $t_{i,d_q} = t_{i,d}$. Therefore, Eq. (3.7) can be rewritten as:

$$sim(d, d_q) = |d \cap d_q| \ , \tag{3.8}$$

also known as the *matching coefficient* (Manning and Schütze, 2002, p. 299). Nevertheless, the resulting similarity value is not ranged. More important: long texts will be, in general, more similar than shorts, simply because it is more likely that they share more contents. Therefore, more robust length-independent measures, are necessary.

Jaccard coefficient

This measure was originally proposed for the analysis of flora in the Alps (Jaccard, 1901). The Jaccard coefficient is defined as:

$$sim(\vec{d}, \vec{d_q}) = \frac{\left|\{j | d^j = 1 \wedge d_q^j = 1\}\right|}{\left|\{j | d^j = 1 \vee d_q^j = 1\}\right|} \ , \tag{3.9}$$

i.e., "the proportion of coordinates that are 1 in both $\vec{d}$ and $\vec{d_q}$ to those that are 1 in $\vec{d}$ or $\vec{d_q}$" (Markov and Larose, 2007, p. 39). The Jaccard coefficient can be safely defined in terms of sets operations as:

$$sim(d, d_q) = \frac{|T(d) \cap T(d_q)|}{|T(d) \cup T(d_q)|} \ , \tag{3.10}$$

where $T(d)$ is the set of terms occurring in d; i.e., the intersection between the vocabularies in d and d_q is normalised by their union; the amount of shared terms between d and d_q with respect to the number of terms in the entire vocabulary.[16]

[16]The rest of Boolean models are expressed in terms of sets, rather than vectors for simplicity.

Albeit its simplicity, the quality of the obtained results by means of this measure is high and represents one of the most widely used Boolean models in IR. The Jaccard Coefficient is also known as *Tanimoto coefficient* (Manning and Schütze, 2002) and, for the specific case of text re-use analysis, *resemblance* (Broder, 1997).

Dice's coefficient

This is another measure originally proposed for ecological studies (Dice, 1945). Once again, the similarity is normalised by means of the entire vocabularies, i.e., the total number of non-zero entries in each document, but this time independently. It is computed as:

$$sim(d, d_q) = \frac{2\,|T(d) \cap T(d_q)|}{|T(d)| + |T(d_q)|} \ , \tag{3.11}$$

where the factor 2 allows obtaining a similarity ranged between 0 and 1 (Manning and Schütze, 2002).

Overlap coefficient

In the overlap coefficient $sim(d_q, d) = 1$ if $T(d_q) \subseteq T(d)$ or $T(d) \subseteq T(d_q)$; i.e., every term in the smaller set exists in the bigger one. It is computed as:

$$sim(d, d_q) = \frac{|T(d) \cap T(d_q)|}{\min\left(|T(d)|\,, |T(d_q)|\right)} \ . \tag{3.12}$$

Containment

It measures the number of matches between the two term sets and scales them by the size of only one of them (Broder, 1997). It is calculated as:

$$sim(d, d_q) = \frac{|T(d) \cap T(d_q)|}{|T(d)|} \ . \tag{3.13}$$

Whereas this measure is still ranged in $[0, 1]$, it is not symmetric; i.e., in general, $sim(d, d_q) \neq sim(d_q, d)$. This measure makes sense under two circumstances: (i) $|d_q|$ is very different to d (for instance, d_q is a sentence and d is an entire document); and (ii) d_q has to be compared to many documents $d \in D$. As $|T(d_q)|$ is constant, the obtained results are not affected by discarding it.

Boolean cosine measure

This measure tries to decrease the impact of non-zero dimensions; i.e., those cases where $|t(d_q) = 0|$ is very different to $|T(d) = 0|$. The idea is similar to that of the containment measure: to let for the comparison of objects of significantly different size. It is computed as:

$$sim(d, d_q) = \frac{|T(d) \cap T(d_q)|}{\sqrt{|T(d_q)| \times |T(d)|}} \ . \tag{3.14}$$

3.3.1.2 Real-Valued Models

In order to indicate the importance of a term in a text, weights can be assigned to the dimensions rather than binary values. By considering weights, the text representations stop being sets to become real vectors. As a result, a collection of documents can be viewed as a multi-dimensional space. As described by Jurafsky and Martin (2000, p. 648) the vector that represents the query is just another point in that space and the most related —relevant— documents will be those located closer.

Dot product

The dot product consists of multiplying every dimension of the two vectors and summing the resulting products:

$$sim(\vec{d}, \vec{d_q}) = \sum_{i=1}^{N} t_{i,d} \cdot t_{i,d_q} \ . \tag{3.15}$$

Note that this is the similarity model proposed in Eq. (3.7), but considering real values. As in that case, the dot product is useless for similarity estimation because of its sensitiveness to dimensions magnitudes. Normalisation is necessary.

Weighted cosine measure

In order to decrease the impact of the vector's length in the similarity calculation, the vector dimension's weight can be normalised. This is made by means of the entire vector length, defined as

$$|\vec{d}| = \sqrt{\sum_{i=1}^{N} t_{i,d}} \ , \tag{3.16}$$

where $t_{i,d}$ represents the weight of the i-th term in document d.

The principle of the cosine measure is calculating the inner product between two normalised feature vectors (Hariharan and Srinivasan, 2008). The resulting value is the cosine of the angle between them (Baeza-Yates and Ribeiro-Neto, 1999, p. 27). By combining Eqs. (3.15) and (3.16), we obtain the weighted cosine measure between two documents:

$$sim(\vec{d}, \vec{d_q}) = \frac{\sum_{i=1}^{N} t_{i,d} \times t_{i,d_q}}{\sqrt{\sum_{i=1}^{N} t_{i,d}^2} \times \sqrt{\sum_{i=1}^{N} t_{i,d_q}^2}} \ . \tag{3.17}$$

According to Baeza-Yates and Ribeiro-Neto (1999, p. 27), this measure represents the correlation between the vectors.

Word chunking overlap

This may be considered a "classic" model for copy-detection (Shivakumar and García-Molina, 1995). It is based upon the so called *asymmetric subset measure* for document pairs. In this case $\vec{d}$ is not composed of the entire set of terms in d, but of a subset only. Such a subset is composed of terms with similar frequency in d and d_q. The similarity (overlapping in terms of Shivakumar and García-Molina (1995)) between d and d_q is defined as:

$$subset(d, d_q) = \frac{\sum_{t \in c(d,d_q)} w_{t,d_q} \cdot w_{t,d}}{\sum_{t \in d_q} w_{t,d_q}^2} \ , \tag{3.18}$$

where w_{t,d_q} is the weight of term t in d_q, in this case represented by the term frequency (tf); $c(d, d_q)$ is the *closeness set*, containing those terms $t \in d \cap d_q$ such that $w_{t,d} \sim w_{t,d_q}$. A term t belongs to the closeness set iff:

$$\epsilon - \left(\frac{tf_{t,d}}{tf_{t,d'}} + \frac{tf_{t,d'}}{tf_{t,d}} \right) > 0 \ . \tag{3.19}$$

The parameter ϵ defines how close the frequency of t in both documents must be in order to be included in the closeness set. The value used by Shivakumar and García-Molina (1995) when comparing *netnews* articles was $\epsilon = 2.5$, because it offered a good balance between precision and recall.[17]

The word chunking overlap is another non symmetric measure. If necessary, a symmetric similarity value between d and d_q can be obtained as:

$$sim'(d, d_q) = max \left\{ subset(d_q, d), subset(d, d_q) \right\} \ . \tag{3.20}$$

Moreover, obtaining a value $sim'(d, d_q) > 1$ is possible. If this measure is considered as part of a ranking process, where d_q is compared to many documents $d \in D$, it can be normalised in order to fit a similarity range $[0, 1]$:

$$sim(d, d_q) = \frac{sim'(d, d_q)}{\max_{d' \in D} sim'(d', d_q)} \ . \tag{3.21}$$

3.3.2 Probabilistic Models

As Manning and Schütze (2002, p. 303) point out, the Euclidean distance —the vector space models are sustained on— is "appropriate for normally distributed quantities, not for counts and probabilities". But the feature vectors of the cosine and other measures are precisely composed of (normalised) counts. By dividing the frequency of a term in a document $f(t) \in d$ by $\sum_i t_i \in d$, we obtain a conditional probability $p(t \mid d)$. By this mean, the feature vector of the VSM becomes a probability distribution, estimated by maximum likelihood (Manning and Schütze, 2002, p. 303). Here we consider two

[17] Cf. Section 4.3 for a definition of these evaluation measures.

probabilistic models: the Kullback-Leibler divergence (in fact, an adaptation that makes it a distance), and a machine translation-based model.

3.3.2.1 Kullback-Leibler Distance

Kullback and Leibler (1951) proposed the after known as *Kullback-Leibler divergence* (KL_d), also known as *relative entropy*. From an information theory point of view, it "[...] is the average number of bits that are wasted by encoding events from a distribution p with a code based on a 'not-quite-right' distribution q" (Manning and Schütze, 2002, p. 72). Indeed, KL_d estimates how different two probability mass functions $p(x)$, $q(x)$ over an event space are. Therefore, it can be considered as a pseudo-(dis)similarity measure. It is defined as

$$KL_d(p \mid\mid q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \ , \tag{3.22}$$

where $\mathcal{X}$ represents the event space and, being a difference measure, $KL_d(p \mid\mid q) = 0$ iff $p = q$. The Kullback-Leibler divergence is asymmetric, i.e., $KL_d(p \mid\mid q) \neq KL_d(q \mid\mid p)$. Therefore, different researchers have proposed symmetric versions, known as *Kullback-Leibler symmetric distance* (KL_δ). Among them, we can include:

$$KL_\delta(p \mid\mid q) = KL_d(p \mid\mid q) + KL_d(p \mid\mid q) \ , \tag{3.23}$$

$$KL_\delta(p \mid\mid q) = \frac{1}{2}\left[KL_d\left(p \mid\mid \frac{p+q}{2}\right) + KL_d\left(q \mid\mid \frac{p+q}{2}\right)\right] \ , \tag{3.24}$$

$$KL_\delta(p \mid\mid q) = \sum_{x \in \mathcal{X}} (p(x) - q(x)) \log \frac{p(x)}{q(x)} \ , \text{ and} \tag{3.25}$$

$$KL_\delta(p \mid\mid q) = \max(KL_d(p \mid\mid q), KL_d(q \mid\mid p)) \ , \tag{3.26}$$

corresponding to the versions of Kullback and Leibler (1951), Jensen (Fuglede and Topse, 2004), Bigi (2003), and Bennett, Li, Vitányi, and Zurek (1998) respectively. A comparison of these four versions concluded that there is no significant difference among them (Pinto *et al.*, 2007). As a result, we use Eq. (3.25). The reason is that it is an adaptation of Eq. (3.22), with an additional subtraction only (the other three versions perform a double calculation of KL_d, which is computationally more expensive).

KL_δ represents a difference rather than a similarity measure. Therefore, a value of $KL_\delta(p \mid\mid q) = 0$ implies that $p = q$, i.e., the probability distributions are identical. Assuming the probability distribution p represents d_q (p_{d_q}), q represents d (q_d), and d_q is going to be compared to many $d \in D$, a similarity value between p_{d_q} and q_d, for all $d \in D$ can be estimated as:

$$sim(d_q, d) = -\left(\frac{KL_\delta(P_{d_q} \mid\mid Q_d)}{max_{d'} KL(P_{d_q} \mid\mid Q_d)} - 1\right) \ . \tag{3.27}$$

3.3.2.2 Machine Translation

We focus on empiricist approaches to machine translation: *statistical machine translation* (SMT).[18] The task of SMT can be described as follows. Let L and L' be two languages. Given x, written in language L, find the most likely translation y in language L'; i.e., we want to estimate $p(y \mid x)$, the probability that a translator will produce y as the translation of x (Brown, Della Pietra, Della Pietra, and Mercer, 1993b, p. 264).[19]

On the basis of the Bayes' theorem, $p(y \mid x)$ can be defined as:

$$p(y \mid x) = \frac{p(y) \cdot p(x \mid y)}{p(x)} \ , \tag{3.28}$$

where $p(x)$ can be discarded as it is independent of y. As a result, the "fundamental equation of machine translation" (Brown *et al.*, 1993b, p. 265) is defined as:

$$\hat{y} = \operatorname*{argmax}_y p(y \mid x) = \operatorname*{argmax}_y p(x \mid y) \cdot p(y) \ , \tag{3.29}$$

where $p(y)$, the *a priori* probability, is known as *target language model probability* and $p(x \mid y)$, the conditional distribution, is known as *translation model probability*. Since translating d_q into L' is not our concern, but identifying texts written in L' which are possible translations of d_q, these models can be adapted as follows (Barrón-Cedeño *et al.*, 2008; Pinto *et al.*, 2009; Potthast *et al.*, 2011a):

1. The language model, whose purpose is to generate intelligible text in the output, is replaced by a *length model* $\varrho(d')$ that depends on text lengths instead of language structures, and
2. the adapted translation model is a non-probabilistic measure $w(d_q \mid d')$.

Based on these adaptations we define the following similarity measure:

$$\varphi(d_q, d') = p(d' \mid d_q) = \varrho(d') \ w(d_q \mid d') \ . \tag{3.30}$$

This is another non-ranged measure. However, the partial order induced among documents resembles that of other similarity measures. Following, we describe the length model $\varrho(d')$ and the adapted translation model $w(d_q \mid d')$.

Length model

Rather than considering $p(y)$ to be a language model, it is adapted into a length model $\varrho(d')$. Although it is uncommon to find a pair of translated sentences, texts, or documents

[18] Those depending on large bilingual corpora and bilingual dictionaries; based on "information gathered wholesale from data" (Brown et al., 1993a).

[19] Due to the nature of this model, it can be used for estimating similarity between texts in the same language, by assuming $L = L'$, or texts in different languages, by considering $L \neq L'$. Here we describe the general approach and go into specificities in those cases where it has been applied as a monolingual or cross-language similarity model.

d and d' such that $|d| = |d'|$, it is expect that their lengths will be closely related by a certain length factor for each language pair. In accordance with Pouliquen, Steinberger, and Ignat (2003), the length model probability is defined as:

$$\varrho(d') = e^{-0.5\left(\frac{\frac{|d'|}{|d_q|}-\mu}{\sigma}\right)^2}, \tag{3.31}$$

where μ and σ are the mean and standard deviation of the character lengths between translations of documents from L into L'. Therefore, this is in fact a normal distribution with a maximum mean value of 1. Note that in those cases where a potential translation d' of a document d_q has not the expected length, the similarity $\varphi(d_q, d')$ is reduced.

Length models estimations for different language pairs can be seen in Section 6.3. In those cases where $d_q \in L$ and $d \in L'$ are written in the same language (i.e., $L = L'$), the length model becomes constant and $\varrho(d') = 1$ (Barrón-Cedeño *et al.*, 2009a). Another option would be considering a paraphrases collection to estimate a "translation" distribution, but this kind of resource is not always available for every language.

The comparison of documents by considering the sentences' length only, is well-known since the 1990s for corpora alignment. Good results have been obtained by measuring lengths at character (Gale and Church, 1993) as well as word level (Brown, Lai, and Mercer, 1991). Nevertheless, these approaches assume that the corpus they are processing is parallel. This implies that the translation of a sentence in L will exist in L' within the corpus. However, when looking for re-used fragments this is far to be truth. The contents of two documents could be completely unrelated and only one single sentence be copied (after translation) into the other. As a result, in a sea of potentially unrelated text, considering more factors is necessary.

Translation model

The translation model is considered to be an "enormous table that associates a real number between zero and one with every possible pairing of a passage in L and a passage in L' (Brown *et al.*, 1993b, p. 264).

For the translation model probability $p(x \mid y)$, the likelihood of the translation (x, y), we can select one out of five translation models (Brown *et al.*, 1993b), the known as the IBM models family. Barrón-Cedeño *et al.* (2008) opted for using model 1 (IBM M1) because of its generality and simplicity. The reason behind this decision is that, given two strings d_q and d', it is assumed that all the connections for each word position are equally likely. As a result, the order of the words in the two texts does not affect $p(x \mid y)$ (Brown *et al.*, 1993b, p. 268). As a text can be translated into many different ways and "the choice among them is largely a matter of taste" (Brown *et al.*, 1993b, p. 264), we take advantage of this positional independence to come out with a general model that aims at estimating good translation probabilities with position independence.[20]

The translation model depends on a *statistical bilingual dictionary*. This dictionary is

[20]Models 2–5 consider the order of the words (Brown *et al.*, 1993b, p. 268) and, therefore, would result in a less general similarity estimation model.

estimated by means of the well-known IBM M1 alignment model (cf. Appendix A) (Brown *et al.*, 1993b; Och and Ney, 2003), which has been successfully applied in monolingual (Berger and Lafferty, 1999) and cross-language information retrieval tasks (Pinto *et al.*, 2007). In fact, adaptations of the M1 have been already applied to monolingual measures of similarity between sentences (Metzler *et al.*, 2005). In order to generate a bilingual dictionary, M1 requires a sentence-aligned parallel corpus.[21] Given the vocabularies of the corresponding languages $\mathcal{X} \in L$ and $\mathcal{Y} \in L'$, the bilingual dictionary contains the estimates of the translation probabilities $p(x, y)$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. This distribution expresses the probability for a word x to be a valid translation of a word y.

The translation probability of two texts d and d' is defined as:

$$p(d_q \mid d') = \prod_{x \in d_q} \sum_{y \in d'} p(x, y) \ , \tag{3.32}$$

where $p(x, y)$ is the bilingual dictionary. The probabilistic measure estimated by means of Eq. (3.32) makes sense when comparing texts at sentence level. However, when comparing entire documents, computing the product of a large amount of probabilities implies a high risk of *numerical underflow*. One possible solution would be performing the calculations in *log space* (Jurafsky and Martin, 2000, p. 200). Nevertheless, still under the log space, the estimated probabilities may be more similar than the similarity between the texts actually is. Barrón-Cedeño *et al.* (2008); Pinto *et al.* (2009) propose an adaptation, resulting in the following tailored version:

$$w(d_q \mid d') = \sum_{x \in d_q} \sum_{y \in d'} p(x, y) \ . \tag{3.33}$$

This measure is no longer probabilistic; i.e., obtaining $w(d \mid d') > 1.0$ is possible. The weight $w(d \mid d')$ increases if valid translations (x, y) appear in the implied vocabularies. For each word $x \in d_q \setminus d$, a penalisation ϵ is applied to $w(d_q \mid d)$. The penalisation value has been empirically defined as $\epsilon = -0.1$.

In Eq. (3.33) the longer the texts d and d' are, the more similar they could be. In order to avoid this behaviour, the mean over $x \in d$ could be considered rather than a sum. However, we are precisely interested in obtaining higher values for longer texts. The reason behind this interest is that having two long texts with a high translation similarity may correctly imply a higher likelihood of being re-used. In cases where the length of two sentences is particularly different, the adapted translation model could estimate they are more similar than they indeed are, though. It is in this cases where the length model's contribution is more relevant. If the difference between $|d|$ and $|d'|$ is higher than expected for a re-used pair, the translation similarity will be importantly downgraded by the length model.

When handling texts in the same language, the statistical dictionary may be considered "virtual". In a monolingual setting, $p(x, y) = 1$ iff $x = y$, 0 otherwise (Barrón-Cedeño *et al.*, 2009a). As a result, in this case the machine translation-based similarity

[21] The estimation is carried out on the basis of the EM algorithm (Baum, 1972; Dempster, Laird, and Rubin, 1977); cf. (Brown *et al.*, 1993b; Pinto *et al.*, 2009) for an explanation of the bilingual dictionary estimation process.

model can be applied by considering Eq. (3.33) only. Another option would be considering some semantic relationship to determine a "monolingual translation probability". This option is highly similar to the proposed by Alzahrani and Salim (2010), using Wordnet synsets (cf. Section 7.3).

As the obtained result may exceed the range $[0, 1]$, the same normalisation as for word chunking overlap may be applied (Eq. (3.21)).

3.4 Stylometric Measures

According to Mallon (2001, p. 108), originality lies in how the author "configures the infinitely varied molecular matter of the words", his stylistic trademarks (Mallon, 2001, p. 139). This "trademark" can be well applied in order to determine whether a document d_q contains fragments which are outliers of the rest of the document, something that may well imply such fragments have been borrowed. Indeed, many scholars point out writers have their own authors' fingerprint and, as mentioned already in Section 2.5, a change in style could be a trigger for plagiarism suspicion. As pointed out by Bull *et al.* (2001, p. 5) "the most common trigger that arouses academics' suspicions of plagiarism in assignments is a change of writing style within text and differences in syntactic structure and in the use of terminology".[22]

As a result, giving the expert evidence about abrupt changes in style, vocabulary and complexity are good indicators that a document deserves further analyses. Here we somehow follow the typology proposed by Meyer zu Eißen, Stein, and Kulig (2007, p. 361) to describe the different features used to reflect these changes.

3.4.1 Text Statistics

These computations are carried out at character level. The number of punctuation marks as well as word lengths are often considered (Meyer zu Eißen *et al.*, 2007). Another option is trying to characterise the vocabulary richness in a text. The diversity in the vocabulary of a text can be estimated with different stylometry measures. An overview is presented in Table 3.1.

The type/token ratio (ttr) represents the percentage of different words in a document; $\max(ttr) = 1$ implies that no single word in d appears more than once. The higher the value of ttr, the more diverse the vocabulary in a text is. Another option to estimate the vocabulary richness of d is counting the number of *hapax legomena* and *dislegomena* (i.e., words with frequency 1 and 2 in d). As seen in Section 2.3 this is a well-known feature in forensic linguistics as well.

[22] The idea of a "stylistic fingerprint" has to be taken with caution, though. For instance, Olsson (2008, pp. 31–32) refers to linguistic fingerprints as a myth stressing that "the proof of its existence is notable for its absence" (mainly because the ideas behind it have not been proven). A linguistic fingerprint that evolves over time due to knowledge acquisition (versus DNA or fingerprints) cannot evolve in a style measurement. Still much of the related literature during the last decades offer experimental results exploiting this concept.

Table 3.1: Summary of vocabulary richness measures. *syl*= syllable, *tok*= token, *typ*= type, tok_i= word occurring i times (e.g. tok_1 refers to hapax legomena).

Name	Equation	Description
Type/token ratio	$ttr = \frac{\|typ\|}{\|tok\|}$	Number of types divided by the number of tokens.
Hapax legomena	$hl = \|tok_1\|$	Number of hapax legomena in a text.
Hapax dislegomena	$hd = \|tok_2\|$	Number of hapax dislegomena in a text.
Honoré's R	$R = \frac{100 \log(\|tok\|)}{1-\|tok_1\|/\|typ\|}$	A representation of the number of tokens normalised by the relationship between hl and types.
Yule's K	$K = \frac{10^4\left(\sum_{i=1}^{\infty} i^2 \|tok_i\| - \|tok\|\right)}{\|tok\|^2}$	The occurrence of a word is assumed to behave as a Poisson distribution and is normalised by the number of tokens.

As Stamatatos (2009a) points out, these simple measures of vocabulary richness depend on text-length. In agreement with the well-known Zipf's law, the longer a text, the less new words are going to appear. As a result, some other measures have been proposed, such as Honore's R and Yule's K. The R function (Honore, 1979) "tests the propensity of an author to choose between the alternatives of employing a word used previously or employing a new word" (Holmes, 1992). The K function (Yule, 1944), in turn, is a different alternative to R. The advantage of Yule's K is that it is constant with respect to the length of the text (Sichel, 1975).[23]

At a lower level, one of the most successful characterisations in authorship attribution (Stamatatos, 2009a), which has been applied to intrinsic plagiarism detection (Stamatatos, 2009b), is based on character n-grams profiling, often considering $n = \{2, 3, 4\}$ (cf. Section 3.1.3). The success of this simple characterisation is that it is able to capture function words (as most of them are shorter than 4 characters), suffixes, and prefixes (cf. Section 5.1.1). Roughly, the idea of the character n-grams profiling is characterising a document as a vector of n-grams and their frequencies in order to further estimate how similar they are to other documents or text fragments (cf. Section 5.1.1).

Another way of analysing the complexity and diversity in a document's vocabulary is on the basis of the known as frequency class of a word, an "indicator of a word's customariness" (Meyer zu Eißen and Stein, 2006). Meyer zu Eißen and Stein (2006) define it as follows. Let C be a text corpus and $|C|$ be the size of C in terms of words. Let $|C_i|$ denote the frequency of word $w_i \in C$, $r(w_i)$ be the rank of w_i in a the list of words in C sorted by decreasing frequency, and w^* the most frequent word in C. The word frequency class $c(w_i)$ of $w_i \in C$ is defined as $\lfloor \log_2(f(w^*)/f(w_i)) \rfloor$. As a result, the most frequent word, w^*, corresponds to the word frequency class 0.

[23] As seen in Holmes (1992).

Table 3.2: Summary of text complexity measures. *syl*= syllable, *tok*= token, tok_c= complex words (with three or more syllables), *sent*= sentence.

Name	Equation	Description
Gunning fog index	$I_G = 0.4 \left(\frac{\|tok\|}{\|sent\|} + 100 \cdot \frac{\|tok_c\|}{\|tok\|} \right)$	A combination of the ratio of words per sentence and the ratio of complex words and the entire set of tokens.
Flesch reading ease	$F_e(d) = 206.835 - 1.015 \left(\frac{\|tok\|}{\|sent\|} \right) - 84.6 \left(\frac{\|syl\|}{\|tok\|} \right)$	A combination of the ratio of words per sentence and the ratio of syllables per word.
Flesch–Kincaid grade level	$F_g(d) = 0.39 \left(\frac{\|tok\|}{\|sent\|} \right) + 11.8 \left(\frac{\|syl\|}{\|tok\|} \right) - 15.59$	A combination of the ratio of words per sentence and the ratio of syllables per word.

3.4.2 Syntactic Features

These features measure the writing style at sentence level. As the length of a sentence can be considered as a reflection of how complex understanding it is, this factor represents a good text complexity measure. These lengths are often considered as the number of words the sentence contains (cf. Table 3.2 to see how to combine it to compute text complexity estimators).

3.4.3 Part of Speech Features

The idea of these features is counting the number of specific words in a text, for instance adjectives or nouns. Another option is considering sequences of POS n-grams. The disadvantage of this characterisation is its strong language dependency and the necessity of POS tagging the text (Argamon and Levitan, 2005)[24].

3.4.4 Closed-Class and Complex Words Features

Particular attention could be paid to the known as function words (stopwords). As these words are topic independent, they compose a good representation of an author's style. The use of foreign and "complex" words are an interesting factor as well. For instance, some people consider that a word with three syllables (in English) can be already considered complex (Gunning, 1968). The existence of complex words, together with the length of the sentences have been often combined to create text complexity measures. Some of the most representative ones are described in Table 3.2.

An example of this kind of feature is the Gunning fog index (Gunning, 1968)[25], which aims at determining how readable a document is. The value resulting of this calculation

[24] As we first observed it in (Stamatatos, 2009a).

[25] As seen in (Wikipedia, 2008).

Feature	Q. mechanics	Physics	C. Chaplin
$\|sent\|$	1,244	903	1874
$\|tok\|$	9,245	6,877	13,332
$\|tok\|/\|sent\|$	7.43	7.62	7.11
$\|tok_1\|$	4,715 (0.51)	3,396 (0.49)	7,654 (0.57)
$\|tok_2\|$	2,037 (0.22)	1,691 (0.25)	3,439 (0.26)
$\|tok_3\|$	1,509 (0.16)	975 (0.14)	1,508 (0.11)
$\|tok_{i\geq 4}\|$	984 (0.11)	815 (0.12)	731 (0.05)
$\|tok_{i\geq 3}\|/\|tok\|$	0.27	0.26	16.79
$\|syl\|/\|tok\|$	1.87	1.88	1.65
I_G	13.76	13.46	9.56
F_e	41.42	39.64	60.14
F_g	9.33	9.62	6.64

Table 3.3: Example of stylometric and complexity measures in three different texts: the Wikipedia articles about Quantum mechanics, Physics, and Charles Chaplin (as for December 4th, 2011) *syl*= syllable, *tok*= token, tok_i= token with i syllables –absolute values and (percentage), *sent*= sentence.

can be interpreted as the number of years of formal education required to understand the document contents. The Flesch–Kincaid readability tests (Kincaid, Fishburne, Rogers, and Chissom, 1975)[26] were created with the same purpose: determining how complex a text is to understand. The first of the tests is known as Flesch reading ease. The higher the value of this measure, the easier to understand the text. The second test is known as Flesch–Kincaid grade level. The obtained value represents the U.S. grade level necessary to understand a text.

Most of the aforementioned measures are relatively language independent (sometimes only a tokeniser is necessary). However, in other cases more resources are required. For instance, for the POS features a POS tagger, for the closed-class words a lexicon, and to compute the word frequency class a reference corpus. An approximation to identifying closed-class words which is still language independent is considering the k most frequent words in a collection of documents (Stamatatos, 2009a).

Examples of the obtained results by some of these measures are included in Table 3.3. Assuming that, in general, the longer a word the more complex it is, the articles on quantum mechanics and physics can be considered more complex than the one on Charles Chaplin (the rate of words with only one and two syllables is higher in the last article). The token/sentence ratio confirms the assumption: the third article contains shorter (less complex) sentences on average. The three complexity measures (Gunning fog index and both Flesh tests) conclude that more academic preparation is required to understand the first two articles.[27]

[26] As seen in (Wikipedia, 2011b).

[27] The figures in Table 3.3 were obtained with `http://juicystudio.com/services/readability.php`. The similar Stylysis Web service is available at: `http://memex2.dsic.upv.es:8080/StylisticAnalysis/en/`.

3.5 Chapter Summary

In this chapter we established the principles upon which automatic text re-use and plagiarism detection are based. First, we described different techniques for text pre-processing, including character, word, and sentence level normalisation operations. Afterwards, we described a set of text representation schemas to compose the set of terms that represent a document's contents. We included bag of words, word and character n-grams, and pseudo-cognates. We also considered a special representation model that may improve the comparison speed of models based on the rest of representation options: hashing.

Once the text representation models were presented, we discussed a set of weighting models which aim at reflecting how relevant a term for a document is. Both Boolean and real valued weighting schemas were described, including the well-known term frequency, document frequency, and term frequency-inverse document frequency models.

Having an option of text representation together with a weighting schema, comparison between documents can be performed. A total of four text similarity estimation families were discussed, including vector space, probabilistic, fingerprinting, and signal processing techniques. The close relationship between each similarity model and the corresponding text characterisation and weighting schemas was stressed. These three elements —representation, weighting and similarity estimation— will be exploited in Chapters 5 and 6, when describing external plagiarism detection.

A different way of analysing a document is on the basis of its style and complexity. The chapter is closed with a discussion on the features that represent such characteristics. We included lexical features (mainly vocabulary richness), character features (in the form of character n-grams profiles), syntactic features (based on syntax), and structural features (mainly based on words and sentences complexity).

Related publications:

- Barrón-Cedeño, Rosso, Pinto, and Juan (2008)
- Pinto, Civera, Barrón-Cedeño, Juan, and Rosso (2009)
- Barrón-Cedeño, Rosso, Agirre, and Labaka (2010c)
- Potthast, Barrón-Cedeño, Stein, and Rosso (2011a)

Chapter 4

Corpora and Evaluation Measures

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description [based upon it] would be no more than a mere list.
>
> Noam Chomsky

One of the main drawbacks in automatic plagiarism detection has been the lack of resources for its analysis. As in any other NLP and IR task, a dataset and a set of measures compose the evaluation framework necessary to develop, tune, and, perhaps even more important, compare different approaches under a common setting. This chapter covers some of the most interesting corpora available for the analysis of text re-use and plagiarism, including some we have helped on building. An overview of the measures used when evaluating text re-use detection is also offered.

The chapter opens with an overview of corpora and evaluation measures used in research on text re-use detection (Section 4.1). The rest of the chapter is clearly divided in two parts. Available corpora for manual analysis and the development of automatic models for text re-use and plagiarism detection are described in Section 4.2. The formalisms available when aiming at evaluating automatic text re-use detection are then discussed in Section 4.3.

Key contributions The author of this dissertation participated in the development of the co-derivatives corpus (Section 4.2.2). He co-operated with researchers from the Bauhaus-Universität Weimar in the conception of the PAN-PC series of corpora (Section 4.2.3); the most important contributions were made for the PAN-PC-09, where strategies for exploiting freely available text resources and for simulating plagiarism instances were defined. For the CL!TR corpus (Section 4.2.5), the author adapted the process followed by Clough and Stevenson (2011) to generate simulated cases of cross-language re-use (though the gathering process was carried out by researchers from the AU-KBC Research Centre). Regarding the evaluation measures, the author participated in the conception of a preliminary version of the measures described in Section 4.3.3

Table 4.1: Plagiarism detection evaluation in 105 papers on text re-use and plagiarism detection. Published in Potthast *et al.* (2010a).

Evaluation Aspect	Pct.
Experiment Task	
local collection	80%
Web retrieval	15%
other	5%
Performance Measure	
precision, recall	43%
manual, similarity	35%
runtime only	15%
other	7%
Comparison	
none	46%
parameter settings	19%
other algorithms	35%

Evaluation Aspect	Pct.
Corpus Acquisition	
existing corpus	20%
home-made corpus	80%
Amount of documents	
$[1, 10)$	11%
$[10, 10^2)$	19%
$[10^2, 10^3)$	38%
$[10^3, 10^4)$	8%
$[10^4, 10^5)$	16%
$[10^5, 10^6)$	8%

(and applied in Chapter 7).

4.1 Overview of Corpora and Evaluation Measures Exploitation

Potthast *et al.* (2010a) performed a survey of evaluation resources and strategies in automatic plagiarism detection. They found more than one hundred scientific papers dealing with different kinds of text re-use, including versioning, co-derivation, and plagiarism. Their findings are summarised in Table 4.1.

Regarding the experimental plagiarism detection task, 80% of the research work has been carried out considering a local collection of documents. It comes out that very often such a collection is private and hardly published. Around 15% of the papers perform their experiments over the Web. Nevertheless, in this context no direct comparison to other models is easy, due to the Web's inherent dynamic nature. The nature of the task is highly related to the corpus acquisition strategy (right hand side in Table 4.1): in 80% of the cases the experiments are carried out on a home-made (potentially newly created, hardly used again) corpus. Nearly 70% of the research work is carried out considering less than $1,000$ documents. On the one side, as Potthast *et al.* (2010a) point out, these corpora are composed of student coursework or documents with manually inserted cases of simulated plagiarism. On the other side, larger corpora (used by roughly the rest 30%) come from writing environments where different versions of a text can be produced (e.g. news wire articles).

An overview of the corpora discussed in Section 4.2 is presented in Table 4.2.[1] Corpora covering different domains and languages have been generated for the study of text-

[1]Note that we use the ISO 639-1 language codes.

Table 4.2: Overview of discussed corpora. A black (white) square appears if a corpus (partially) accomplishes with some feature. CS→computer science, en→English, de→German, es→Spanish, hi→Hindi. Corpora we helped to generate appear highlighted.

corpus	focus	domain	language(s)	real	simulated	synthetic	fully annotated	cross-language
METER	journalistic re-use	press (politics & show-business)	en	■			□	
Co-derivatives	co-derivation	multiple (encyclopedic)	de, en, es, hi	■				
PAN-PC	plagiarism	multiple	de, en, es		■	■	■	□
Short answers	plagiarism	CS (encyclopedic)	en		■			
CL!TR	cross-language re-use	CS and tourism (encyclopedic)	en, hi		■			■

reuse and plagiarism detection. One of the strengths of the METER and co-derivatives corpora is that they contain real cases of re-use. By producing simulated (manually-created) and synthetic (algorithmically) re-use, more corpora (at a larger scale) can be generated, even allowing for fully annotating the borders between re-used and original text.

Yet another issue is the evaluation strategy followed by the researchers. More than 40% of the research works performs an evaluation based on the well-known precision and recall measures. Nevertheless, 35% follows a manual strategy, where an expert reviews the cases in order to notice how effective the model was. As a result, the experiments performed are expensive, hardly replicable, and nearly incomparable.

The lack of better defined evaluation frameworks in research on text re-use has caused the comparison among different models hard to be made. As a result, more than half of the papers analysed do not include a comparison of their proposed model against others, causing the actual appreciation of their quality to remain an open issue. The research on mono- and cross-language detection of text re-use and plagiarism requires two main factors: (*a*) publicly available corpora containing actual or simulated cases of re-use and (*b*) clear (potentially standard) evaluation measures properly defined and objective. In the sections to come we analyse resources that accomplish with these requirements.

4.2 Corpora for Plagiarism and Text Re-Use Detection

Clough (2003) appreciates that "real examples of students plagiarism are hard to come by due to restrictions in student confidentiality". He considers that building a test

collection for plagiarism detection —even if simulated— would offer many benefits. Such a collection could...

1. Stimulate research on automatic plagiarism detection;
2. Enable communities to compare different approaches;
3. Help us to better understand plagiarism; and
4. Be used to help teach students how to cite and paraphrase correctly by looking at examples of plagiarism.

With the aim of filling these gaps, different corpora have been created. Nevertheless, as seen throughout this section, these corpora have helped to fill gaps number 1 and 2 only. Gaps 3 and 4 have been less approached due to the difficulties here discussed.

In general, documents containing actual cases of plagiarism are not freely available. Whereas the confidentiality issues are not present when analysing text re-use (taking the plagiarism label out), it is still hard to come out with a corpus where the re-used documents or, even better, the re-used text fragments and their corresponding sources are identified. Three scenarios exist to resolve these problems:

1. Considering a local collection of suspicious documents and the Web as their potential source;
2. Using a corpus of real text re-use; and
3. Creating a corpus with simulated cases of re-use/plagiarism.

As seen in Section 2.5, the Web is a preferred source for plagiarism. As a result, scenario 1 is worth considering when developing a plagiarism detection model. Nevertheless, the disadvantages of this option are twofold. On the one hand, the suspicious documents must be manually analysed in order to determine whether they actually contain re-used fragments and the source texts have been identified properly. This has to be done either before or after the experimentation in order to perform an objective evaluation. On the other hand, comparison of models by different researchers becomes practically impossible, given how dynamic the Web is. Albeit a Web retrieval scenario is more realistic, from an experimental point of view it does not seem to be the best option because of the lack of control of the dynamic Web contents for an objective comparison.

Scenario 2 aims at avoiding any confidentiality and legal issue. The main difficulty in this case is the necessity of manually reviewing a set of documents in order to annotate every case of re-use together with its actual source. Fortunately, a corpus created under this principle exists and is freely available: the METER corpus (cf. Section 4.2.1).

Scenario 3 has gained popularity in recent years. One reason is the rise of controlled frameworks aiming at creating evaluation environments where an objective evaluation can be carried out and different models can be compared. Either generated in the framework of a competition or simply trying to encourage direct comparison among models, existing corpora with simulated cases of re-use/plagiarism include the series of PAN corpora, the co-derivatives corpus, the short plagiarised answers corpus and the CL!TR corpus. The PAN corpora include a mixture of artificially created and human simulated cases of plagiarism. The co-derivatives corpus is a sub-collection of Wikipedia articles' revisions, hence containing cases of real re-use. Both the short plagiarised

answers and CL!TR corpus contain cases of human simulated plagiarism only; the former one is monolingual (English), whereas the latter one is cross-language (English-Hindi). These corpora are described in Sections 4.2.2 to 4.2.5.

4.2.1 METER Corpus

METER stands for *Measuring TExt Reuse*, a project held at the University of Sheffield.[2] Its main aim was analysing the phenomenon of text re-use in British press.[3] The interests were twofold: (*i*) monitoring the amount of texts from the Press Association (PA), a British provider of news contents, that actually reach the reader through British press (Wilks, 2004); and (*ii*) defining re-use analysis algorithms in order to detecting and measuring it.

As pointed out by Clough (2003), the journalistic process for notes generation and publication is as follows. News agencies provide their subscribers with a huge collection of news wires. By paying a subscription fee, press and media get the right for reusing agencies' materials, particularly their texts (agencies provide pictures, audio and videos as well). Subscribers earn the right for publishing the contents *verbatim* or modifying them as it fits their own interests. Clough *et al.* (2002, p. 1678) identify some characteristic constrains when preparing a note for publishing. The most interesting are the following:

(a) Prescriptive writing practices;
(b) Editorial bias;
(c) A newspaper's house style;
(d) Readability and audience comprehension;
(e) Short deadlines; and
(f) Limits of physical size during page layout.

Constrains (a) to (c) have to do with the newspaper profile and pre-defined writing style (either it is a left or right wing publication). Constrain (d) has to do with how complex the syntactic structures in the publication are, related to the aimed audience (for instance, if it is quality or popular press). Constrains (e) and (f) regard to the press common environment: events have to be covered and published as soon as possible and all the relevant information must fit in a limited space. These last two constrains are common in every authoring environment, in particular academia. Students, researchers and other writers often face (short) deadlines as well (cf. Section 2.4). Deadlines for students and researches are always well-defined and, whereas longer than in press, they still represent a difficulty to deal with. About researchers, scientific authoring has to permanently deal with deadlines. Conferences and journals define minimum and maximum lengths for the received notebooks. Moreover, scientific and dissemination forums exist, requiring an adaptation of complexity and style at writing.

As with academic plagiarism, Clough (2003, p. 3) appreciates that the arrival of electronic media has promoted a culture of cut & paste in journalism.[4] At the end of the day, the kind of rewriting (paraphrasing) journalism, school, and other authoring

[2] `http://nlp.shef.ac.uk/meter/`

[3] Another attempt to analyse re-use in media was carried out by Lyon, Malcolm, and Dickerson (2001), which used a broadcast news corpus including 334 TV reports.

[4] Cf. Section 2.4.1.3 for an overview of journalistic plagiarism cases.

environments provoke is quite similar. Yet another similarity between journalistic text re-use and plagiarism exists: the news-worker could be considered to be an "experienced plagiary", one with high editing skills (Clough, 2003, p. 6).

The *METER corpus* (Clough *et al.*, 2002) is one of the main outcomes of this project. It is considered as a seminal effort on automatically analysing the phenomenon of text re-use. The corpus is composed of a collection of news stories produced by the PA, as well as notes about the same events published by nine British newspapers.[5]

As aforementioned, the newspapers are allowed to re-use the PA notes as a source for their own publications. Therefore, the specific kind of borrowing this corpus is composed of is *journalistic re-use*; contrary to plagiarism, an acceptable activity. The interesting fact behind this corpus is that, for a news story published by the PA, newspaper articles are identified covering the same event, either independently or not. As a result, a good part of the vocabulary, for instance named entities[6], are expected to co-occur.

Hard (politics, diplomacy, disasters) and soft news (editorials, commentaries, feature stories) are included. One topic per kind of event was considered, namely: (*i*) law and courts, and (*ii*) show business (Clough *et al.*, 2002, p. 1680). Three factors led to considering these topics. Firstly, they use to appear on a daily basis in press, gathering lot of attention. Secondly, they use to be covered by most of the media. Thirdly, court stories tend to be more rigid, including limited vocabulary, while show business are more flexible. As a result, re-use of hard news is supposed to be more easily detected than soft news. As expected, quality press contains more reports on hard news; the opposite occurs for tabloids.

The corpus was annotated by a professional journalist. Two levels of text unit were defined when annotating, with three degrees of re-use each. The levels are *whole document* and *word sequence*. For whole document level, three degrees of derivation were defined:

1. *Wholly-derived*. The PA note is the only source of the newspaper text.
2. *Partially-derived*. The PA note is one of the sources of the newspaper text, but not the only one.
3. *Non-derived*. The PA note is not considered at all.

At word sequence level, specific text fragments are identified at three different levels respect to their derivation relationship to the corresponding PA note:

1. *Verbatim*. The copy is made word-for-word (exact copy).
2. *Rewrite*. The text is paraphrased.
3. *New*. The text is not directly related to the PA text.[7]

[5]The newspapers included comprise both quality press and broadsheets. Quality press representatives include *The Times*, *The Guardian*, and *The Telegraph*. Broadsheets (also known as tabloids) include *The Sun*, *Daily Star*, *Daily Mail*, *Express*, *The Mirror*, and *The Independent*. The considered news were published in paper between 12 July 1999 and 21 June 2000.

[6]A word or sequence of words that represent a name of a person, a location, an organisation, etc.

[7]Clough *et al.* (2002, p. 1680) point that this fragments do include fragments from the PA, but they are "not used in the same context".

Table 4.3: Statistics of the METER corpus. Figures shown for the entire corpus as well as the courts and show business partitions. Global is the combination of newspapers and PA notes. The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{chars}|$ total number of characters, $|D_{tokens}|$ total number of tokens, $|D_{types}|$ total number of types, $|d_{chars}|$ mean characters per file, $|d_{tokens}|$ mean tokens per file, and $|d_{types}|$ mean types per file.

	$\|D\|$	$\|D_{chars}\|$	$\|D_{tokens}\|$	$\|D_{types}\|$	$\|d_{chars}\|$	$\|d_{tokens}\|$	$\|d_{types}\|$
PA	773	$1.6M$	$285k$	$16k$	$2,053 \pm 1,207$	368 ± 242	21.17 ± 25.52
Newspapers	945	$1.7M$	$337k$	$17k$	$1,827 \pm 1,448$	356 ± 285	18.01 ± 23.29
Global	1,718	$3.3M$	$621k$	$23k$	$1,928 \pm 1,350$	361 ± 266	13.11 ± 18.70
Court partition							
PA	661	$1.4M$	$245k$	$14k$	$2,068 \pm 1,268$	370 ± 254	21.25 ± 26.92
Newspapers	770	$1.5M$	$287k$	$14k$	$1,921 \pm 1,494$	373 ± 294	18.66 ± 24.83
Global	1,431	$2.8M$	$532k$	$19k$	$1,989 \pm 1,396$	372 ± 276	13.29 ± 19.84
Show business partition							
PA	112	$220k$	$40k$	$6k$	$1,963 \pm 750$	358 ± 156	50.72 ± 39.61
Newspapers	175	$248k$	$49k$	$6k$	$1,415 \pm 1,141$	282 ± 229	36.50 ± 40.40
Global	287	$467k$	$89k$	$9k$	$1,629 \pm 1,042$	312 ± 207	30.27 ± 33.86

The corpus is composed of 1,718 texts (more than 600,000 words). Some statistics for the entire METER corpus as well as the *courts* and *show business* partitions are included in Table 4.3. In total, 945 newspaper articles and 773 PA reports are included. For the *court* partition 770 (661) newspaper (PA) notes exist. For the *show business* partition 175 (112) newspaper (PA) notes exist. As expected, court news are in general longer, whereas the variety of vocabulary is bigger for show business.

As described by Clough *et al.* (2002, p. 1682), the corpus was created attempting to include documents of varying lengths (from one-sentence summaries, to brief stories, and long reports). A representation of the METER corpus length distribution in terms of characters, tokens and types is depicted in Table 4.4. The distribution of derivation relationship between the reports at document level is summarised in Table 4.5. Only 445 newspaper articles are annotated at word sequence level. Unfortunately, the link between verbatim and rewritten sequences and the corresponding source text in the PA document are not provided in the corpus version we had access to.

When analysing the METER corpus for text re-use, the PA notes are potential source documents, whereas the newspapers are candidate re-users (Clough, Gaizauskas, Piao, and Wilks, 2001). A sample of source (PA) and re-used (newspaper) case is included in Table 4.6.

It is worth noting that, according to Clough *et al.* (2002, p. 1680), it is very likely that the newspapers' texts consider the PA copy as one of their sources. This is mainly because large amounts of money are invested in getting access to this resource and the PA has a wide coverage of events. As a result, the texts in the newspaper texts have a high a-priori probability of being re-used at some extent from the PA note.

The text re-use in this corpus has been assumed to represent plagiarism cases in different research works, such as that of Sánchez-Vega, Villaseñor-Pineda, Montes-y Gómez,

Table 4.4: Documents length variety in the METER corpus. The number of documents in a given range of characters (left), tokens (middle), and types (right) are shown. For each histogram, the left hand side number represents the high threshold in the range ($< 2,000$ in the characters histogram includes documents such that $1,000 < |d| \leq 2,000$). The bottom numbers represent the absolute number of documents, whereas the numbers over the histograms represent the percentage of documents within the corpus.

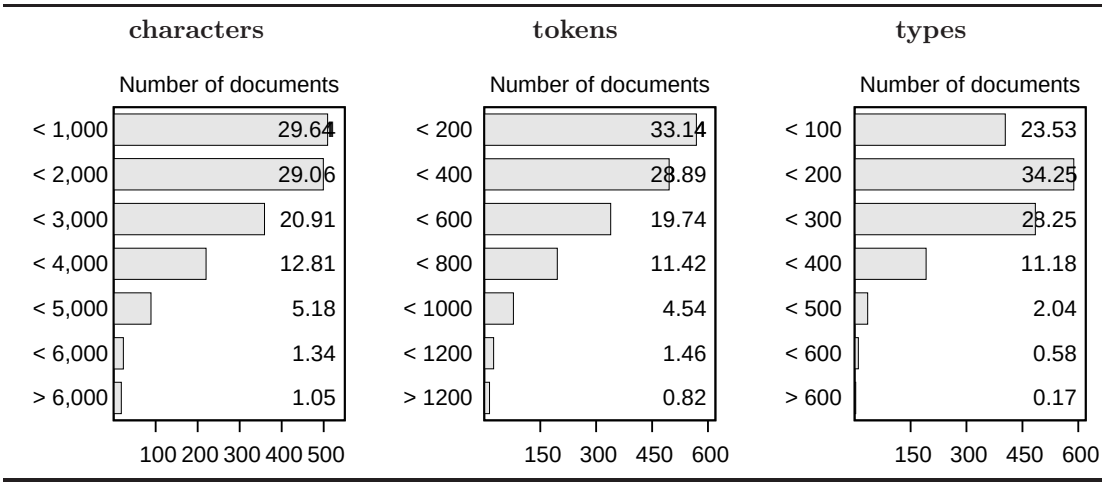


Table 4.5: Percentage of text re-use between the PA and newspaper notes in the METER corpus. As published by Clough *et al.* (2002).

Derivation degree	Court	Show business
Wholly-derived	34%	22.2%
Partially-derived	44.4%	54.9%
Total derived	78.4%	77.1%

and Rosso (2010). Our experiments over the METER corpus are described in Sections 5.2 to 5.4.

4.2.2 Co-derivatives Corpus

According to Bernstein and Zobel (2004) two documents d_1 and d_2 are co-derived if some portion of d_1 is derived from d_2 or both d_1 and d_2 are derived from d_3. They include, among other kinds of co-derivatives, document revisions, digests and abstracts. One of the biggest collections of revisions at hand is Wikipedia. Therefore, this co-derivatives corpus was generated on the basis of Wikipedia articles revisions (Barrón-Cedeño *et al.*, 2009a). In order to compose a multilingual experimental framework, a collection D of articles a was defined for a total of four languages. D was built under the following three main considerations:

Language Four are included: English, German, Spanish, and Hindi (en, de, es, hi).

Articles selection The documents consist of the set of most popular articles in the corresponding language. The popularity is measured as the frequency of access to the articles; 500 articles were considered per language ($2,000$ in total).

History The collection of co-derivatives is composed of 10 revisions $a_{k,1} \ldots a_{k,10}$ per article a_k. Such revisions were equally distributed in time among the 500 most

Table 4.6: A news story as covered by the PA and *The Telegraph*. Verbatim and rewritten fragments are identified with bold and italics in the PA text. The XML annotation format originally found in the corpus is depicted in the newspaper text.

PA version: Titanic restaurant case discontinued

Celebrity **chef Marco Pierre White** *today* **won** *the battle of* **the Titanic and Atlantic restaurants**. **Oliver Peyton, owner of the Atlantic** Bar and Grill, **had tried to** *sink Marco's* **new Titanic restaurant housed in the same West End hotel in London by seeking damages against landlords Forte Hotels and** *an* **injunction in the High Court**. But today the Atlantic announced in court it had reached a confidential agreement with the landlords and was discontinuing the whole action.
Mr Peyton, whose action began on Monday, had **claimed that the Titanic was a replica of the Atlantic**, with the same art deco style and attracting the same clientele **and should not be allowed to trade in competition** because he has exclusive rights under his lease **at the Regent Palace Hotel** off Piccadilly Circus.

The Telegraph version

```
<Rewrite>  THE </Rewrite>
<Verbatim> chef Marco Pierre White </Verbatim>
<Rewrite>  yesterday </Rewrite>
<Verbatim> won </Verbatim>
<Rewrite>  a dispute over </Rewrite>
<Verbatim> the Titanic and Atlantic restaurants. </Verbatim>
<Verbatim> Oliver Peyton, owner of the Atlantic, had tried to </Verbatim>
<Rewrite>  close White's </Rewrite>
<Verbatim> new Titanic restaurant, housed in the same West End hotel in London,
   by seeking damages against </Verbatim>
<Rewrite>  the </Rewrite>
<Verbatim> landlords, Forte Hotels, and </Verbatim>
<Rewrite>  a </Rewrite>
<Verbatim> High Court injunction.</Verbatim>
<Rewrite>  He </Rewrite>
<Verbatim> claimed that the Titanic was a replica of the Atlantic and should not
   be allowed to trade in competition at the Regent Palace Hotel. </Verbatim>
```

recent revisions; i.e., the number of revisions available in the "live" Wikipedia. Note that there is no reason to have the same topics in every language.

The most recent revision of article a_k by the crawling time (2009) is represented as $a_{k,1}$. A revision was discarded if: (i) a *Wikipedian* rejected $a_{k,t}$ due to vandalism[8], or (ii) the changes between $a_{k,t-1}$ and $a_{k,t}$ were minimal. The former rule was applied in order to avoid considering revisions for which the co-derivation relationship is not guaranteed. Our naïve approach, rather than considering a formal model for automatic vandalism analysis (Potthast, Stein, and Gerling, 2008b), was based on whether an edition survived at the following revision $a_{k,t+1}$. That is, if an edition was rejected at the next review, it was excluded from the revisions sample. The latter rule aims at assuring that each revision $a_{k,t}$ had a different level of similarity with respect to the

[8]Wikipedia articles are often affected by vandalism, which particularly describes the deletion or modification of its contents with malicious intentions (cf. `http://bit.ly/pan_vandalism_2011`).

Table 4.7: Co-derivatives corpus statistics. The figures are shown for the four different languages: English (en), German (de), Hindi (hi) and Spanish (es). The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{chars}|$ total number of characters, $|D_{tokens}|$ total number of tokens, $|D_{types}|$ total number of types, $|d_{chars}|$ mean characters per file, $|d_{tokens}|$ mean tokens per file, and $|d_{types}|$ mean types per file.

Lan	$\|D\|$	$\|D_{chars}\|$	$\|D_{tokens}\|$	$\|D_{types}\|$	$\|d_{chars}\|$	$\|d_{tokens}\|$	$\|d_{types}\|$
de	5050	$186.7M$	$29.4M$	$325k$	$36,973 \pm 34,662$	$5,821 \pm 5,279$	$1,944 \pm 1,468$
en	5050	$264.2M$	$48.5M$	$274k$	$52,311 \pm 31,828$	$9,597 \pm 5,766$	$2,556 \pm 1,297$
es	5050	$128.5M$	$23.4M$	$166k$	$25,444 \pm 29,962$	$4,637 \pm 5,466$	$1,339 \pm 1,288$
hi	5050	$19.3M$	$5.2M$	$78k$	$3,819 \pm 6,714$	$1,025 \pm 1,761$	308 ± 384

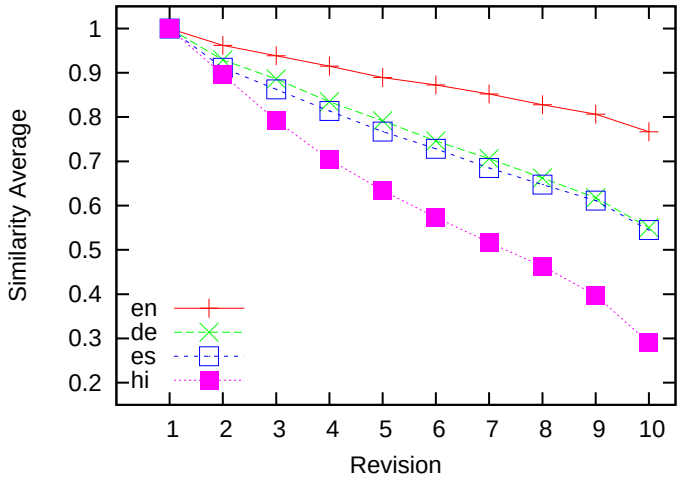


Figure 4.1: Evolution of mean similarities in the co-derivatives corpus. Similarity measured between $a_{k,1}$ and its preceding revisions, estimated on the basis of the Jaccard coefficient (cf. Section 3.3.1.1).

others, particularly respect to $a_{k,1}$. In order to do this discrimination we took advantage of the tags available in Wikipedia itself.

The co-derivatives corpus can be formally described as follows. Let D be a collection of documents in different languages, namely $\{D_{de}, D_{en}, D_{es}, D_{hi}\} \in D$. Each subcollection D_{xx} includes K=500 topics. For every topic a_k, ten revisions $\{a_{k,1}, a_{k,2}, \ldots, a_{k,10}\}$ exist. For experimental purposes, D_q can be defined as a set of query-documents $\{a_{1,1}, \ldots, a_{K,1}\}$; i.e., the most recent article per topic at hand ($D_q \subset D_{xx}$). Defining $D_q \subset D$ aims at considering samples of co-derivatives which are in fact exact copies.

Two versions of the corpus are available: original and pre-processed. The original version contains the articles' text without any further processing. The pre-processed version includes the same documents after spaces normalisation, sentence detection, tokenisation and case folding. Some statistics of the co-derivatives corpus are included in Table 4.7.

Figure 4.1 shows the average evolution of similarity between $a_{k,1}$ (i.e., the defined as query document) and the different articles revisions $a_{k,1} \ldots a_{k,10}$. As expected, the similarity decreases for more distant revisions. On the one side, the evolution of the English revisions is clearly slighter than for other languages. On the other side, the revisions in Hindi show an obviously stronger evolution. Evolution of Spanish and German revisions seems quite similar. The tendency of the similarity in the four languages might be explained by the maturity of the articles (a topic for further research). These differences

have to be considered when working with the different languages in this corpus.

A representation of the co-derivatives corpus length distribution in terms of characters, tokens and types is depicted in Fig. 4.2 for the four considered languages. For the German language, the most of the articles are no longer than 4,000 tokens. Spanish articles show to be less developed: most of the articles are shorter than 3,000 tokens. Hindi articles are still shorter: nearly 80% of them are shorter than 1,300 characters. English articles are much longer than for the other three languages. Indeed, more than half of the articles in this language have between 5,000 and 15,000 tokens. As a result, the difference between the four partitions in this corpus is not only related to the evolution of the revisions they contain. The amount of text to compare is an important factor as well.

The corpus is freely available on our website.[9] Our experiments over the co-derivatives corpus are described in Section 9.2.

4.2.3 PAN-PC Corpora

Both the METER and co-derivatives corpora include real cases of text re-use (the former journalistic and the latter of encyclopedic revisions). Nevertheless, they still have a few weaknesses. For instance, most of the documents in the METER corpus are short (the mean length in terms of tokens is around 360). For the case of the co-derivatives corpus, making a low level analysis (in terms of specific text fragments) is nearly impossible, as it lacks of any annotation about what specific text fragment from a given revision prevails in another one.

Aiming at filling these and other gaps, the *PAN-PC* series of corpora was generated. These corpora have been created in the framework of the *PAN: Uncovering Plagiarism, Authorship and Social Software Misuse* initiative, a workshop held annually since 2007 (PAN-PC stands for PAN Plagiarism Corpus). In the last three editions PAN adopted the format of a competition, including plagiarism and Wikipedia vandalism detection as well as authorship verification (Stein *et al.*, 2011a).[10] One of the main outcomes of these competitions have been annotated corpora for automatic plagiarism detection. Section 4.2.3.1 describes the ideals pursued and some principles considered while building the corpora. Section 4.2.3.2 describes the strategies applied when generating the re-use cases. Sections 4.2.3.3 to 4.2.3.5 discuss the specificities of every edition of the PAN-PC corpus. Finally, Section 4.2.3.6 includes a discussion on lacks and weaknesses in these corpora series, together with proposals for improving future editions.

4.2.3.1 PAN-PC Conception

The characteristics pursued from the beginning of the conception of the PAN-PC corpora are discussed following.

[9] `http://users.dsic.upv.es/grupos/nle/downloads.html`
[10] `http://pan.webis.de`

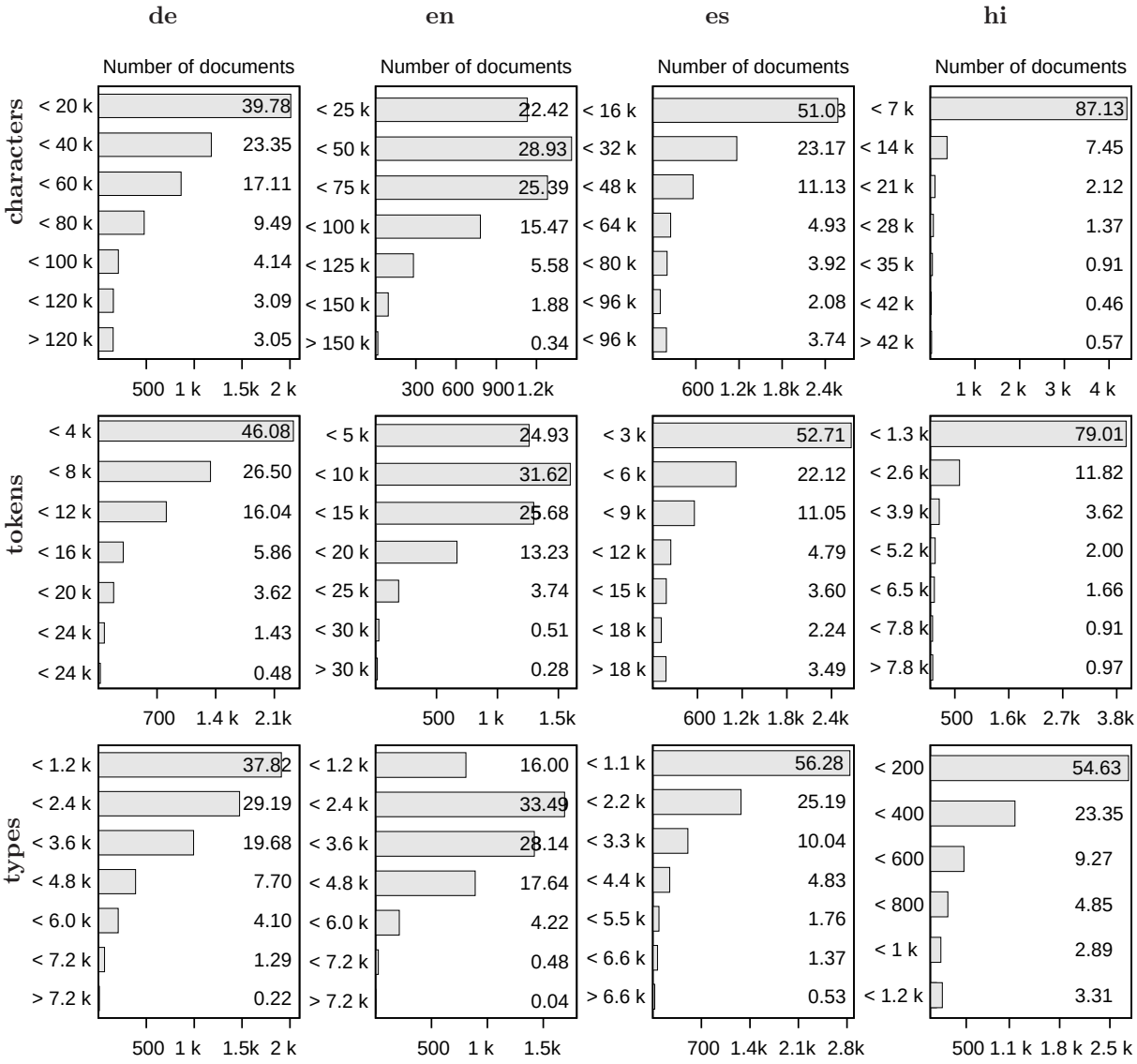


Figure 4.2: Documents length variety in the co-derivatives corpus. The number of documents in a given range of characters, tokens, and types are shown for each language. For each histogram, the left hand side number represents the high threshold in the range ($< 40k$ in the characters histogram includes documents such that $20,000 < |d| \leq 40,000$). The bottom numbers represent the absolute number of documents, whereas the numbers over the histograms represent the percentage of documents within the corpus.

A free corpus It may allow for the direct comparison of different detection approaches and represents a good way of fostering the development of more and better models. Nevertheless, an important issue in plagiarism-related corpora is that, in general, real cases of plagiarism are hard to include in a free corpus, mainly for ethical issues. As a result, the PAN-PC corpora include synthetic cases only; i.e., the cases are generated either automatically or manually, but always simulating plagiarism. For licensing issues the base texts[11] come from public domain documents only, i.e., copyright free documents.

[11]Base documents are those used either as source or suspicious document before the plagiarism case is generated. As described afterwards, a text fragment from a base document (source) is selected which, after a given modification process, is inserted into a different (suspicious) base document.

One of the biggest text collections at hand accomplishing with this characteristic is the one of the *Project Gutenberg*.[12] By considering this kind of material as the collection of base documents, it is possible to freely distribute the resulting corpus without any agreement or permission requests, and free of charge.[13]

Cases identification One of the aims of a plagiarism detector is, rather than simply determining whether a document is plagiarised from another one, identifying the specific borrowed fragment together with its potential source fragment. In order to evaluate this factor, corpora are required for which such fragments are explicitly marked. As every case in the PAN-PC corpora is synthetically generated, the borders between borrowed and original text can be precisely set (the same occurs for the source fragment). The suspicious text files are provided with an XML file. The information it contains is: (*i*) offset and length of the plagiarised fragment (if any), (*ii*) name of the source file of the borrowed text (if available), and (*iii*) offset and length of the source text fragment (again, if available).

Scale In order to compose a challenging (realistic) task, thousands of documents should be included. The synthetic nature of the proposed cases let for generating large document collections. Whereas still far of being realistic —achieving Web scale— considering settings of thousands of documents propitiate the interest in generating efficient models for plagiarism detection.

Detection-oriented nature As previously mentioned, two main approaches to text plagiarism detection exist: intrinsic and external. Therefore, the corpora include cases of plagiarism for which the source text is either available or not in the text collection. By means of these reasoning, three document sub-collections are included. Collection E consists of suspicious documents in which the source of the plagiarism cases is available in collection S. Collection S contains the potential source documents for the cases in E.[14] Collection I consists of suspicious documents in which the source of the plagiarised fragments is not available. Taking advantage of the nature of external plagiarism detection, collections E and S are not subject to any limitation about the number of authors of the base documents. On the contrary, the nature of intrinsic plagiarism detection forces the base documents of collection I to include documents that are unlikely to contain re-used text and, more important, a high probability of being written by one single author. In summary, both E and I include inserted borrowed text fragments. The fragments in E come from S. The fragments in I come from another, non included, text collection S' ($S \cap S' = \emptyset$).

Re-use variety On the quest of realism, not only scale, but generating different kinds of re-use is necessary. The range of modification should go from cut & paste through *different levels of paraphrasing and up to translation.* Some of these cases can be generated algorithmically, whereas some other require manual generation. Artificial cases

[12]Project Gutenberg is "the first and largest single collection of free electronic books" with roots in 1971 (Project Gutenberg, 2011).

[13]One of the first proposals on using Project Gutenberg documents to simulate plagiarism is that of Barrón-Cedeño and Rosso (2008).

[14]The extreme case where document d_1 contains borrowed material from d_2 and is used as source when writing d_3 is not considered.

are relatively easy and cheap to come with. The cost is that the variety of paraphrasing types that can be generated by these means is narrow.[15] As a result manually created samples have been included in some of the corpora. Human made cases are expensive both temporally and economically. As a result, just a few cases of this nature can be included (cf. Sections 4.2.3.4 and 4.2.3.5).[16]

The different kinds of re-use aimed at being included in these corpora are roughly the following: (*i*) exact copy: a document contains a 1:1 copy of (a fragment of) another document, (*ii*) modified copy: a document contains a modified/rewritten copy of (a fragment of) another document, and (*iii*) translated copy: a document contains a translation of (a fragment of) another document.

The range of rewriting aimed at considering may be well resumed as:

> *The degree of rewriting can vary from direct copying from a source with no attribution, the insertion or deletion of grammatical units, e.g. sentences or phrases, the insertion or deletion of words within a sentence, e.g. noun phrase modifiers, the reordering of words in a sentence or the reordering of sentences in a discourse, inversion of original clauses, substitution of equivalent words or phrases, changes in form such as tense and voice (e.g. active to passive voice), making abstract ideas more concrete (specification), making concrete ideas more abstract (generalisation), merging or separating sentences (either adjacent or dispersed throughout the original) and rewriting direct quotes as indirect (and vice-versa).*

This description is borrowed from Clough (2003, p. 3), who described what a proper corpus for plagiarism analysis should be.

In order to simulate this paraphrasing process, some monolingual cases were automatically obfuscated.[17] The automatic obfuscation strategies are discussed in Section 4.2.3.2.

Language variety Translation represents perhaps the most drastic obfuscation strategy as there is no straightforward relationship between source and borrowed texts' vocabulary and, in some cases, grammar. Cases of cross-language plagiarism have occurred since centuries ago (cf. Section 2.2.1). Moreover, as seen in Section 2.5.2, it is likely that plagiarism will cross language borders. Nevertheless, few attempts exist that approach the detection of translated plagiarism (cf. Chapter 6). In order to foster the development of cross-language models, cases of borrowing between different languages are worth considering. The three languages included in the corpora are English, German, and Spanish.

[15] Whereas artificially created cases of plagiarism are not considered to be ideal, they are considered an option when generating corpora for research on plagiarism detection (Clough, 2003).

[16] Determining what kind of modification is most applied when plagiarising remains an open issue. Cf. Chapter 8 for a discussion on paraphrase plagiarism: its generation and detection by state of the art automatic plagiarism detectors.

[17] Only cases for external detection are obfuscated. Cases generated for intrinsic detection are inserted without any modification in order to avoid modifying stylistic and complexity features of the texts, which are precisely the factors considered when trying to detect plagiarism with this approach (cf. Section 5.1.1).

Most of the documents are written in English and a few source documents in German and Spanish have been considered from which translated fragments are borrowed into English documents.[18] This process aims at resembling the fact that people whose native language is not English write in this language very often.

Positive and negative examples In real world scenarios not every plagiarism suspicious document contains borrowed fragments. Therefore, both collections E and I include two partitions: one of originally created documents and one of documents with plagiarism cases. The distribution is roughly $50\% - 50\%$.

4.2.3.2 Cases Generation

A good part of the plagiarism cases generation process in the three editions of the PAN-PC corpus is artificial; i.e., generated automatically, with software. The generation process can be divided into two main steps: (i) extraction-insertion and (ii) obfuscation.

Extraction-insertion It consists of two operations. Firstly, given a document $d \in S \cup S'$ a text fragment s that is going to be re-used in another document is selected. Secondly, s is inserted into a document $d \in E \cup I$. In the three PAN-PC corpora the selection of extracted fragments as well as the position where they are inserted is random. Attention is paid trying to select entire sentences though.

Obfuscation In order not to consider cut & paste borrowings only, different methods have been applied for modifying s before inserting it into a suspicious document. The aim is simulating the different paraphrasing strategies described in Section 2.1.1 and further investigated in Chapter 8. For the PAN-PC-09 corpus, only automatic obfuscation strategies were applied. In both PAN-PC-10 and PAN-PC-11 a short amount of cases was manually obfuscated. The automatic strategies applied to a text fragment s when plagiarising are the following:

1. **Same polarity substitutions**. The vocabulary in s is substituted by the corresponding synonym, hyponym or hypernym.
2. **Opposite polarity substitutions**. Some of the words in s are substituted by the corresponding antonym.
3. **POS preserving shuffling**. The tokens in s are re-ordered such that its original POS sequence is preserved.
4. **Random operations**. The words in s are shuffled, removed, inserted, or replaced randomly. In some cases, new vocabulary is inserted into s from its new context, the suspicious document s is inserted in.
5. **Machine translation**. s is translated either from German or Spanish into English.[19]

[18] German and Spanish were selected because they are native languages of most of the people behind this corpus (German for the researchers at the Webis group in the Bauhaus-Universität Weimar and Spanish for the NLE Lab at the Universidad Politécnica de Valencia), hence facilitating the task.

[19] The other side of the problem, i.e., writers that translate a source text from English into their native

Table 4.8: Statistics of the PAN-PC-09 corpus, containing 41,223 documents with 94,202 plagiarism cases (Potthast *et al.*, 2009).

document statistics				**obfuscation statistics**	
document purpose		document length			
source documents	50%	short (1-10 pp.)	50%	none	35%
suspicious documents		medium (10-100 pp.)	35%	paraphrasing	
– with plagiarism	25%	long (100-1000 pp.)	15%	– automatic (low)	35%
– without plagiarism	25%			– automatic (high)	20%
				translation	10%

Strategies 1 and 2 were so called in order to directly mapping them to paraphrase types (these modifications were originally grouped into the label "semantic word variation" (Potthast *et al.*, 2009, 2010a), something not strictly correct from the paraphrases point of view (cf. Section 8.1). Strategy 3 aims at mimicking syntax-based paraphrase changes (cf. Chapter 8 for more information on this and other types of paraphrases). Strategy 4 does not always produce human-readable text. While not the best option, this characteristic is somehow justified by the fact that standard text similarity models are often based on bag of words (cf. Section 3.1) and are still able to capture the similarity between non-readable text and its original version. Moreover, consider the case of a non-native writer whom writings do not accomplish with the grammatical rules all the time. Strategy 5 aims at simulating translated plagiarism. Chapter 6 is fully devoted to the research work carried out for detecting cross-language plagiarism.

4.2.3.3 PAN-PC-09

The first corpus developed for the PAN series of competitions is the called *PAN-PC-09*.[20] The PAN-PC-09 was originally divided in the three partitions mentioned in page 89: S, E, and I. The amount of cases of artificial plagiarism in E and I together is around $90,000$. The main statistics of this corpus are summarised in Table 4.8.

The collection comprises a total of $41,223$ documents between source and potentially plagiarised texts.[21] The partition S is composed of 50% of the documents and $E \cup I$ includes the rest 50%. About the documents lengths, they mimic books, thesis, papers, and assignments. As a result, documents with lengths ranging from one to thousand pages were generated. Figures of the suspicious and source documents lengths are shown in Table 4.13, where these values are compared to those of the 2010 and 2011 editions.

The amount of plagiarised text a document could contain is highly varied. The percentage of plagiarised text a suspicious document may contain in the PAN-PC-09 ranges from 0% to 100%. The length of the borrowed fragments is evenly distributed between 50 and $5,000$ words. Beside cut & paste cases, all the obfuscation strategies

language, is not represented in these corpora. English documents are used as source in the CL!TR corpus (cf. Section 4.2.5).

[20] `http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-09.html`

[21] The base documents from Project Gutenberg include $22,135$ in English, 527 in German, and 211 in Spanish.

Table 4.9: Statistics for the PAN-PC-10 corpus, containing 27,073 documents with 68,558 plagiarism cases: 70% (30%) of the documents are for external (internal) detection (Potthast *et al.*, 2010d).

document statistics					
document purpose		plagiarism per document		document length	
source documents	50%	hardly (5%-20%)	45%	short (1-10 pp.)	50%
suspicious documents		medium (20%-50%)	15%	medium (10-100 pp.)	35%
– with plagiarism	25%	much (50%-80%)	25%	long (100-1000 pp.)	15%
– without plagiarism	25%	entirely (> 80%)	15%		
plagiarism case statistics					
obfuscation		case Length		topic match	
none	40%	short (50-150 words)	34%	intra-topic cases	50%
artificial		medium (300-500 words)	33%	inter-topic cases	50%
– low obfuscation	20%	long (3000-5000 words)	33%		
– high obfuscation	20%				
simulated	6%				
translated	14%				

described in Section 4.2.3.2 are applied in this corpus.

4.2.3.4 PAN-PC-10

The PAN-PC-10[22] follows the same philosophy as its predecessor, but it represents an improved version. The statistics of the PAN-PC-10 corpus are given in Table 4.9. The most interesting differences to its predecessor are: (*i*) the new obfuscation strategies considered, and (*ii*) the topic relationship between source and suspicious document. Following, we summarise and further discuss the main improvements achieved in the PAN-PC-10 corpus.

Base documents sanitisation Cases of unnoticed derivation in the base documents were minimised as much as possible (e.g. anthologies and single books from the same author were not considered together). If two sentences shared an exact sequence of eight words (after stopword deletion), one of them was discarded. Additionally, the Project Gutenberg base documents were manually reviewed to guarantee their quality.

Better obfuscation strategies More parameters were considered in order to determine how a plagiarised fragment s had to be automatically obfuscated. For instance, the longer s was, the less modifying operations were applied. It is expected that the longer a borrowed text, the less likely the plagiariser would modify it (as at the end of the day, her aim is taking a short cut); hence this decision was taken trying to make the corpus more realistic. Once again, cross-language plagiarised texts were produced by machine translation.

Manually created cases The most interesting obfuscation model included in this corpus is manual generation. The extraction-insertion process (cf. Section 4.2.3.2) was

[22] `http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-10.html`

still made automatically, but the obfuscation process was carried out by human beings. The fragments s were submitted to the crowd-sourcing platform *Amazon Mechanical Turk*.[23] *Turkers* were explicitly asked to strongly paraphrase s having in mind that they were plagiarising it (Potthast *et al.*, 2010a). As a result of this request, these cases of plagiarism were the hardest to detect in the collection. Later on, a sample of these cases was further annotated at paraphrase level in order to better understand the capabilities and drawbacks of current plagiarism detection technology, composing the P4P corpus (cf. Section 8.2).

Topical relationship Possibly one of the main weaknesses of the PAN-PC corpora has been the low relationship between the source document s is extracted from and the suspicious document it is inserted in. In the PAN-PC-09 corpus this pair of documents was selected in a completely random way. As seen in Table 4.9, in the PAN-PC-10, in 50% of the plagiarism cases s was inserted in a document that was somehow related to the source it was extracted from. Before generating any case of plagiarism, the base documents were clustered by means of a repeated bisections process. More than twenty clusters were generated identifying some topics (e.g. history, science, or religion). In the intra-topic plagiarism cases both source and suspicious documents belong to the same cluster. In the inter-topic plagiarism cases, the source and suspicious documents belong to different clusters.[24]

4.2.3.5 PAN-PC-11

Whereas the focus when developing the PAN-PC-09 was on generating a challenging corpus in terms of dimension (cf. Section 4.2.3.3), the 2010 and 2011 versions were developed aiming at including more diverse kinds of plagiarism. The purpose was coming out with more realistic instances of plagiarism as well as taking into account the feedback of participants of the previous two editions of the PAN competition. In the PAN-PC-11 more relevance is given to paraphrase plagiarism (either automatically or manually obfuscated).

Some statistics regarding the composition of the PAN-PC-11[25] corpus are given in Table 4.10. As there observed, only 18% of the plagiarism cases are cut & pasted and more than 70% have some kind of monolingual obfuscation. The other relevant change is the introduction of new, manually post-paraphrased, translated samples.

Post-paraphrased translated cases In a similar fashion to that used in the PAN-PC-10 for generating manual cases of plagiarism, Mechanical Turk was used to request people for further obfuscating cases of translated plagiarism (Potthast *et al.*, 2011b).

Tables 4.11 and 4.12 show the statistics for the intrinsic and external partitions of the PAN-PC-11. In the intrinsic partition, more than 80% of the suspicious documents include 20% of plagiarised text at most (only 1% of the documents includes up to 50%).

[23] A marketplace for work that requires human intelligence. One of the most popular frameworks for crowd-sourcing `http://www.mturk.com`. A worker in this framework is often called "turker".

[24] The process was carried out by means of a hierarchical partitional clustering process available in `CLUTO` (Zhao, Karypis, and Fayyad, 2005) (`http://glaros.dtc.umn.edu/gkhome/views/cluto`).

[25] `http://www.uni-weimar.de/cms/medien/webis/research/corpora/pan-pc-11.html`

Table 4.10: Statistics of the PAN-PC-11 corpus, containing 26,939 documents and 61,064 plagiarism cases (Potthast *et al.*, 2011b).

Document Statistics					
Document Purpose		Plagiarism per Document		Document Length	
source documents	50%	hardly (5%-20%)	57%	short (1-10 pp.)	50%
suspicious documents		medium (20%-50%)	15%	medium (10-100 pp.)	35%
– with plagiarism	25%	much (50%-80%)	18%	long (100-1000 pp.)	15%
– without plagiarism	25%	entirely (>80%)	10%		

Plagiarism Case Statistics			
Obfuscation		Case Length	
none	18%	short (<150 words)	35%
paraphrasing		medium (150-1150 words)	38%
– automatic (low)	32%	long (>1150 words)	27%
– automatic (high)	31%		
– manual	8%		
translation ({de, es} to en)			
– automatic	10%		
– automatic + manual correction	1%		

Table 4.11: Statistics of the intrinsic partition of the PAN-PC-11, containing 4,753 documents and 11,443 plagiarism cases.

Document Statistics					
Document Purpose		Plagiarism per Document		Document Length	
suspicious documents		hardly (5%-20%)	84%	short (1-10 pp.)	49%
– with plagiarism	50%	medium (20%-40%)	15%	medium (10-100 pp.)	39%
– without plagiarism	50%	much (40%-50%)	1%	long (100-1000 pp.)	12%

Plagiarism Case Statistics			
Obfuscation		Case Length	
none	87%	short (<150 words)	47%
translation ({de, es} to en)		medium (150-1150 words)	39%
– automatic	11%	long (>1150 words)	14%
– automatic + manual correction	2%		

In the external partition, only 2% of the borrowed text fragments are not modified at all. The rest 98% are paraphrased or translated. As a result, this corpus is even more challenging than the previous two.

4.2.3.6 Potential Future Improvements to the PAN-PC Corpora

Whereas the PAN-PC corpora represent a cutting edge effort to massify the development of plagiarism detection models and their objective evaluation, improvement is still necessary. We identify three directions that can be approached in the near future.

Better automatic obfuscation models are necessary in order to make the corpora more realistic. Random shuffling of words remain being a considered model and the text

Table 4.12: Statistics of the external partition of the PAN-PC-11, containing 22,186 documents and 49,621 plagiarism cases.

Document Statistics					
Document Purpose		Plagiarism per Document		Document Length	
source documents	50%	hardly (5%-20%)	46%	short (1-10 pp.)	49%
suspicious documents		medium (20%-50%)	15%	medium (10-100 pp.)	36%
– with plagiarism	25%	much (50%-80%)	25%	long (100-1000 pp.)	15%
– without plagiarism	25%	entirely (>80%)	14%		

Plagiarism Case Statistics			
Obfuscation		Case Length	
none	2%	short (<150 words)	32%
paraphrasing		medium (150-1150 words)	38%
– automatic (low)	40%	long (>1150 words)	30%
– automatic (high)	39%		
– manual	9%		
translation ({de, es} to en)			
– automatic	9%		
– automatic + manual correction	1%		

Table 4.13: Length statistics of the PAN-PC corpora. The figures are shown for the suspicious and source partitions. Global is the combination of both. The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{chars}|$ total number of characters, $|D_{tokens}|$ total number of tokens, $|d_{chars}|$ mean characters per file, $|d_{tokens}|$ mean tokens per file. M stands for 1×10^6 G stands for 1×10^9.

	$\|D\|$	$\|D_{chars}\|$	$\|D_{tokens}\|$	$\|d_{chars}\|$	$\|d_{tokens}\|$
PAN-PC-09					
Suspicious	14,429	2.5 G	491 M	$170,421 \pm 298,135$	$34,023 \pm 59,284$
Source	14,428	3.1 G	648 M	$214,733 \pm 269,950$	$44,882 \pm 56,347$
Global	28,857	5.6 G	1.1 G		
PAN-PC-10					
Suspicious	15,925	3.4 G	698 M	$211,950 \pm 264,009$	$43,801 \pm 54,579$
Source	11,148	1.8 G	348 M	$157,883 \pm 269281$	$31,235 \pm 53087$
Global	27,073	5.1 G	1.0 G		
PAN-PC-11					
Suspicious	11,093	1.7 G	351 M	$152,137 \pm 207,928$	$31,607 \pm 43,447$
Source	11,093	2.5 G	497 M	$224,463 \pm 343,848$	$44,798 \pm 68,508$
Global	22,182	4.2 G	848 M		

it generates is non-readable. More advanced models manage to generate only a few kinds of paraphrase operations. We consider that looking at the area of paraphrases generation is necessary (Barzilay and Lee, 2003). For instance, models based on (monolingual) machine translation (Quirck, Brockett, and Dolan, 2004) could be worth considering. Yet another option is exploiting writing assistance tools (Potthast, Trenkmann, and Stein, 2010b) in order to look for likely co-occurrences that might well be interchangeable within a given context.

Making intrinsic cases doable to be better detected can be made by decreasing the number of cases and their extent in single documents. Whereas it could be not "strictly" realistic, models for intrinsic plagiarism detection are not aimed at detecting cases where 50% of the document is plagiarised, and from different sources. In order to encourage the development of better models, the problem could be simplified in PAN 2012.

Relationship between a fragment and its context is necessary as well. As seen in Chapter 5, the first step when looking for plagiarism is, given a suspicious document d_q, retrieving the most related documents from a collection. In order to get good results from this step, the plagiarised fragment $s \in d_q$ should be on the same (or highly) similar topic than its context. A preliminary effort in this direction was made in the PAN-PC-10, but more has to be done. Moreover, this has to be done if a plagiarism detection models that actually performs document level IR as a first stage wants to be used (cf. Section 5).

Inclusion of more realistic manual cases An open question, even for the manually generated cases, is at what extent they represent actual cases of plagiarism. In general, volunteers are instructed to "simulate plagiarism", but they are not immersed in the common environment of a plagiarist (e.g. time pressures). With the mechanism applied during the PAN-PC-10 and -11 (composing the cases through Mechanical Turk), such environment is impossible to mimic.[26]

4.2.4 Short Plagiarised Answers Corpus

The short plagiarised answers corpus was created aiming at simulating the "types of plagiarism practised by students in an academic setting" (Clough and Stevenson, 2011). Every case it contains was manually generated. In order to do that, five definitional questions on computer science were provided to a group of volunteers. The volunteers were asked to simulate plagiarised and non-plagiarised answers to the different questions. The source of the plagiarised answers was a set of Wikipedia articles on the implied topics.[27]

Three levels of rewriting were requested: near copy, light revision, and heavy revision. Additionally, non-plagiarised answers were requested as well. These levels seem to be inspired by a previous study on rewriting strategies of language learners. Keck (2006, p. 268) considers a paraphrase taxonomy with four types, depending on the amount of words the source and the rewritten text have: (*i*) near copy (50% or more words are shared), (*ii*) minimal revision (20-49% of words are shared), (*iii*) moderate revision (1-19% of words are shared), and (*iv*) substantial revision (no words are shared at all).[28]

[26]This important issue was pointed out by one of the reviewers.

[27]As explained by Clough and Stevenson (2011), one of the advantages of considering Wikipedia articles as source is the possibility of creating cross-language cases by taking advantage of Wikipedia's multilingual nature. Cf. Section 4.2.5 for a cross-language corpus generated over the same principles of this corpus.

[28]This paraphrases typology is much more superficial than the one we consider in Chapter 8 that includes more possible cases if paraphrasing.

Table 4.14: Questions used to generate the cases of the short plagiarised answers corpus (Clough and Stevenson, 2011, p. 10).

(a)	What is inheritance in object oriented programming?
(b)	Explain the PageRank algorithm that is used by the Google search engine.
(c)	Explain the vector space model for Information Retrieval.
(d)	Explain Bayes Theorem from probability theory.
(e)	What is dynamic programming?

Volunteers were instructed to generate short answers (between 200 and 300 words) from the five questions in Table 4.14. No instructions were given about which parts of the article to copy. Further instructions to generate each kind of answer were provided. The instructions to generate plagiarised answers are the following (Clough and Stevenson, 2011, p. 11):

Near copy. Answer the question by cut & pasting from the relevant Wikipedia article.

Light revision. Base your answer on text found in the Wikipedia article. Alter the text in some basic ways including substituting words and phrases with synonyms and altering the grammatical structure (i.e., paraphrasing). Do not radically alter the order of information found in sentences.

Heavy revision. Base your answer on the relevant Wikipedia article. Rephrase the text to generate an answer with the same meaning as the source text, but expressed using different words and structure. This could include splitting source sentences into one or more individual sentences, or combining more than one source sentence into a single one (no constraints were placed on how the text could be altered).

In order to generate non-plagiarised answers, volunteers were provided with additional learning materials. The instructions were:

Non-plagiarism. Answer the question by considering the learning materials provided or some other material (lecture notes, sections from textbooks, etc.). Read these materials and then attempt to answer the question using your own knowledge. You could look at other materials to answer the question but do not look at Wikipedia.

Note that the last strategy still generates a re-used answer. Nevertheless, as the source is not provided in the corpus, from an experimental point of view, they can be safely considered originals. Each volunteer was asked to answer the five questions once, using a different strategy each time. A total of 95 answers were obtained from nineteen volunteers. This fine-grained annotation allows for analysing the capabilities of a model on detecting different kinds of text re-use; extending plagiarism detection into a multi-class problem. For instance, Chong, Specia, and Mitkov (2010) discriminate the three different kinds of re-use and original fragments within this corpus on the basis of a machine learning approach. However, they found that this task is very hard, and a binary classification (plagiarised versus original) can be performed.

Statistics of the corpus are included in Table 4.15. It is worth noting that light and heavy revisions tend to be shorter than cut & paste plagiarism (around 235 tokens with

Table 4.15: Short plagiarised answers corpus statistics. Figures shown for the source (Wikipedia) articles as well as for the plagiarised and original answers. The column headers stand for: $|D|$ number of documents in the corpus partition, $|D_{tokens}|$ total number of tokens, $|D_{types}|$ total number of types. $|d_{tokens}|$ mean tokens per file, and $|d_{types}|$ mean types per file.

	$\|D\|$	$\|D_{tokens}\|$	$\|D_{types}\|$	$\|d_{tokens}\|$	$\|d_{types}\|$
Wikipedia articles	5	2,164	683	433 ± 126	200 ± 45
Plagiarism suspicious	95	21,914	2,284	231 ± 72	119 ± 30
– Near copy	19	4,933	872	260 ± 85	131 ± 33
– Light revision	19	4,414	830	232 ± 70	122 ± 29
– Heavy revision	19	4,544	938	239 ± 61	122 ± 27
– Non-plagiarised	38	8,023	1,525	211 ± 65	110 ± 29

respect to 260). This may resemble the fact that humans tend to summarise contents when rewriting them (this trend is further discussed from a paraphrasing point of view in Chapter 8).

4.2.5 CL!TR 2011 Corpus

The Cross-Language !ndian Text Re-use corpus is the first large-scale corpus for analysis of *cross-language* text re-use (Barrón-Cedeño *et al.*, 2011).[29] On the contrary to the PAN-PC corpora, as in the case of the short plagiarised answers corpus, every case it contains was manually generated. The potentially re-used documents are on computer science (questions are the same as those listed in Table 4.14) and tourism, particularly from Incredible !ndia (questions are those included in Table 4.16). The same strategies described by Clough and Stevenson (2011) were used when generating this corpus (cf. Section 4.2.4). Specific instructions were provided that volunteers had to simulate the situation where plagiarism occurs across languages. The approach was adapted to create a cross-language version, though: forty volunteers (most of them Hindi native speakers; none of them English native speaker) were provided with source texts in English and asked to answer the questions in Hindi.[30] Volunteers were instructed to generate the translated answers as follows:

Near copy. To translate the text with a machine translator and paste the result without further modification.

Light revision. To (optionally) translate the text with a machine translator and do simple editing, such as Hindi grammar correction and simple lexical substitutions.

Heavy revision. To translate the text manually (automatic translation was forbidden), and perform as many grammatical, lexical, and structural changes as possible.

Not re-used. To study another learning material (considering Wikipedia contents was not allowed), understand the topic and write the answers "in their own words" in Hindi.

[29] It is freely available in `http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html`

[30] A similar approach was followed by Somers, Gaspari, and Niño (2006) to build a corpus that aimed at simulating misuse of machine translation by language students (cf. Section 6.2.2).

Table 4.16: Questions used to generate the tourism-related cases of the CL!TR 2011 corpus.

(a)	Where is Agra and describe the important places to visit in Agra?
(b)	Fatepur Sikri has a history. Describe the history behind Fatepur Sikri.
(c)	There are many popular river fronts in Varanasi. State the importance and the myth associated with each of the river fronts.
(d)	Give a brief description about Madurai Meenakshi temple.
(e)	What are the contributions of John Sullivan in developing Ooty into Queen of Hill Stations?
(f)	What are the geographical features of Ladakh? Which are the most important places to visit in Ladakh?
(g)	Give a brief description about major attractions in Kanyakumari.
(h)	Give a detailed description about interesting places in Kodaikanal.
(i)	Kashmir is called the "Paradise of Earth". How do you substantiate this by describing the various places of interest?
(j)	Describe the main attraction of Ranthambore town of Rajasthan.

Table 4.17: CL!TR 2011 corpus statistics. Figures shown for the two sets: D_{en} and D_{hi}. The headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{tokens}|$ total number of tokens, $|D_{types}|$ total number of types. k= thousand, M = million.

Partition	$\|D\|$	$\|D_{tokens}\|$	$\|D_{types}\|$
D_{hi}	388	216 k	5 k
D_{en}	5,032	9.3 M	644 k

The CL!TR corpus includes a set of potentially re-used documents written in Hindi, D_{hi}, and a set of potential source documents written in English, D_{en}. The documents in D_{hi} are likely to be re-used from a document in D_{en}. D_{hi} includes a total of 388 documents. D_{en} includes a total of 5,032 Wikipedia articles (this is yet another important difference respect to the plagiarised answers corpus and the PAN-PC corpora: around 5,000 randomly selected Wikipedia articles were added to the collection in order to make the task of re-use detection more realistic). Some statistics are shown in Table 4.17.

The corpus is divided in training and test partitions. In both partitions the collection of Wikipedia articles (D_{en}) is the same one. The collection D_{hi} is divided into two subcollections: (i) training, composed of 198 documents; and (ii) test, composed of 190 documents. The distribution of simulated-plagiarism and original documents in D_{hi} is shown in Table 4.18. Although its size has to be increased, the CL!TR corpus was the first attempt of a manually created cross-language text re-use collection of cases based on Wikipedia. This corpus was used for the PAN@FIRE: Cross-Language !ndian Text Re-use detection competition. An overview of such a challenge is included in Section 9.5.

Table 4.18: CL!TR 2011 documents distribution. The re-use suspicion documents are included for the training and test partition. For both partitions the set of Wikipedia articles is the same.

Training partition		**Test Partition**	
Re-used	130	Re-used	146
– Light revision	30	– Light revision	69
– Heavy revision	55	– Heavy revision	43
– Exact copy	45	– Exact copy	34
Original	68	Original	44
Total	198	Total	190

System	Actual relevant	Actual ¬ relevant
selected	tp	fp
¬ selected	fn	tn

Table 4.19: Target and decision contingency matrix (borrowed from Manning and Schütze (2002, p. 268)).

4.3 Evaluation Metrics

Now we discuss the different metrics for quantitatively evaluating text re-use and plagiarism detection. As text re-use and plagiarism detection (as well as co-derivatives detection), can be considered as IR tasks, "traditional" IR metrics can be considered for their evaluation. The existence of standard, generally accepted, evaluation measures for these specific tasks have been a gap for long time. This has caused, together with other factors, difficulties in the direct comparison of different models.[31] This section surveys some of the evaluation metrics used in these tasks. In order to do that, we call the collection of k retrieved documents r_q (the documents $r$etrieved after $q$uery q).

4.3.1 Recall, Precision, and *F*-measure

In a retrieval task, the objective is twofold: (i) we would like to retrieve the most of documents considered as relevant for an information necessity; and (ii) we would like to retrieve the least of documents that are considered non-relevant. Two classical measures in IR which aim at estimating how well these objectives are achieved are the well-known recall (rec) and precision ($prec$). In order to define them, we somehow follow the explanation of Manning and Schütze (2002, p. 268). First of all, we define the set of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) as in the contingency matrix of Table 4.19. The four sets are as follows:

tp is the set of relevant objects, and the system selects them as such,
tn is the set of irrelevant objects, and the system selects them as such,
fp is the set of irrelevant objects, but the system selects them as relevant, and
fn is the set of relevant objects, but the system selects them as irrelevant.

Note that tp and tn represent the hits of the model; fp and fn represent the errors. In terms of these four sets, recall can be defined as follows:

$$rec = \frac{tp}{tp + fn} \ , \tag{4.1}$$

i.e., the proportion of objects that are properly identified as relevant respect to the total number of relevant objects. Similarly, precision is defined as:

$$prec = \frac{tp}{tp + fp} \ , \tag{4.2}$$

[31] The main other factor is the lack of real plagiarism, annotated, corpora (cf. Section 4.2).

i.e., the proportion of objects properly identified as relevant respect to the total number of identified objects. In our case the objects are indeed documents (fragments). Equations (4.1) and (4.2) can be rewritten, according to Baeza-Yates and Ribeiro-Neto (1999), by considering the following sets. Let R be the set of relevant documents (fragments). Let A be the set of documents (fragments) returned by the retrieval process. Let R_a be the set of documents in the intersection of R and A. Recall and precision are defined as:

$$rec = \frac{|R_a|}{|R|} = \frac{|\text{relevant documents retrieved}|}{|\text{relevant documents}|} \text{ and} \tag{4.3}$$

$$prec = \frac{|R_a|}{|A|} = \frac{|\text{relevant documents retrieved}|}{|\text{documents retrieved}|} \ . \tag{4.4}$$

In summary, *rec* measures the proportion of relevant documents that were accurately retrieved among those in a collection D; *prec* measures the proportion of documents which are actually relevant among those that were retrieved from a collection D. Both *prec* and *rec* are ranged in $[0, 1]$ with 1 representing the best performance. Note that in some cases an estimation of *rec* is nearly impossible. For instance, in the case of Web IR, where the amount of relevant documents for a given query is unknown (cf. (Baeza-Yates and Ribeiro-Neto, 1999, p. 81)). Indeed, as Manning, Raghavan, and Schütze (2008, Ch. 8) consider, "what matters is rather how many good results there are on the first page or even the first three pages". As a result, only the top k retrieved documents can be considered for evaluation purposes, resulting in "precision at k" (*prec*@k) and, if the required information is available, "recall at k" (*rec*@k). In general, a low value is selected, such as $k = 10$.

These measures can be combined into the harmonic mean of precision and recall (Baeza-Yates and Ribeiro-Neto, 1999, p. 82): the so called F-measure (this is the known as F_1-measure). Whereas different variants of this measure exist, giving more or less relevance to *prec* and *rec*, we opt for using the harmonic mean, defined as:

$$F = 2 \, \frac{prec \cdot rec}{rec + prec} \ . \tag{4.5}$$

The range of F-measure is $[0, 1]$. Note that, *rec* (*prec*) can be perfect ($= 1$) by selecting every document as relevant (irrelevant) (Manning and Schütze, 2002, p. 269). This is why considering a combination of both measures offers a better picture of an obtained result.

We used *prec*, *rec*, and F-measure in different experiments on monolingual journalistic text-reuse (cf. Sections 5.2 and 5.3), cross-language text re-use detection (cf. Section 6.4), and detection of mono- and cross-language co-derivatives over Wikipedia (cf. Section 9.2).

4.3.2 Highest False Match and Separation

Plenty of other evaluation measures exist, but there are a few of them that were originally proposed to evaluate the detection of a kind of text re-use: *versioning* (i.e., co-

derivation (Hoad and Zobel, 2003)). This is the case of the two measures here described: highest false match (HFM) and separation (sep) (Hoad and Zobel, 2003). Rather than simply determining whether a relevant document was properly retrieved, these measures aim at estimating the distance of the correctly retrieved documents in r_q with respect to those incorrectly retrieved. This is a relevant factor when trying to estimate how likely is that the model will make a mistake, as we estimate how good the relevant and irrelevant documents are differentiated in the final ranking. In order to calculate these measures, one condition has to be met: every relevant document must be included in r_q.

Given the ranking of documents r_q, the maximum similarity value s^* is defined as:

$$s^* = \max_{d \in r_q} sim(d_q, d) \ , \tag{4.6}$$

i.e., the maximum similarity value between d_q and any document in D. We also define d^- to be the highest ranked irrelevant document concerning d_q, i.e.,

$$d^- = \max_{d \in r_q^\times} sim(d_q, d) \ , \tag{4.7}$$

where $r_q^\times \in r_q$ represents the subset of irrelevant documents retrieved. Moreover, we define d^+ as the lowest ranked document in r_q that is considered relevant to d_q, i.e.,

$$d^+ = \min_{d \in r_q^+} sim(d_q, d) \ , \tag{4.8}$$

where, as expected, $r_q^+ \in r_q$ represents the subset of relevant documents retrieved. The final computation required is that of the *lowest true match*, computed as:

$$LTM = 100 \cdot sim(d^+, d_q)/s^* \ . \tag{4.9}$$

On the basis of these parameters, HFM is defined as the similarity percentage assigned to d^-:

$$HFM = \frac{100 \cdot sim(d^-, d_q)}{s^*} \ , \tag{4.10}$$

and the separation is defined as $sep = LTM - HFM$, which can be simply computed as :

$$sep = \frac{100 \cdot (sim(d^+, d_q) - sim(d^-, d_q))}{s^*} \ . \tag{4.11}$$

It is worth noting that obtaining a value of $sep > 0$ implies that the highest rated documents in r_q are all those related to d_q. Obtaining $sep < 0$ means that other documents were ranked before those relevant to d_q. As mentioned by Hoad and Zobel (2003), a high HFM is acceptable if sep is high, and a low HFM is acceptable if sep is also low. Therefore, considering the ratio of HFM to sep can be used to assess the quality of a retrieval process.

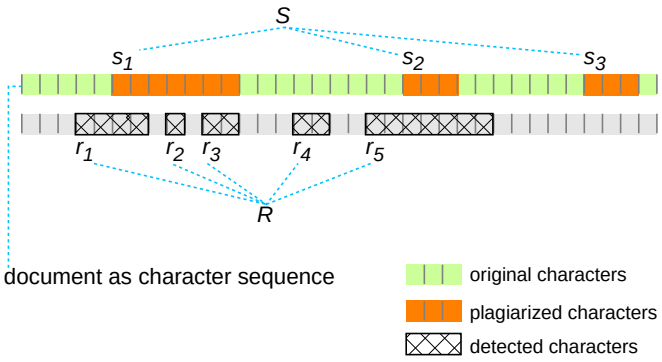


Figure 4.3: A suspicious document d_q as character sequence. It includes actually plagiarised text fragments (S) and detections provided by a given method (R). Squares consist of many characters (borrowed from Potthast *et al.* (2009)).

We used HFM and *sep* measures in experiments on monolingual co-derivatives (versions) detection over Wikipedia articles (cf. Section 9.2).

4.3.3 Especially Fitted Measures for Plagiarism Detection

As aforementioned, *rec* and *prec*, standard evaluation measures in IR, seem to be the option to evaluate this task. However, these measures are designed to evaluate the retrieval of entire documents. Plagiarism detection is different though, since its concern is retrieving specific text fragments: either a plagiarised fragment and its source (external approach) or only the plagiarised fragment (intrinsic approach). Therefore, special measures are required to accurately evaluate the output given by a plagiarism detection model. Obviously a measure that aims at evaluating whether a specific re-used text fragment has been properly identified requires a gold standard where such information is available.[32]

In order to clarify the evaluation process at fragment level, let us consider the example of Fig. 4.3. In this case, d_q includes the plagiarised text fragments $s_{1,\ldots,3}$ (S). An analysis carried out by a given detection method considers that the plagiarised text fragments are $r_{1,\ldots,5}$ (R). The output is said to be perfect if $S \cap R = S \cup R = 1$, i.e., all the plagiarised fragments are accurately retrieved excluding any original fragment. This is not the case in the example (and in most of the cases). Some text fragments are correctly detected while some others are not. Additionally, some original fragments are wrongly detected as plagiarised.

We are interested in evaluating the following three main factors:

1. Original text fragments are not reported as plagiarised;
2. Plagiarised and —if available— source fragments are retrieved; and
3. Plagiarised fragments are not detected over and over again.

[32]The only large scale corpora for plagiarism detection evaluation that includes this information is the series of PAN-PC (cf. Section 4.2.3).

4.3.3.1 Especially Fitted Recall and Precision

The version of *prec* defined to represent the fraction of retrieved fragments that are actually plagiarised is defined as:

$$prec_{PDA}(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|\bigcup_{s\in S}(s\sqcap r)|}{|r|} \ , \tag{4.12}$$

and aims at evaluating how well factor 1 is accomplished. The version of *rec* that represents the fraction of plagiarised fragments that were properly retrieved (factor 2) is defined as:

$$rec_{PDA}(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{|\bigcup_{r\in R}(s\sqcap r)|}{|s|} \ . \tag{4.13}$$

These are tailored versions where every contiguous fragment of plagiarised characters is considered to be a basic retrieval unit. In that way, both *rec* and *prec* express whether a specific plagiarised fragment has been properly recognised and, if available, its source fragment as well. In both equations $\sqcap$ computes the positionally overlapping characters, $\cup_{x\in X}$ defines the union for every $x \in X$, and $|x|$ represents the cardinality of x. Precision and recall can be used to compute the F-Measure as defined in Eq. (4.5).

To clarify the calculation, consider again the situation depicted in Figure 4.3. In order to calculate *prec*, we substitute the corresponding values in Eq. 4.12 as follows:

$$\begin{aligned} prec_{PDA}(S,R) &= \frac{1}{|R|}\cdot\left(\frac{|r_1\sqcap s_1|}{|r_1|}+\frac{|r_2\sqcap s_1|}{|r_2|}+\frac{|r_3\sqcap s_1|}{|r_3|}+\frac{\overset{0}{\overbrace{|\emptyset|}}}{|r_4|}+\frac{|r_5\sqcap s_2|}{|r_5|}\right) \\ &= \frac{1}{5}\cdot\left(\frac{2}{4}+\frac{1}{1}+\frac{2}{2}+\frac{3}{7}\right)=\frac{1}{5}\cdot\left(\frac{41}{14}\right)=0.5857 \ . \end{aligned}$$

In order to calculate *rec*, we substitute the corresponding values in Eq. 4.13 as follows:

$$\begin{aligned} rec_{PDA}(S,R) &= \frac{1}{|S|}\cdot\left(\frac{|(s_1\sqcap r_1)\bigcup(s_1\sqcap r_2)\bigcup(s_1\sqcap r_3)|}{|s_1|}+\frac{|s_2\sqcap r_5|}{|s_2|}++\frac{\overset{0}{\overbrace{\emptyset}}}{|s_3|}\right) \\ &= \frac{1}{3}\cdot\left(\frac{5}{7}+\frac{3}{3}\right)=\frac{1}{3}\cdot\left(\frac{12}{7}\right)=0.5714 \ . \end{aligned}$$

By substituting the obtained values in Eq. (4.5), the F-measure of our example is:

$$F(S,R) = 2\cdot\frac{0.5857\cdot 0.5714}{0.5857+0.5714}=2\cdot\frac{0.3347}{1.1571}=0.5785 \ .$$

4.3.3.2 Granularity

The last factor to consider when evaluating a detection process is factor 3: whether a fragment is detected over and over again or is detected in pieces. This factor is evaluated by means of a novel measure: granularity (Potthast *et al.*, 2010a). This measure punishes those cases where overlapping plagiarised passages are reported. The granularity of R, for a set of plagiarised sections S, is defined by the average size of the existing covers: a detection $r \in R$ belongs to the cover C_s of an $s \in S$ iff s and r overlap. Let $S_R \subseteq S$ denote the set of cases so that for each $s \in S : |Cs| > 0$. The granularity of R given S is defined as:

$$gran(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |C_s|, \tag{4.14}$$

where $S_R = \{s \mid s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$[33] and $C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$[34]. The domain of the granularity is $[1, |R|]$, where 1 marks the desirable one-to-one correspondence between R and S, and $|R|$ marks the worst case, where a single $s \in S$ is detected over an over again.

Going back to the example of Figure 4.3, the granularity would be calculated as follows:

$$gran(S, R) \quad = \quad \frac{1}{|S_R|} \sum_{s \in S_R} |C_s| = \frac{1}{2} \cdot (3 + 1) = 2 \ .$$

4.3.3.3 Plagdet

The especial versions of precision and recall and granularity are combined into an overall value known as *plagdet* (*plag*iarism *det*ection):

$$plagdet(S, R) = \frac{F}{\log_2(1 + gran)} \ , \tag{4.15}$$

Following with our example, we can substitute the obtained values of F-measure and granularity into Eq. (4.15) to obtain the final performance evaluation:

$$plagdet(S, R) \quad = \quad \frac{0.5785}{\log_2(1 + 2)} = \frac{0.5785}{1.5850} = 0.3650 \ .$$

[33]This can be read as the set of elements s in S such that an r in R exists for which the intersection between s and r is not empty.

[34]This can be read as the set of elements r in R, such that the intersection between s and r is not empty.

Note that a logarithm is applied to the granularity in order to decrease its impact in the final calculation. Moreover, if the fragments are detected without overlapping, i.e., $gran = 1$, the F-measure is not modified at all.

These measures are used when evaluating the results of the International Competition on Plagiarism Detection and, in general, every approach that aims at detecting plagiarism over the PAN-PC corpora (cf. Chapter 7).

4.4 Chapter Summary

In this chapter we described the two key factors in an evaluation framework for text re-use and plagiarism detection: corpora and evaluation measures.

Corpora of journalistic text re-use and Wikipedia co-derivation, with real, human-made, cases of re-use were discussed, including the METER and co-derivatives corpus. Corpora of simulated re-use and, in particular, simulated plagiarism were also described: the first large scale corpora for the study of plagiarism detection —the PAN-PC—, a corpus of manually created cases from Wikipedia with different levels of paraphrasing —the short plagiarised answers corpus—, and a corpus of manually created cross-language cases from Wikipedia with automatic and further paraphrased translations —the CL!TR corpus. Their annotation, size, nature and purpose were discussed. These corpora have been used to investigate the detection of text re-use and plagiarism.

The second part of the chapter contains an overview of state of the art evaluation metrics for automatic text re-use and plagiarism detection. Standard evaluation metrics in information retrieval (precision, recall, and F-measure) were recalled. Afterwards, a couple of metrics specially designed to evaluate text re-use detection, in particular co-derivatives were discussed: highest false match and separation.

Finally, three evaluation measures specially designed for automatic plagiarism detection, particularly when the aim is detecting specific text fragments, were discussed: two fitted versions of precision and recall, the granularity measure, and their combination into the plagdet measure. These measures were designed for the evaluation of the plagiarism detection systems participating at the PAN competitions on plagiarism detection.

Related publications:

- Barrón-Cedeño, Rosso, Lalitha Devi, Clough, and Stevenson (2011)
- Potthast, Stein, Barrón-Cedeño, and Rosso (2010a)
- Barrón-Cedeño, Potthast, Rosso, Stein, and Eiselt (2010a)
- Barrón-Cedeño and Rosso (2010)
- Barrón-Cedeño, Eiselt, and Rosso (2009a)

Chapter 5

Monolingual Detection of Text Re-Use and Plagiarism

> There is no more infuriantingly tiresome job in academic life than searching for that needle in the book stacks, the elusive source you *know* young Heather must have copied but that you simply can't find. And there you are spending three hours of a Saturday afternoon in a fruitless search while Heather's back in Wyatt Hall blowing a joint and watching a Billy Idol video.
>
> Thomas Mallon

In this chapter we discuss models for automatic text re-use and plagiarism detection. Though the models are indeed re-use detectors, and the final decision is taken by the expert, the literature refers to the problem as "plagiarism detection". Therefore, we use both terms as "pseudo-synonyms".[1]

In order to detect cases of re-use, a key factor is selecting a good set of text features through which re-used fragments can be discriminated from originals. Clough (2000; 2003) identified a set of features that may trigger suspicion. A system that aims at detecting re-use automatically may be based on (some of) them. We divide such features into three groups: *intrinsic*, *external*, and *hybrid*. In order to describe them, we consider d_q to be the query document (i.e., the suspicious document), $\mathcal{A}$ its claimed author, and $\mathcal{A}'$ a different author.

The intrinsic features are the following:

1. Changes of vocabulary. Inconsistencies within the written text itself; e.g. changes in vocabulary, style or quality may imply that d_q contains text fragments written by $\mathcal{A}'$ (coming from an external resource).
2. Incoherent text. If the contents in d_q do not flow consistently or smoothly, it could include external text fragments (though it could just be the result of $\mathcal{A}$'s poor authoring skills or multi-authorship).

[1] As previously mentioned (cf. Section 2.2) strictly plagiarism is considered a hyponym of text re-use as it represents a specific case of the broader phenomenon.

3. Preference for the use of either long or short sentences. Authors have a certain tendency to write long or short sentences.
4. Readability of written text. It is unlikely that $\mathcal{A}$ and $\mathcal{A}'$ would write with the same level of complexity and share the same readability score.
5. Dangling references. References appearing in the text but not in the bibliography (or vice versa), may indicate cut & paste re-reuse with reference omission, or plagiarism of secondary sources (cf. Section 2.2.2).
6. Inconsistent references. Use of inconsistent referencing in the bibliography, suggesting again cut & paste.

These features clearly point at style and complexity measures to determine whether $\mathcal{A}$ has actually written a suspicious text (cf. Section 3.4). The external features are the following:

7. Amount of similarity between texts. Documents on the same topic are expected to share some contents (e.g. proper names, terms). Nevertheless, it is very unlikely that they share large amounts of the same or similar text, or even matching words, if written independently. This is even more relevant if they have long sequences of tokens in common (cf. page 23).
8. Distribution of words. It is unlikely that the distribution of word usage throughout independent texts would be the same.
9. Common spelling mistakes. It is very unlikely that $\mathcal{A}$ and $\mathcal{A}'$ make the same (number of) spelling mistakes.
10. Common punctuation. It is unlikely that $\mathcal{A}$ and $\mathcal{A}'$ would use punctuation in exactly the same manner.
11. Common syntactic structure of the text. Suspicion should be triggered if two texts share the exact same syntactic structure.
12. Dependence on certain words and phrases. $\mathcal{A}'$ may prefer using particular words or phrases. Consistent use of these words and phrases in a text written by $\mathcal{A}$ may indicate a potential case of re-use.
13. Frequency of words. It is unlikely that two independent texts share the same vocabulary distribution.

These features point to the comparison of the suspicious text to a collection of reference documents (cf. Section 3.3). Lastly, the hybrid features are the following:

14. Uses of vocabulary. Analysing the vocabulary in d_q, with respect to material previously produced by $\mathcal{A}$. A large amount of new vocabulary, or technical vocabulary beyond that expected from $\mathcal{A}$, triggers suspicion.
15. Unexpected improvement. A large improvement in writing style compared to previous submitted work is unexpected.

Intrinsic features have to do with the evolution of text within d_q. Unexpected differences among fragments in d_q may be caused by the insertion of text originally written in an external resource. Features 1 to 6 aim at analysing d_q's contents in order to identify

outliers. Section 5.1.1 overviews how intrinsic detection models exploit some of these features to detect cases of re-use, without considering other document than d_q.

External features are those for which texts written by $\mathcal{A}'$ are compared to d_q, looking for borrowed contents. Features 7 to 13 aim at comparing d_q to d, a document written by another author, looking for unexpected similar fragments, mistakes, or structure. Section 5.1.2 overviews how external detection models exploit some of these features to detect cases of re-use.

We consider hybrid features as a combination between intrinsic and external: d_q is compared to a document d, but the initial hypothesis is that $\mathcal{A}$ wrote both. This approach is only possible if previous material written by $\mathcal{A}$, the claimed author of d_q, is at hand. The idea of features 14 and 15 is determining whether d_q actually wrote the text in d_q by considering his authoring background, his "fingerprint" (cf. Section 3.4). Forensic linguists seem to exploit these features very often for both tasks: plagiarism detection and authorship identification (cf. Section 2.3). Nevertheless, to the best of our knowledge, no automatic models for plagiarism detection have been created exploiting them. This is a problem closer to authorship identification;[2] see Stamatatos (2009a) and Argamon and Juola (2011) for an overview of state of the art authorship identification approaches that could be exploited when detecting plagiarism from this point of view.

As an aside note, the pertinence of some of the mentioned features is subjective, even doubtful. Consider, for instance, feature 14. Is not expected that a student —or any other writer— increases her vocabulary and authoring abilities over time?

Maurer *et al.* (2006) identify three main strategies text re-use can be detected with:

- Applying style analysis to determine whether d_q has been actually written by the claimed author;
- Comparing d_q against a collection of potential source documents D; and
- Identifying a characteristic text fragment in d_q and querying it into a search engine.

The former schema regards to exploiting the intrinsic features before described, where d_q is analysed isolated. The second and third schemas regard to exploiting the external features before described, where d_q is analysed together with a collection D of documents.

Intrinsic analysis is a binary classification task. The contents in d_q are analysed seeking inconsistencies across the same document (e.g. changes in vocabulary and complexity, bad flowing, etc). The aim is performing a supervised or unsupervised classification process of d_q's contents to identify original and borrowed fragments. As a result, the output is a set of plagiarism suspicion text fragments, but no other evidence is provided.

External analysis is a task closer to IR. The contents in d_q are compared to the contents in a document collection D. The objective is identifying text fragments which are more similar than expected. As a result, the evidence expected consists of pairs of text fragments: $s_q \in d_q$ and $s \in d$, where s is suspected to be the source of s_q.

Selecting one of the approaches depends on two main factors: the available resources and the required evidence. Intrinsic analysis has very low requirements: d_q, the suspicious

[2]Indeed, automatic plagiarism detection and authorship attribution are considered "two sides of the same coin" (Stein, Stamatatos, and Koppel, 2008).

document itself. As a result, the analysis can be carried out in short time. Its main weakness, though, is that a suspicious text fragment with no further information might not always be proof enough.[3] External analysis has much higher requirements. It needs a collection of documents to compare d_q against. Building such a collection represents a non trivial task. An option would be considering an open collection, such as the Web. The advantage of external detection is that the evidence provided includes the claimed re-used fragment together with its source. The main disadvantage is the processing required to make the necessary comparisons (and at the end of the day if the source document for a given case of plagiarism is not in the reference collection, the case would go unnoticed).

A combination of both approaches seems reasonable. In the first stage, an intrinsic analysis could identify plagiarism suspicion text fragments. In the second stage, such fragments could be submitted to the external process, looking for their potential source. Nevertheless the quality of state of the art models for intrinsic detection is still far to be reliable (Potthast *et al.*, 2011b). As a result, rather than helping to decrease the load of work for the external stage, it could deprive it of the necessary information.

A person is able to identify potential cases of plagiarism by simply detecting unexpected irregularities through a document. Disruptive changes in style, vocabulary, or complexity are triggers of suspicion. Intrinsic models try to mimic this ability by analysing different text features. Intrinsic plagiarism detection is defined as follows:

> **Intrinsic plagiarism detection.** Let d_q be a document presumably written by author $\mathcal{A}$. Determine whether all the contents in d_q have been actually written by $\mathcal{A}$. If not, extract those text fragments that could have been written by a different author $\mathcal{A}'$.

As said by Jude Carroll, in this case we are looking for changes from good to bad language (BBC, 2011). The general architecture of intrinsic plagiarism detection is depicted in Fig. 5.1. The typical process can be divided into four steps (Potthast *et al.*, 2011b):

Document chunking. The contents in d_q are split into a set of text chunks, in general, equally lengthened. In most cases a sliding window is used for selecting the chunks with a high level of overlapping (Stamatatos, 2009b).

Retrieval model. A function is used that maps texts onto feature representations (e.g. those described in Section 3.4), along with a similarity measure to compare them. The retrieval model aims at assessing the similarity among the different text chunks in d_q. Different measures can be used, such as the cosine measure (cf. Section 3.3.1.2) or more robust ones, specially designed for this task (Stamatatos, 2009b).

Outlier detection. At this stage the aim is identifying "chunks of d_q that are noticeably different from the rest" (Potthast *et al.*, 2011b). Different strategies can be applied, from considering the deviation between a chunk characterisation respect to the

[3]Note that a high percentage of professors consider that a disruptive change in style or complexity is reason enough to consider that a document contains plagiarism (cf. Section 2.5.1).

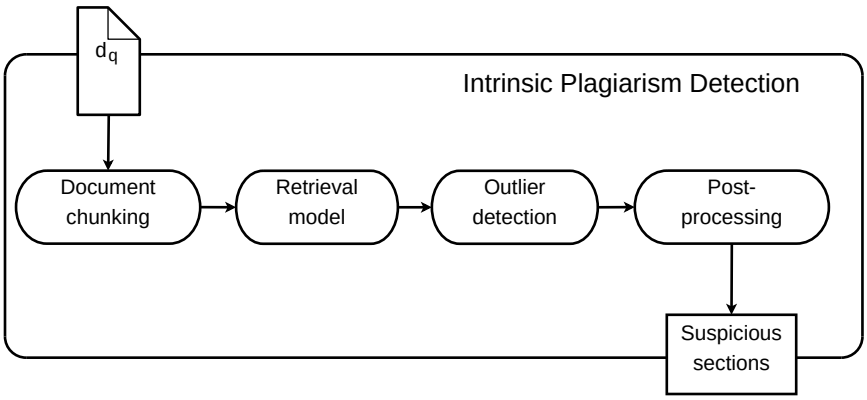


Figure 5.1: General architecture of intrinsic plagiarism detection (derived from Potthast *et al.* (2011b); Stein and Meyer zu Eissen (2007)).

entire d_q's to applying some clustering technique. The aim of the second approach is coming up with two clusters, one of which would presumably contain non-original passages.

Post-processing. Overlaps between neighbouring and closely detected chunks in the text are merged. This step aims at providing a cleaner output to the user.

Respect to external plagiarism detection, Metzler *et al.* (2005) propose the concept of *similarity spectrum*. On the one side, the standard problem of IR is located, finding matches of documents on the basis of topical similarity. On the other side, we have the location of identical documents, duplicates. Both extremes of the spectrum are relatively solved. The open avenues are in between, when trying to differentiate between documents with a particularly high topical similarity, but without any link to each other, and documents that include common chunks, borrowed from the other. The interest for developing models for external plagiarism detection is not new. Lancaster and Culwin (2005) differentiate two settings where external plagiarism detection can be performed: (*i*) in *intra-corporal plagiarism detection*, the plagiarism suspicion and potential source documents are included in the corpus at hand; and (*ii*) in *extra-corporal plagiarism detection*, the source of a plagiarised fragment is outside the collection (e.g. on the Web). The problem of external plagiarism detection can be defined, at document level, as follows:

> **Document level external plagiarism detection.** Let d_q be a suspicious document. Let D be a set of potential source documents. Determine whether d_q contains borrowed text from a specific $d \in D$.

If d_q happens to contain fragments from d and no proper citation is provided, a case of plagiarism may be at hand. At a lower granularity, the fragment level, external plagiarism detection can be defined as follows:

> **Fragment level external plagiarism detection.** Let d_q be a suspicious document. Let D be a set of potential source documents. Determine whether the fragment $s_q \in d_q$ has been borrowed from the fragment $s \in d$ ($d \in D$).

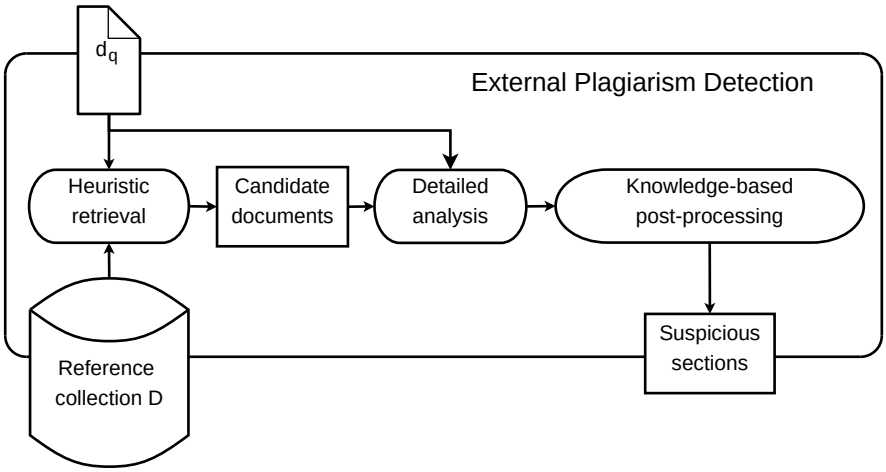


Figure 5.2: General architecture of external plagiarism detection (derived from Stein *et al.* (2007)).

In both cases the expected output consists of a pair $\{s_{plg}, s_{src}\}$; i.e., the plagiarism suspicious fragment —or document— together with its source. In contrast to other IR tasks, where we are interested in measuring the amount of material two documents have in common, plagiarism can only be identified by also measuring the amount of text that differs (Hoad and Zobel, 2003).

An overview of the general architecture of external plagiarism detection is depicted in Fig. 5.2. The typical process can be divided into the following three stages:[4]

Heuristic retrieval. Retrieve a small set of documents D^* such that $D^* \ll D$ and $d \in D^*$ is likely to be the source of the potentially plagiarised text in d_q.

Detailed analysis. d_q is compared (section-wise) to every document $d \in D^*$. Plagiarism suspicion fragments and their potential sources are identified.

Knowledge-based post-processing. Fragments identified as plagiarised which are too short to be considered relevant are discarded. Neighbouring cases identified are merged to compose a single case.[5]

In Section 5.1 we review the literature on automatic plagiarism detection, both intrinsic and external. We have designed a total of three experiments in order to analyse different settings on monolingual detection of text re-use. One of our major concerns is to tackle the identification of modified borrowing (i.e., paraphrasing), and we explore different text representation strategies. Our experiments include a query by example retrieval problem, detection of plagiarised sentences and their source text, and an entire process for automatic plagiarism detection. They are discussed in Sections 5.2 to 5.4.

Key contributions The impact of applying different text pre-processing and characterisation, as well as term weighting models, in the text re-use process is thoroughly

[4]Recently, in the framework of the 2012 edition of the International Competition on Plagiarism Detection, the step "heuristic retrieval" has been renamed as "candidate document retrieval" and "detailed analysis" as "detailed comparison" (cf. `http://pan.webis.de`).

[5]This stage was originally proposed to discard actual cases of re-use which included a proper citation, "hence establish[ing] no plagiarism offence" (Stein *et al.*, 2007, p. 825) (cf. Section 5.1.2.3).

analysed (Section 5.2). A model is proposed that, given a suspicious document, selects a subset of good potential source documents in order to further search for potential cases of re-use (Section 5.4).

5.1 Past Work

As seen in Section 4.1, Potthast *et al.* (2010a) made a compilation of research papers on topics related to detection of co-derivatives, versions, plagiarism, and other kinds of text re-use, finding more than one hundred papers on this issue. Here we only discuss some of the most representative approaches to intrinsic and external plagiarism detection. During the last three years, the trend of development of plagiarism detection models has been highly influenced by the International Competition on Plagiarism Detection. The approaches tried within that framework are discussed in Chapter 7.

5.1.1 Approaches for Intrinsic Plagiarism Detection

Here we describe some of the most successful approaches to intrinsic plagiarism detection.

5.1.1.1 Averaged Word Frequency Class

It seems to be one of the best features for discriminating between original and re-used text fragments within a document. The experiments of Meyer zu Eißen and Stein (2006) analysed documents for plagiarism including, among others, the following text statistics and syntactical features:

1. Averaged word frequency class, which aims at estimating the vocabulary complexity and diversity (cf. Section 3.4.1).
2. Average sentence length, which aims at reflecting how complex the sentences in a document are (cf. Section 3.4.2).
3. Part-of-speech features, aiming at determining the variety in language (cf. Section 3.4.3).
4. Average stopword number, which aims at considering the set of articles, prepositions and other function words in the text (cf. Section 3.4.4).

These features are used to perform a discriminant analysis[6]. The results, reported on an artificial corpus of 450 documents, show that averaged word frequency class is the best discriminating feature. A very similar setting was applied by Zechner, Muhr, Kern, and Granitzer (2009).

In a different paper describing experiments over the same corpus, Meyer zu Eißen *et al.* (2007) report trying with more features, such as the Gunning fog index, Flesch-Kincaid grade level, Honoré's R, and Yule's K (cf. Table 3.2 in page 74). Nevertheless,

[6]A method to find a combination of features which separate two or more classes of objects

the averaged word frequency class showed, once again, to be the most discriminative one. Later in that year, Stein and Meyer zu Eissen (2007) went further, including an authorship identification module to improve the accuracy of the intrinsic detection output.

5.1.1.2 Character *n*-Gram Profiles

They have shown to be a good option as well, due to the simplicity of the model (Stamatatos, 2009b). In this approach, the profile p_d of a document d is a bag of tf-weighted character 3-grams (cf. Section 3.1.3). In order to profile the different fragments $s \in d$, sliding windows of length m and step n are used (Stamatatos proposes $m = 1,000$, $n = 200$). The dissimilarity between every p_s and p_d is then computed on the basis of the *normalised* d_1:

$$nd_1(p_s, p_d) = \frac{\sum_{t \in p_s} \left(\frac{2(tf_{t,p_s} - tf_{t,p_d})}{tf_{t,p_s} + tf_{t,p_d}} \right)^2}{4|p_s|} \quad , \tag{5.1}$$

where $tf_{t,\cdot}$ is the normalised frequency of term t —a character n-gram– in $\cdot$, and $|p_s|$ represents the size of the profile of text s. The possible result is $0 \leq nd_1 \leq 1$, with 0 representing the maximum similarity. Similarly to the containment concept (cf. Section 3.3.1.1), nd_1 is asymmetric, aiming at comparing text fragments of highly different lengths.

If a higher than expected standard deviation exists between p_s and the mean respect to every p_s and p_d, a potential case of plagiarism has been found. Variants of this measure have been used by other researchers, such as Kestemont, Luyckx, and Daelemans (2011) and Oberreuter, L'Huillier, Ríos, and Velásquez (2011).

5.1.1.3 Kolmogorov Complexity Measures

Kolmogorov complexity measures have been applied to intrinsic detection as well. Seaward and Matwin (2009) opt for characterising the sentences in a text on the basis of different text statistics as well as syntactic, and part of speech features (cf. Sections 3.4.1 to 3.4.3). Each feature is used singleton to represent a sentence as a binary string. A representation including nouns versus non-nouns would imply putting a 1 if a word is indeed a noun and 0 otherwise. The same for short versus long words and so on (Seaward and Matwin, 2009, p. 57). They approximate the computation of the Kolmogorov complexity of a string of bits on the basis of the Zlib lossless compressing algorithm (Li and Vitanyi, 1997); a high compression rate denotes high complexity. The result of the chunks' identification, codification, and compression is inserted into a classifier that aims at discriminating between plagiarised and original text fragments.

5.1.1.4 Assumptions and Drawbacks

Some assumptions and drawbacks of intrinsic plagiarism detection can be identified. The first assumption is that finding style and complexity outliers in a document is reason enough to plagiarism suspicion. As discussed in Section 2.5.1, this seems to be truth. Most professors consider that abrupt changes throughout a text are good triggers to suspect borrowing. The second assumption is not so favourable for this approach. It is assumed that, beside the potential borrowed text fragments in d_q, the document has been written by one single author. The reason for this assumption is precisely on the principles this approach is based upon: that style and complexity in the writings of different authors vary. Consequently, it cannot be applied to analyse collaboratively written documents (Maurer *et al.*, 2006).

From our point of view this drawback could not be so serious though. As already discussed, models for automatic plagiarism detection are indeed for text re-use detection. The final decision has to be taken by the expert. The intrinsic approach does not provide other than a suspicion trigger: a text fragment that seems not to fit with its context. Now, let us assume that an ideal model exists that is able to detect style and complexity outliers in a text —no matter who wrote what. The problem would be, rather than detecting a specific borrowed text fragment, detecting the borders between the text fragments written by different authors. Such a model could be useful on two different tasks, depending of the nature of d_q: (i) if d_q is claimed to be written by a single author, potentially borrowed text fragments could be identified, and (ii) if d_q is written by multiple authors, the different sections written by each of them could be identified.

During this research work we have not worked on intrinsic plagiarism detection in depth. Beyond the discussed literature in this section, refer to Stein, Lipka, and Prettenhofer (2011b) for an overview of intrinsic plagiarism detection and Stamatatos (2009a) for an overview of models for authorship attribution, a "sister" task of intrinsic analysis. We have created a Web application called `Stylysis`. It chunks a text into paragraphs and computes many of the features used in intrinsic plagiarism detection: average sentence and word lengths, Gunning fog index, Honoré's R, Yule's K, and Flesch-Kincaid readability test. It is available at `http://memex2.dsic.upv.es:8080/StylisticAnalysis`.

5.1.2 Approaches for External Plagiarism Detection

Here we discuss some models for external plagiarism detection. Most of the literature is focussed on the detailed analysis stage. Indeed, most of them do not consider any other stage. Therefore, we discuss detailed analysis in the following subsection and turn back to the heuristic retrieval stage in Section 5.1.2.2. Post- and pre-processing are discussed in Sections 5.1.2.3 and 5.1.2.4.

5.1.2.1 Detailed Analysis

In this stage, documents are compared in order to determine whether the source of a text fragment contained in d_q is in d. It is here where the concept of *fingerprint* comes

Given d_q and $DB_{\mathcal{H}}$:

break d_q into shingles $d_{q,i}$
for each chunk $d_{q,i}$ in d_q:
    Compute $\mathcal{H}(d_{q,i})$
    if $\exists\, \mathcal{H}(d_{q,i})$ in $\mathcal{H}^*$
        return $d \mid \mathcal{H}(d'_q) \in d$

Figure 5.3: COPS matching algorithm. d_q is a plagiarism suspicious document, $\mathcal{H}$ is a hashing function, and $\mathcal{H}^*$ is a database of hashed shingles previously generated from a collection of documents D.

out. Fingerprinting is an approach proposed to characterise and compare documents. A fingerprint aims at being a representation of document's contents (also known as the document's signature (Monostori, Finkel, Zaslavsky, and Hodász, 2002)). It uses to be a selection of the overlapping strings occurring in the document, which could be a series of characters, words, sentences, etc. (Heintze, 1996) (e.g. Brin, Davis, and Garcia-Molina (1995) consider sentences, whereas Bernstein and Zobel (2004) and Lyon, Barret, and Malcolm (2004) use word n-grams, and Heintze (1996) and Schleimer, Wilkerson, and Aiken (2003) character n-grams). These are the known as shingles, a contiguous sequence of tokens in a document (either characters of tokens). As in humans, the fingerprint should include those characteristics inherent of a specific document which, if present in another document, may imply that they are somehow related.

For efficiency issues, the shingles that compose the fingerprint may be codified with a hash function (Stein and Potthast, 2007), for instance, computed on the basis of the Rabin-Karp hashing algorithm (Karp and Rabin, 1987). As seen in Section 3.1.5, these functions compute a numerical value from a string that is very likely to be unique, i.e., different strings would not generate the same hash value.

COPS and SCAM are probably some of the first efforts to detect duplicates in a collection of texts (Brin *et al.*, 1995; Shivakumar and García-Molina, 1995). The objective of these projects from the University of Stanford was detecting document duplicates at different extents. These systems were designed as a countermeasure to re-use, as people were worried about how easy electronic media made "illegally copying and distributing information".

COPS (*COpy Protection System*) (Brin *et al.*, 1995) was a system for registering any generated document in order to keep control of the places it could be republished in. The proposed registration schema is as follows: (*i*) $\mathcal{A}$ creates a new work d and registers it into a server, (*ii*) d is broken into small shingles (for instance, sentences), and (*iii*) each shingle is stored into a large data base. In order to reduce the spacial and temporal cost, hash values of the shingles are saved rather than the actual text. When a new document d_q has to be analysed, the process represented in Fig. 5.3 is followed. If a shingle in d_q is shared with d, both documents would generate the same hash, resulting in a match. If the selected shingle is long enough to be considered a trigger of suspicion by itself (a sentence, a paragraph, or even an entire document) finding a match could be the output of the process. Otherwise, a similarity can be computed. Brin *et al.* (1995) consider that a low similarity threshold is appropriate for detecting borrowed paragraphs. Larger thresholds are the option when looking for documents with high extents of common text only.

SCAM (Stanford Copy Analysis Mechanism) (Shivakumar and García-Molina, 1995) was proved with different settings, considering the entire vocabulary in the texts, the k most frequent words, sentences, modulo m hash values, etc. They propose the so called *relative frequency model* as well. Instead of representing a document on the basis of its entire contents, SCAM selects a subset of terms which could better reflect the similarity between d_q and d. Such terms are those with similar frequency in the implied documents. This idea results on the word chunking overlap we already discussed in page 67. Experiments over UseNet and a set of websites showed that the measure gave promising results.

In COPS and SCAM the interest is more on the architecture of a system for the registration and control of documents. Indeed this is probably their biggest contribution: an architecture that comprehends from the registration of documents to generate a reference collection, up to the stages necessary to analyse a suspicious document.

Scalable Document Fingerprinting is the title of a seminal paper that contemplated already most of the issues approached in subsequent years (Heintze, 1996). The selected document fingerprints in this approach, as Heintze already calls them, are text sequences between 30 and 45 characters long. Interestingly, he mentions that no right length for the shingles exists and that empirically calculating it is not possible, as it depends on the problem to face and how sensitive the model should be respect to modifications.

Heintze (1996) differentiates between two kinds of fingerprinting: full and selective. In *full fingerprinting*, every sequence of characters in d is considered.[7] In *selective fingerprinting* a sub-sample of d's shingles is chosen to represent it. Selective fingerprinting can be also divided into *fixed size* and *proportional size*. In the former one, a fix amount of sub-strings is selected from d, regardless its size. In the latter one, the number of sub-strings to select is in accordance with the size of d. Rather than randomly selecting a set of strings (an option he mentions), Heintze considers using different heuristics. He selects those shingles in which the first five characters appear with lowest frequency in d. The reasoning behind this decision is that the lower the intra-document frequency of a term is, the lower the extra-document frequency is expected to be. By filtering very likely chunks, the number of false positives retrieved, in general, decreases.[8] When using long shingles, such as in Heintze's proposal, a system practically expects detecting cut & paste re-use only, where the source and borrowed texts are identical. Considering shorter representations is often enough to detect cases of modified re-use, but at the risk of getting a higher amount of false positives.

The works of Shivakumar and García-Molina (1995), Brin *et al.* (1995), and Heintze (1996) can be considered as the seminal works on automatic plagiarism detection of free text. Most of the research work in the years to follow included variations of their principles. For instance, consider the following two fingerprinting approaches.

[7] As stressed by Heintze himself and Lyon *et al.* (2004), in general full fingerprinting representations result in objects larger than the original document!

[8] Moreover, some shingles are discarded in advance from the representation, in particular those that appear in many documents as they could refer to agencies, acknowledgements or any other common phrases.

Given d**:**

$G = \{$sequence of n-grams in $d\}$
Initialize $\mathcal{H}^*_{tmp}$ `//Temporal array of hashes`
for each n-gram $\in G$:
    add $\mathcal{H}(n$-gram$)$ to $\mathcal{H}^*_{tmp}$
// `Winnowing`
Initialize $\mathcal{H}^*$ `//The selected hashes`
for each sliding window $W_\mathcal{H}$ such that $|W_\mathcal{H}| = th$:
    insert $(min(W_\mathcal{H})$, offs$(min(W_\mathcal{H})))$ into $\mathcal{H}^*$
$d_{fing} = \mathcal{H}^*$

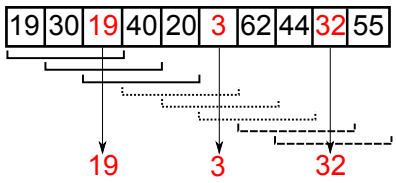


Figure 5.4: Winnowing fingerprinting algorithm. *min* selects the minimum hash value among those in window $W_\mathcal{H}$ (if more than one hash contains the minimum value, the rightmost is selected), offs returns the offset of the selected hash value. In the graphic representation $th = 3$. The first three windows select the hash 19, the dotted windows select 3, and the dashed windows select 32.

Winnowing is a selective fingerprinting model (Schleimer *et al.*, 2003). As in the case of Heintze (1996), the shingles selected are character n-grams. Additionally to the traditional pre-processing operations, such as case folding and diacritics elimination (cf. Section 3.1.1), spaces and punctuation marks are discarded, i.e., d is considered to be a long string of characters s_d in the alphabet $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$. The representative shingles are selected on the basis of a sliding window mechanism with two parameters:

The noise threshold n**.** It defines the level of the n-grams (as already seen, the larger the value of n, the more sensible the method becomes with respect to modifications); and

The guarantee threshold th**.** It defines the length of the sliding window $W_\mathcal{H}$.

Hash values are computed for every n-gram in the document string. Afterwards, a window of length th is slid throughout the n-gram sequences, always picking the lowest hash value to be part of the fingerprint.[9] If $W_\mathcal{H}$ contains two equal lowest values, the rightmost one is selected. The Winnowing fingerprinting process together with a graphical representation of the sampling process is sketched in Fig. 5.4. As there shown, it is expected that many different windows will select the same hash value to be part of the fingerprinting. As a result, $|\mathcal{H}^*| \ll |W_\mathcal{H}|$.

Both n and th $(n \leq th)$ can be empirically defined. Schleimer *et al.* (2003) use $n = 50$ and $t = 100$ with promising results. This is equivalent to considering shingles of around twelve words. Matches between documents d_q and d can be then sought in order to get duplicated text fragments or a similarity value between the documents could be estimated. Given the flat distribution that results of considering high levels of n (it is very likely that every shingle and its corresponding hash will appear only once in a document) a Boolean measure could be used, such as the Jaccard coefficient (Eq. (3.10) at page 64).

[9] This idea of taking the lowest numerical values had been mentioned already by Heintze (1996), but he did not go further.

Given a collection of documents D:

```
for each d ∈ D:
    for each 1-gram g ∈ d
        H*_1 ← H(g)

for n = {2, ..., l}:
    for each d ∈ D:
        for each n-gram in g ∈ d
            if cnt(H*_{n-1}, g_{[0,n-1]}) == cnt(H*_{n-1}, g_{[1,n]}) == 2:
                H*_n ← H(g)
```

Figure 5.5: `SPEX` algorithm. $\mathcal{H}_n^*$ is the hash table for word n-grams. Each n-gram in $\mathcal{H}_n^*$ has an associated counter; *cnt* returns the counter for a given hash.

Sorokina, Gehrke, Warner, and Ginsparg (2006) adapted Winnowing when trying to detect plagiarism in `arXiv`[10] in two different ways. Firstly, the n-grams are considered at word rather than character level. Secondly, instead of using a sliding window, the windows become the sentences. One n-gram is selected per sentence to be part of d's fingerprint. Yet another adaptation is performed by Kent and Salim (2009). In their proposal each sentence in d is represented by its three least-frequent character 4-grams (considering the entire document), which are merged to compose the sentence's representation in the document's fingerprint.

SPEX algorithm represents a kind of full fingerprinting approach, this time considering word n-grams. The idea behind it is that "if any sub-chunk of any chunk can be shown to be unique, then the chunk in its entirety must be unique" (Bernstein and Zobel, 2004). That is, if a chunk '$t_1\ t_2$' is unique, the chunk '$t_1\ t_2\ t_3$' is going to be unique as well. This principle can be easily applied to a collection D of documents.

In order to perform a fast "common-chunks" finder, `SPEX` discards those word n-grams that occur in one single document only. The only parameter in this algorithm is l, the target shingle length (for instance, Bernstein and Zobel (2004) propose using $l = 8$). Meanwhile the value of n increases, less and less common chunks would be found to be shared between $(d_i, d_j) \in D$. The algorithm to identify those chunks appearing in more than one document $d \in D$ is depicted in Fig. 5.5. The first step is generating a hash table $\mathcal{H}_1^*$ containing the 1-grams in every $d \in D$. Every entry in the hash table has an associated counter: 1 if it appeared in one document or 2 if it appeared in at least two documents. In the second step a new hash table $\mathcal{H}_n^*$ is iteratively generated by considering $\mathcal{H}_{n-1}^*$. The hash of an n-gram is inserted into $\mathcal{H}_n^*$ iff its two (n-1)-grams exist in $\mathcal{H}_{n-1}^*$ with a counter of 2.

At the end of the process, those entries in $\mathcal{H}_l^*$ that are marked to appear in two documents represent the set of potential borrowings. By simply inspecting these cases it is likely that any case of text re-use could be found. If a similarity between documents d and d_q is still necessary, it can be computed by means, once again, of the Jaccard coefficient. Note that a very low value of this measure may well indicate a high probability of derivation. Among the different measures Bernstein and Zobel (2004) propose, the overlap coefficient is included (Eq. (3.12) in page 65).

[10] `arXiv` is an e-print service in different fields of science (cf. `http://arxiv.org`).

Table 5.1: Percentage of common n-grams in texts written by the same authors and on related topics. Average tokens per document: 3,700 (Barrón-Cedeño and Rosso, 2009a).

Documents	1-grams	2-grams	3-grams	4-grams
2	0.1692	0.1125	0.0574	0.0312
3	0.0720	0.0302	0.0093	0.0027
4	0.0739	0.0166	0.0031	0.0004

SPEX assumes that D is a closed, newly created collection of documents, a common setting. For instance, consider a lecturer that aims at detecting potential cases of unauthorised collaborative writing within a class. This is clearly an intra-corporal model (Culwin, 2008).

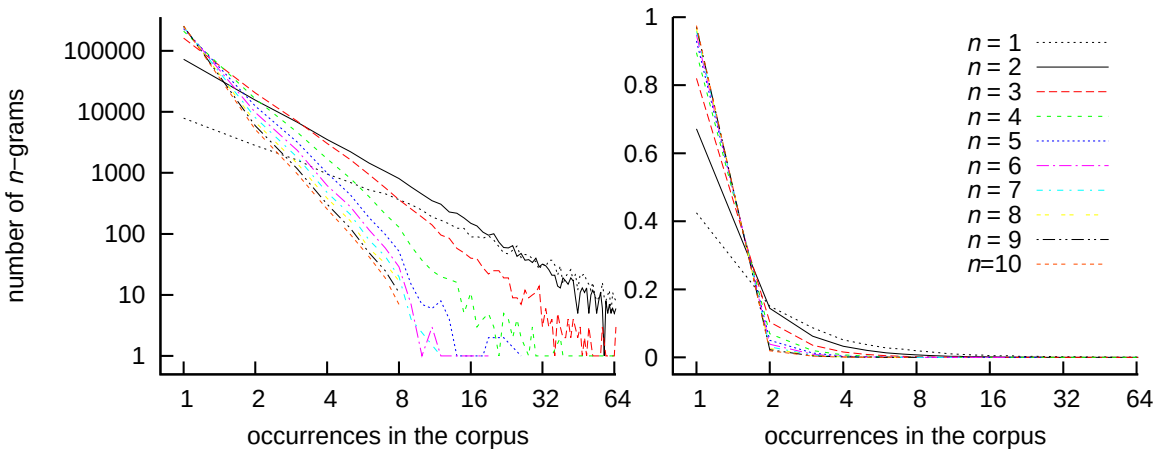
FERRET aims at detecting cases of plagiarism, even after modification (Lyon *et al.*, 2001, 2004). This approach considers word 3-grams as the shingles to compare. The use of this terms is empirically justified by the fact that word 3-grams are unlikely to co-occur in independently produced text.[11] As word 3-grams are likely to be hapax legomena and dislegomena, their frequency in a document can be discarded. Inspired by Broder (1997), two similarity measures based on set operations are considered: *resemblance* and *containment*. The former one is indeed the Jaccard coefficient, and aims at comparing texts of similar length, whereas the latter one is an adaptation that aims at determining whether a short text, e.g. a sentence or paragraph, was borrowed from a longer text, e.g. an entire document (cf. Section 3.3.1.1).

One of the conclusions of Lyon *et al.* (2004) is that the word 3-grams characterisation is the best option when looking for potential cases of plagiarism. Nevertheless, different studies have shown that this is not always true. In particular, Barrón-Cedeño and Rosso (2009a) showed that using 2-grams and 3-grams is roughly equivalent, at least in terms of F-measure (we further discuss on this issue in Section 5.3). Other researchers, e.g. Kasprzak and Brandejs (2010) and Zou, Long, and Ling (2010), consider that higher levels, $n \geq 4$, offer better results.

The idea that short n-grams work better (for detecting modified copy) is based on a simple fact: even if two documents are written by the same author and on the same topic, the expected amount of common [2-3]-grams between them is low. See for instance the figures in Table 5.1. As expected, the higher the n, the less likely that more than two documents will contain a common n-gram. If this happens when considering documents written by the same authors, the proportion of collision in two documents presumably written by different authors is expected to be even lower.

In order to stress this point, Fig. 5.6 shows the amount of common n-grams appearing in the newspapers of the METER corpus' courts partition (cf. Section 4.2.1 in page 81). This experiment is carried out over a corpus that includes different documents on the same event, which are very likely to share contents and even include common text fragments. When considering $n = 1$, a total of 7,901 (42%) n-grams occur only once in the corpus, they are *hapax legomena*. The number of *hapax dislegomena* is 2,800 (15%). The curve goes down very slowly because many 1-grams occur very often. When considering $n = 2$, the percentage of *hapax legomena* is already 67% (14% are *hapax dislegomena*). With 3-grams the amount of *hapax legomena* is already 82%; 10% of the shingles are *hapax dislegomena*. The percentage of hapax legomena and dislegomena increases dras-

[11] Cf. Fig. 5.6 in page 123 for a graphical proof of this fact.



(a) Absolute (b) Normalised

Figure 5.6: Occurrence of word n-grams in the METER corpus. Each curve represents the absolute/normalised amount of n-grams for a given n (y axis) that appear k times in the entire documents collection (x axis). The 770 newspaper texts from the courts section were considered.

tically and is directly proportional to the value of n. For instance, when considering $n = 8$, 97% of the shingles occur only once. These values are already close to those used in the strategies for detecting verbatim copy. However, it is worth noting that for very low values of n, $\{2, 3, 4\}$, a very low percentage of shingles occurs more than once.

Some researchers propose obtaining the shingles from d and d_q in a different way. For instance, Pataki (2006) proposes extracting overlapping shingles from the suspicious documents only. The justification is that non-overlapping shingles from the reference corpus save storing resources and, if the text re-use has implied some modification, this would be caught by d_q's n-grams.

As aforementioned, Heintze (1996) considered that no right length for the shingles exists. This fact is reflected by the variety in shingles used. For instance, character n-grams have been used considering $n = 16$ (Grozea, Gehl, and Popescu, 2009) and $n = 30$ (Micol, Llopis, and Muñoz, 2010). Word n-grams have been used considering with $n = 3$ (Muhr, Kern, Zechner, and Granitzer, 2010; Rodríguez Torrejón and Martín Ramos, 2010) and $n = 7$ (Gupta, Sameer, and Majumdar, 2010).[12] Using a "non-fixed n" seems to be a reasonable option.

Greedy string tiling is, according to Clough (2003), only one of the sequence comparison methods applied when looking for plagiarism. The advantage of this approach (Wise, 1993) is that it finds every common sequence of words between two texts by defining the minimum value of n only (e.g. if $n = 3$, every $[3, 4, \ldots]$-gram in common between the texts is detected). The core of the algorithm is the computation of longest common sub-strings (Wise, 1993).[13]

[12] Some of our experiments considering the word n-grams characterisation are discussed in Sections 5.2 and 5.3.

[13] As seen in Clough (2003).

Thesauri-based expansion is another strategy that can be applied when trying to uncover cases of text re-use, in particular those generated by paraphrasing (cf. Chapter 8). The aforementioned approaches to plagiarism detection simply look for string collisions, i.e., whether a shingle appears in two texts. However, as seen in Section 2.1.1, text re-use often implies paraphrasing, causing the vocabulary in the source to be modified in the re-used text.

Trying to overcome this difficulty, *PPChecker* considers the vocabulary in the two texts and every semantically related word (Kang, Gelbukh, and Han, 2006). The vocabulary in d_q is expanded on the basis of semantic relationships within the Wordnet synsets.[14] The similarity assessment between d_q and d is then performed at sentence level, representing each document on the basis of the BoW model. Rather than using a standard similarity measure, they propose a set of six different computations that end up in an overall similarity estimation. The measures consider different factors such as the union and difference between the vocabularies as well as their membership to a common Wordnet synset. By such a combination they manage to differentiate among verbatim copy and words insertions, removal, or substitution.

A similar expansion is carried out by Runeson, Alexandersson, and Nyholm (2007), which apply synonym substitution in order to detect duplicated reports of defects in a line of products. The thesaurus they exploit is not so general, though, as it comes from a defect management system. Hartrumpf, vor der Brück, and Eichhorn (2010) in turn, propose extracting hypernyms from Wikipedia, obtaining good results over a corpus composed of news, Web pages, and real plagiarism cases.

This kind of expansion comes at two costs: (*i*) the comparison process is much more expensive, having to compute up to six values for each pair of sentences and including more vocabulary in the comparison, and (*ii*) considering the semantically related words of the document's vocabulary inserts noise.

Dot-plot is yet another option to detect both verbatim and paraphrase plagiarism. Clough (2003) describes this technique as "a visualisation of matches between two sequences where diagonal lines indicate ordered matching sequences, and squares indicate unordered matches" and considers that it can be used for: (*i*) identifying regions of duplication within a text, and (*ii*) identifying regions of similarity across texts. This approach was originally aimed at aligning bilingual corpora (Church and Helfman, 1993). The documents are represented in an X, Y plane: d is located in X, while d_q is located in Y. The coordinates can be filled either with character n-grams, tokens, or word n-grams. In particular, Church (1993) considers character 4-grams (note that this model was inspired by that of Simard *et al.* (1992)).

A dot is drawn in the coordinate x, y if $d^x = d_q^y$; i.e., the term in d matches that of d_q (Manning and Schütze, 2002, p. 475). See a very simplified example in Table 5.7, considering a word $[1, 2]$-grams characterisation. If a text fragment from d has been copied (even after modification) into d_q, a shaded area will appear: (*a*) a straight line could indicate an exact copy or (*b*) a shaded square could indicate modified copy. Once the dot-plot is generated, the plagiarism candidate fragments have to be extracted. For in-

[14] `http://wordnet.princeton.edu`

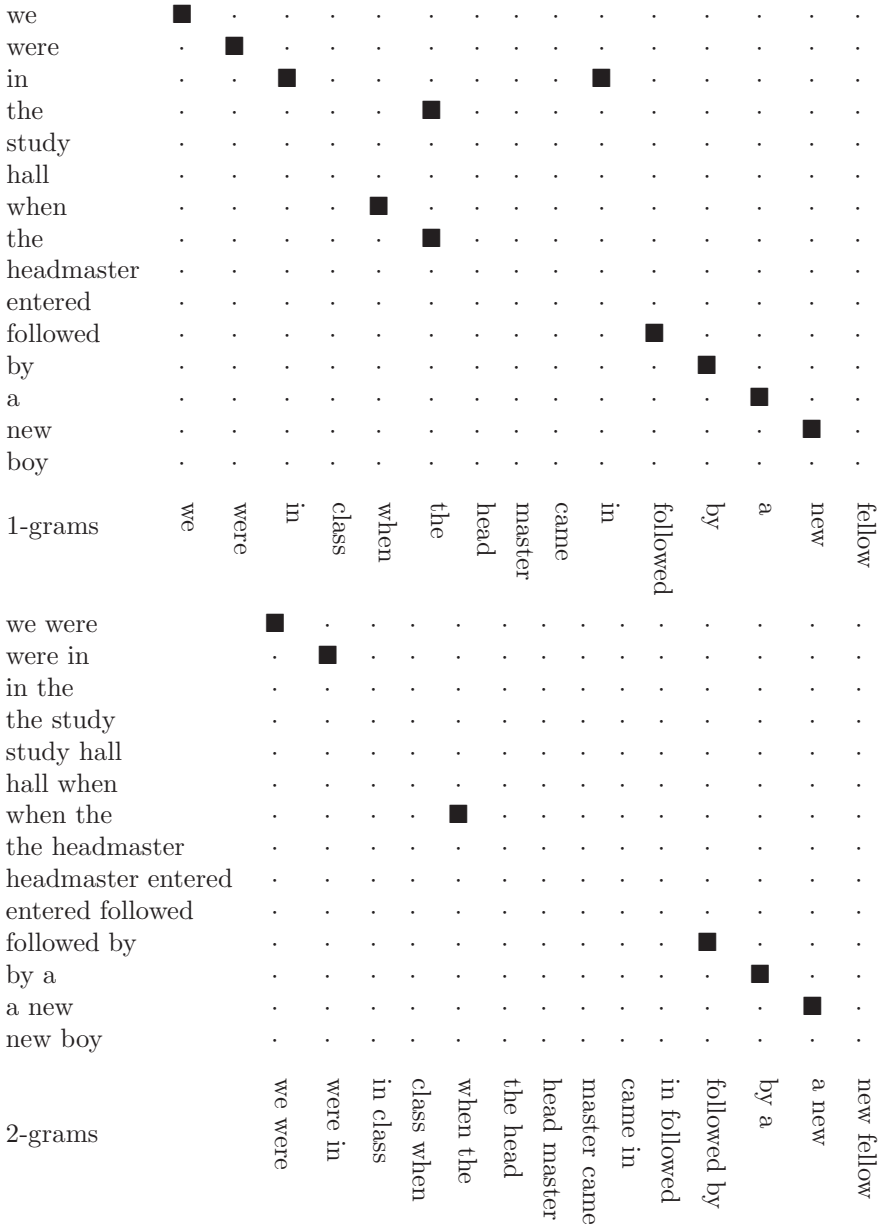


Figure 5.7: Example of dotplot between two sentences for 1-grams and 2-grams. A dot, in this simplified example represented as a square, is located if a term in the two sentences matches. Pre-processing consisted of tokenisation, casefolding, and punctuation marks removal. Sentences borrowed from the paraphrases corpus, created by Barzilay and McKeown (2001).

stance, Church (1993, p. 5) proposes using a low-pass filtering and thresholding approach to better identify the related fragments.

Piao (2001) had already used the dot-plot technique to detect re-use in journalistic material, in the METER corpus. Lancaster and Culwin (2004) in turn used the dot-plot as well, with word $[1, 2, 3]$-grams. They consider the dot-plot as a proper visual output for the user. Other researchers, such as Basile, Benedetto, Caglioti, and Degli Esposti (2009) and Grozea *et al.* (2009) applied dot-plot based techniques in the PAN competition (cf. Section 7.1.1).

Citation-based plagiarism detection Citations and references can be used to determine document similarities and identify potential cases of plagiarism (Gipp *et al.*, 2011). Finding similar patterns in the citations within two (scientific) texts is an indicator of semantic text similarity. Additionally, if two documents cite the same sources through the text, in the same order, suspicion triggers as well.

All of these models have something in common: they consider that the task of detecting duplication and re-use simply implies an exhaustive comparison of (a sample of) the contents in the implied documents. However, a few approaches consider that, before actually looking for a potential case of text re-use, a set of good source candidates for the contents of a suspicious text has to be identified.

5.1.2.2 Heuristic Retrieval

In many cases it is worth considering a preliminary retrieval stage before performing the detailed analysis of the plagiarism detection process. The most related documents $d \in D$ to d_q are retrieved, composing a collection of candidate source documents. We identify two scenarios where this makes sense: (i) when D is not a previously built, closed, collection of documents, for instance, D is the Web, and (ii) the aim is detecting paraphrase plagiarism, where the risk of obtaining a high rate of false positives is high.

Despite its importance, much of the research on automatic plagiarism detection considers either that this stage is solved or unnecessary. The reason is that researchers often assume D to be a local collection of documents (Bernstein and Zobel, 2004; Brin *et al.*, 1995; HaCohen-Kerner, Tayeb, and Ben-Dror, 2010; Iyer and Singh, 2005; Kang *et al.*, 2006), and just a few consider D to represent the entire Web (Bendersky and Croft, 2009). To make the situation worse, no corpus has been created in recent years that impulse the development of better models for this specific stage. On the contrary, the nature of the corpora recently developed has pushed against (cf. Section 4.2.3.6). As a result, not too much research has been focussed to determine the best way for performing this step or even determining how relevant it is for the quality of the output. See Section 5.4 for our contributions on this issue.

We identify three approaches to this stage. (a) the structure of the documents is exploited in order to perform a section-wise comparison; (b) only the contents are considered, without explicitly considering any structural information; and (c) the relationship among the documents considered in the reference collection is exploited.

Structure-based retrieval exploits knowledge about the different sectioning of a document. Si *et al.* (1997) proposed CHECK, which stratifies the document's contents into a tree structure, where the root is the document itself, the first level nodes are the sections, followed by subsections, etc. The bottom elements, the leafs, represent the paragraphs. Every node of the so built three contains a representation of its contents. Such a representation is no other than the set of keywords it contains. This set is composed of the nouns, verbs, and adjectives in the document fragment.

Additionally, the synsets of the selected keywords are looked up in Wordnet in order to increase the chance of detecting re-use after modification. With a top-bottom strategy,

the different nodes in d_q and d are compared on the basis of the dot-product (Eq. (3.15) in page 66). If the estimated similarity surpasses a given threshold, the children of the corresponding nodes are compared. It is only when the leaves are reached that a detailed analysis is carried out, by performing a sentence level comparison. Iyer and Singh (2005) base their approach on practically the same architecture of CHECK.

The structure of HTML files is exploited by Zhang and Chow (2011). They compose pseudo-paragraphs taking advantage of the $< p >$ and other HTML tags. If a text s_i between two of these structural tags is shorter than thirty words, it is merged to s_{i+1}, up to reaching a maximum length of fifty words. Every paragraph is represented into a real valued vector space model (using a variation of tf-idf) for comparison. If two fragments are considered similar enough, a sentence-wise comparison is carried out, this time over a Boolean space.

Content-based retrieval is another paradigm where the selection of relevant documents is faced as a query by example IR problem: retrieving topically similar documents. This is the most common approach to this stage. In most cases an IR search engine — such as Lucene or Terrier[15]— is used to index D and the contents of d_q are queried to retrieve similar documents (Muhr *et al.*, 2010; R. Costa-jussà, Banchs, Grivolla, and Codina, 2010).

Another option is applying related technology, such as a simple inverted index, to retrieve the set of most similar documents to d_q over the vector space model (Palkovskii, Belov, and Muzika, 2010; Rodríguez Torrejón and Martín Ramos, 2010). Different text representations have been used at this stage, including word 1-grams (Barrón-Cedeño *et al.*, 2009b; Zechner *et al.*, 2009), word 8-grams (Basile *et al.*, 2009), and character 16-grams (Grozea *et al.*, 2009).

Clustering-based retrieval is a paradigm nearly explored that seems to make sense when D is a closed set of documents. Once the collection of documents aimed at composing D has been obtained, Fullam and Park (2002) propose applying a clustering process. Only the documents from the most similar cluster to d_q are considered for the detailed analysis stage. Zechner *et al.* (2009) propose a similar approach. The sentences $s \in D$ are clustered and the centroids are identified. A sentence $s_q \in d_q$ is compared to the centroids only, and those documents $d \in D$ which sentences belong to the two most similar clusters are retrieved.

The Web as reference corpus The Web is few times considered on scientific publications about plagiarism detection. An on-line version of FERRET, *WebFerret*, aims at detecting article spinning (Malcolm and Lane, 2008).[16] In this case, word 3-grams are extracted from a Web page (those containing stopwords are discarded), and looked up on the Web.

Zaka (2009) is more ambitious and composes the queries of those sets of words within 200 character long chunks (the reasons behind this number are the restrictions of commercial Web search engines APIs, which sometimes admit queries 255 characters long at

[15]cf. `http://lucene.apache.org/` and `http://terrier.org/`.

[16]Article spinning is a form of plagiarism where *new* Web contents are created from existing material, "systematically rewording it to appear original" (Malcolm and Lane, 2008).

most). The top retrieved documents are downloaded for further compare them to d_q, on the basis of the vector space model.

Liu, Zhang, Chen, and Teng (2007) propose another model that looks for the potential sources of d_q on the Web. In this case, d_q is split into sentences which are sorted on the basis of their contents' term frequency. From top to bottom the sentences are queried to the search engine, downloading the most relevant documents for the detailed analysis stage.

5.1.2.3 Knowledge-Based Post-Processing

Once the plagiarised-source candidates are identified, a post-processing is applied. When this stage was proposed by Stein *et al.* (2007), it aimed at exploiting linguistic (and other kinds of) knowledge to discarding those detected cases which were not actual plagiarism. For instance, consider a case where the proper citation is provided. Nevertheless, we are not aware of any approach which has actually tried to face this problem.[17] Beside the inherent complexity of the problem, no corpora including borrowings with real cases of plagiarism and acknowledged text re-use is at hand. As discussed in Section 4.2.3, in the PAN-PC corpora no re-used fragment includes quotations or references to its source. As a result, this stage is becoming a *heuristic post-processing*, rather than knowledge-based. In heuristic post-processing, the aim is cleaning the output to better present it to the user. Heuristics are used both to merge and discard detected cases.

Two identified cases of plagiarism, namely $(s_{plg,1}, s_{src,1})$ and $(s_{plg,2}, s_{src,2})$ are merged if (a) $(s_{plg,1}, s_{plg,2}) \in d_q$, $(s_{src,1}, s_{src,2}) \in d$ and (b) $dist(s_{plg,1}, s_{plg,2})$ is minor than a given threshold. That is, two cases of plagiarism from the same document are detected having their source in the same document and they appear very close to each other. Moreover, an identified case of plagiarism is discarded if it is considered too short to be considered an actual fault.

The final process in automatic plagiarism detection is the generation of a useful output for the user. Clough (2003) discriminates two kinds of (non-exclusive) outputs:

Quantitative. For instance, the longest common sub-string or the average common sub-string length as well as a similarity measure.

Qualitative. A visual representation of matches, for instance, a dot-plot.

5.1.2.4 Pre-processing

We have left pre-processing, the first stage of the external plagiarism detection process, as the last point to discuss about. The reason is that whether this stage makes sense in this problem remains unclear. In many cases, researchers report pre-processing the documents before analysing them, but other do not. Common operations include tokenisation, case folding and stemming. Moreover, punctuation marks and stopwords

[17] We stress again that the final decision has to be made by an expert.

are sometimes discarded as they are considered not to contain information about the document.

Regarding the last point, stopword deletion, Stamatatos (2011) proposes exactly the opposite: discarding every content word and considering stopwords only, in particular a set of fifty tokens. In order to perform the heuristic retrieval process he uses stopword 11-grams, and 8-grams for the detailed analysis. Stamatatos (2011) justifies the good results by the fact that stopwords are "associated with syntactic patterns" and the "syntactic structure of a document [...] is likely to remain stable during a plagiarism stage".

5.1.2.5 Detecting the Direction of Re-Use

Ryu, Kim, Ji, Gyun, and Hwan-Gue (2008) claim to be the first to detect the directionality of the plagiarism, i.e., which document is the source and which one is the re-used. However, they do so on the basis of the document's own time-stamp, something that can be easily falsified.

Grozea and Popescu (2010b) propose using a dot-plot based approach to determine, between d_q and d, what document borrowed from the other. They do so by turning the problem of plagiarism detection upside down; i.e., instead of looking for the potential source of d_q, they look for the potential re-uses from d (Grozea *et al.*, 2009). As they mention, it is more likely that a character n-gram will appear more often in the rest of the source document it was borrowed from than in the document it is inserted in. On the basis of character 8-grams, they look precisely for this behaviour.

Surveys of the research done on automatic plagiarism detection can be reviewed in Clough (2003) and Maurer *et al.* (2006). In the section to come, we describe our experiments on monolingual text re-use detection.

5.2 Word n-Grams Retrieval

In this experiment we aim at determining how different pre-processing, representation, and weighting strategies affect the process of re-use detection when dealing with real cases of journalistic text re-use. The experiment resembles the situation where d_q is given and the documents in D are ranked according to the similarity $sim(d_q, d)$ for every $d \in D$, i.e., a query by example retrieval. In the ideal case, the document d that covers the same event of document d_q would be located on top of the ranked list (either it was used as source of d_q's contents or not). In this case we are not interested in determining whether d_q and d actually have a re-use relationship.[18]

[18]This experiment is inspired by Experiment 1 in Potthast *et al.* (2011a).

5.2.1 Experimental Setup

In this case, we used the METER corpus (Clough *et al.*, 2002) (cf. Section 4.2.1). The set of query documents D_q is composed of the entire set of newspaper notes (around 950 documents). The reference corpus D is the entire set of PA notes (around 770 documents). One particularity has to be considered in this corpus definition. In some cases, for a document d_q in the collection, more than one related document d exists in D. The reason is that sometimes the PA publishes more than one note on the same event, most of them complementary. We assume that the most relevant document for d_q is the most similar among those in the relevant subset.

The text representation models selected are word n-grams, with n in the range $[1, \ldots, 10]$. $n = 1$ represents a "simple" BoW representation. For higher values we are indeed trying to retrieve documents sharing exact text fragments. As a result, we aspire to properly retrieve the relevant document d from d_q only in those cases where it contains borrowed fragments. A total of four weighting schemas were considered, including Boolean and three real valued (cf. Sections 3.2.1 and 3.2.2). The real valued weighting models are tf-idf, tf, and tp. For the Boolean weighting, $sim(d_q, d)$ was computed with the Jaccard coefficient (cf. Section 3.3.1.1). For the real valued weightings, $sim(d_q, d)$ was computed with the cosine similarity measure (cf. Section 3.3.1.2). The pre-processing options we explored are case folding, stemming, punctuation removal and stopwords removal. The only operation carried out in every case was tokenisation.

5.2.2 Results and Discussion

We do not include a graphical representation of the entire set of experiments. Instead we show and discuss the most interesting cases. Figure 5.8 shows the recall at k curves for the four weighting schemas applied to the different values of n when no pre-processing —other than tokenisation— is applied at all. When considering $n = 1$ it is remarkable that by disregarding any term weighting strategy we obtain the best results ($rec@1 = 0.64$ for Boolean respect to the second best, 0.43 for tf-idf). The Boolean weighting is surpassed up to $rec@5$ only (0.915 for Boolean respect to 0.923 for tf-idf). Practically all the relevant documents are ranked among the top-10 documents when terms are weighted with tf-idf ($rec@10 = 0.999$)

When considering $n = \{2, \ldots, 10\}$, the Boolean weighting, together with the Jaccard similarity, allows for getting the best results for top values of k. Two interesting phenomena occur in these cases. Firstly, considering tf-idf, tf, or tp results in similar retrieval qualities for all the values of n. Secondly, whereas n increases, the four weighting schemas tend to converge more and more, but every time Boolean slightly overcomes the rest.

Figure 5.9 shows the same figures, but this time applying all the pre-processing operations. This time since the lowest values of n, all the real-valued weighting schemas, in combination with the cosine measure, result in practically the same retrieval quality (average $rec@1 = 0.45$). This is more clear for $n = 4$: we already obtain low levels of recall. As longer terms tend to be hapax legomena and dislegomena, no matter the

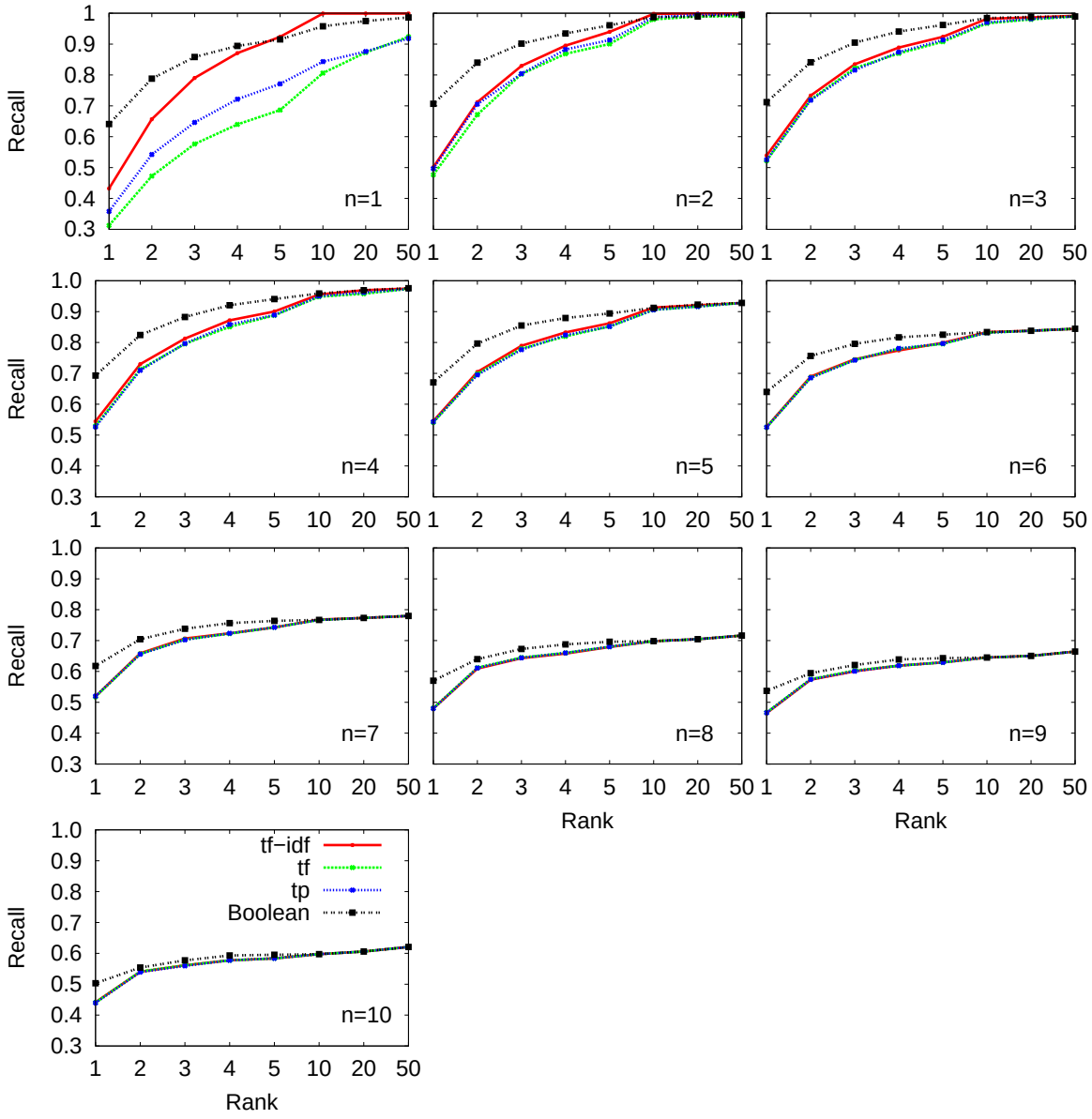


Figure 5.8: Retrieval experiments over the METER corpus without pre-processing. Experiments carried out with word $[1, \ldots, 10]$-grams. The only text pre-processing is tokenisation. The similarities for tf-idf, tf, and tp weighting schemas were computed with cosine. The similarity for Boolean weighting was computed with the Jaccard coefficient. Values of rec@k are shown for $k = \{1, 2, 3, 4, 5, 10, 20, 50\}$.

weighting used, the similarity trends are the same for the three cases. Once again the Boolean weighting clearly outperforms the rest ($rec@1 = 0.68$) and converges only at $k = 10$.

When considering 1-grams and 2-grams, tf and tp are practically as good as tf-idf. In both cases the outcome is slightly better than when applying no pre-processing at all, particularly for $n = 2$. However, the quality decreases drastically as higher n-grams are considered ($n \geq 4$). The reason is very simple: for higher values of n, the elimination of stopwords (even punctuation marks) affects many terms and, as already discussed, the

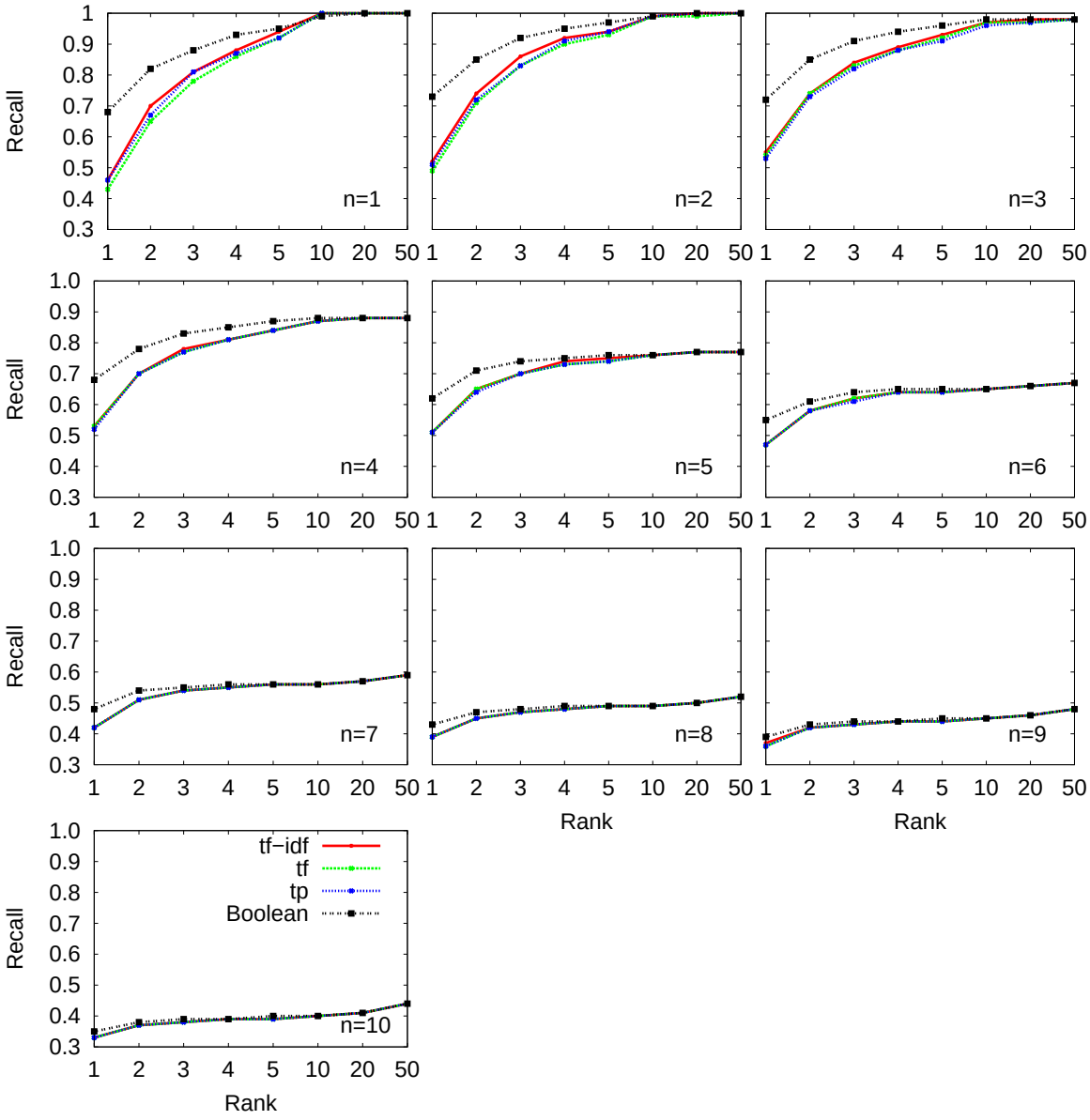


Figure 5.9: Retrieval experiments over the METER corpus with pre-processing, including tokenisation, case folding, and stemming as well as punctuation and stopwords removal. Experiments carried out with word $[1, \ldots, 10]$-grams. The similarities for tf-idf, tf, and tp weighting schemas were computed with cosine. The similarity for Boolean weighting was computed with the Jaccard coefficient. Values of rec@k are shown for $k = \{1, 2, 3, 4, 5, 10, 20, 50\}$.

longer a term, the more sensitive to modifications it is, even those that are product of pre-processing. The same happens with the Boolean weighting. The output is better for $n = [1, 2, 3]$, but since $n = 4$ it becomes worst.

Finally, Fig. 5.10 shows the output when applying tokenisation, case folding and stemming (no tokens are eliminated). The results are very similar to those obtained without pre-processing (for instance, when considering $n = 1$, $rec@1 =$ for tf and tf-idf show a very slight increase of 0.02 in both cases). It seems that neither case folding nor

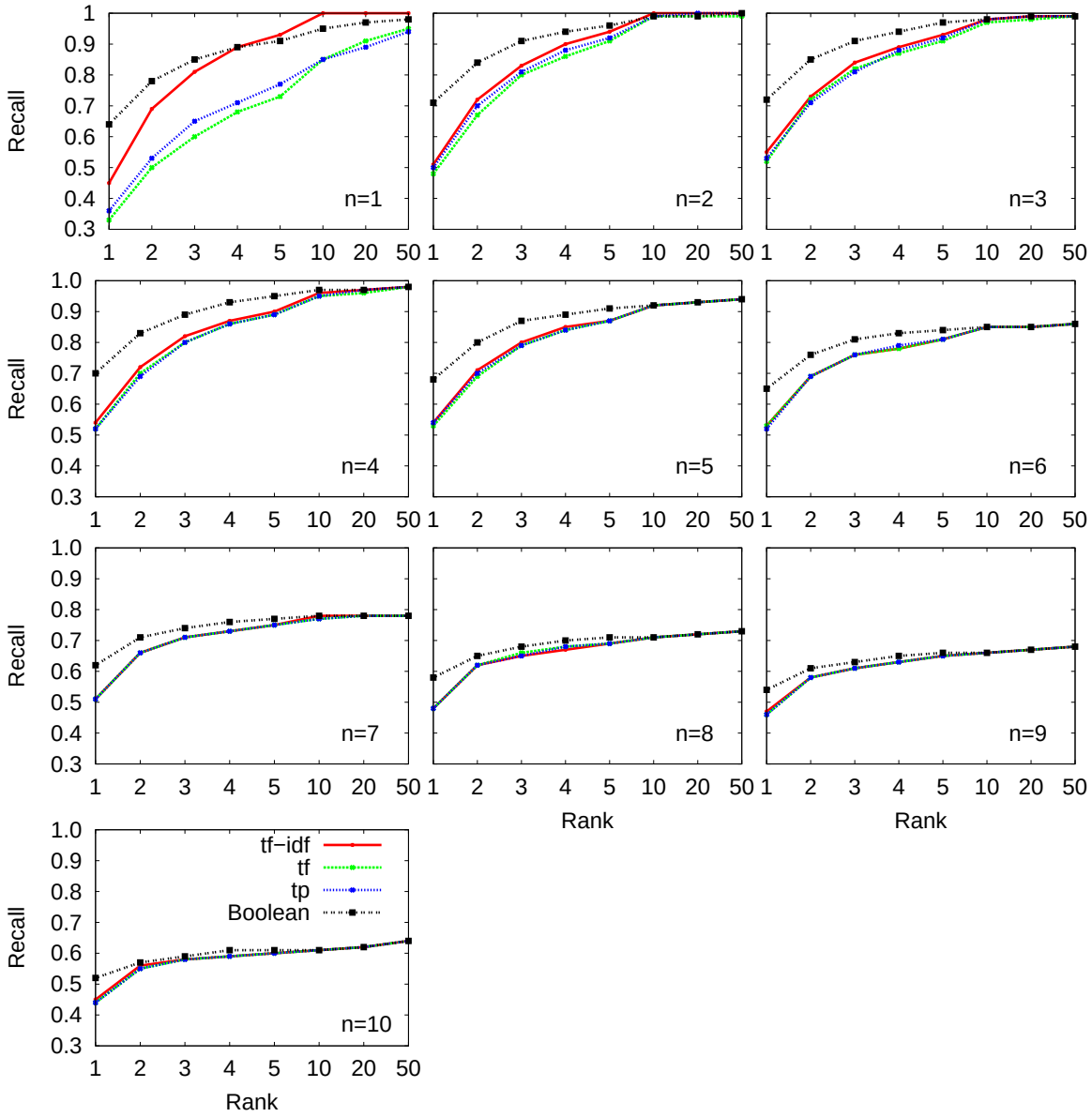


Figure 5.10: Retrieval experiments over the METER corpus with some pre-processing, including tokenisation, case folding, and stemming. Experiments carried out with word $[1, \ldots, 10]$-grams. The similarities for tf-idf, tf, and tp weighting schemas were computed with cosine. The similarity for Boolean weighting was computed with the Jaccard coefficient. Values of rec@k are shown for $k = \{1, 2, 3, 4, 5, 10, 20, 50\}$.

stemming make a big difference, but still improve results slightly. Regarding stopwords, at least since $n > 2$, discarding them results reasonable.

5.3 Containment-based Re-Use Detection

This experiment resembles the situation when we aim at determining whether the contents in a document have been generated by re-use from an external source. Given a

Given d_q and D:

// **Pre-processing**
$casefold(d_q)$; $stem(d_q)$
$casefold(d)$; $stem(d)$

// **Analysis**
for each sentence s_q in d_q:
    n_{s_q} = [n-grams in s_q]
    for each d in D:
        n_d = [n-grams in d]
        compute $sim(n_{s_q}, n_d)$
    if $\max_{d \in D}(sim(n_{s_q}, n_d)) \geq threshold$:
        s_q becomes re-use candidate from $\arg\max_{d \in D}(sim(n_{s_q} \mid n_d))$

Figure 5.11: Sentence re-use detection algorithm. $casefold()$ and $stem()$ perform case folding and stemming. $sim(n_{s_q}, n_d)$ computes the containment between the sets of n-grams in s_q and d. Note that this algorithm is designed with explanatory rather than implementation purposes.

suspicious document d_q and a reference corpus D, the objective is answering the question "*Is a sentence $s \in d_q$ re-used from document $d \in D$?*". In summary, we aim at detecting whether a sentence has been re-used, together with its source document (Barrón-Cedeño and Rosso, 2009a).

5.3.1 Experimental Setup

In this case the corpus D is again the entire set of PA notes in the METER corpus. The corpus D_q, i.e., the set of re-use suspicion documents, is composed of approximately 440 newspaper notes only. The reason is that these documents include fragment-level annotation, identifying the different fragments as *verbatim* or *rewritten* copy from the PA note or *newly* created. For experimental purposes we consider a sentence $s \in d_q$ to be re-used from a PA note if a high percentage of its words belong to verbatim or rewritten fragments (i.e., paraphrased). We assume that s is re-used from d if:

$$|s_V \cup s_R| > 0.4 \cdot |s| \ ,$$

where s_V and s_R are the sets of words in verbatim and rewritten fragments within s, respectively. This assumption avoids considering sentences with incidental common fragments (such as named entities) as re-used. The distribution of verbatim, rewritten and new fragments among the entire set of suspicious sentences is $\{43, 17, 39\}\%$, respectively. For the set of sentences considered as re-used the distribution becomes $\{65, 26, 7\}\%$.

The retrieval process, which considers Boolean weights, is depicted in Fig. 5.11. The pre-processing applied to D and D_q includes tokenisation, case folding, and stemming (no stopword deletion is applied and punctuation marks are considered just another word in the text). Document d_q is then split into sentences s_q and every s_q is represented by the set of n-grams it contains. No splitting is applied to d and is represented by its n-grams. As the cardinality of the compared sets will be, in general, very different ($\vec{s_q} \ll \vec{d}$), an asymmetric comparison is carried out on the basis of the *containment* measure (Broder, 1997) (cf. Eq. 3.13 in page 65). A sentence s_q is considered a re-use candidate from d if $\max_{d \in D} sim(s_q, d)$ is higher than a given threshold.

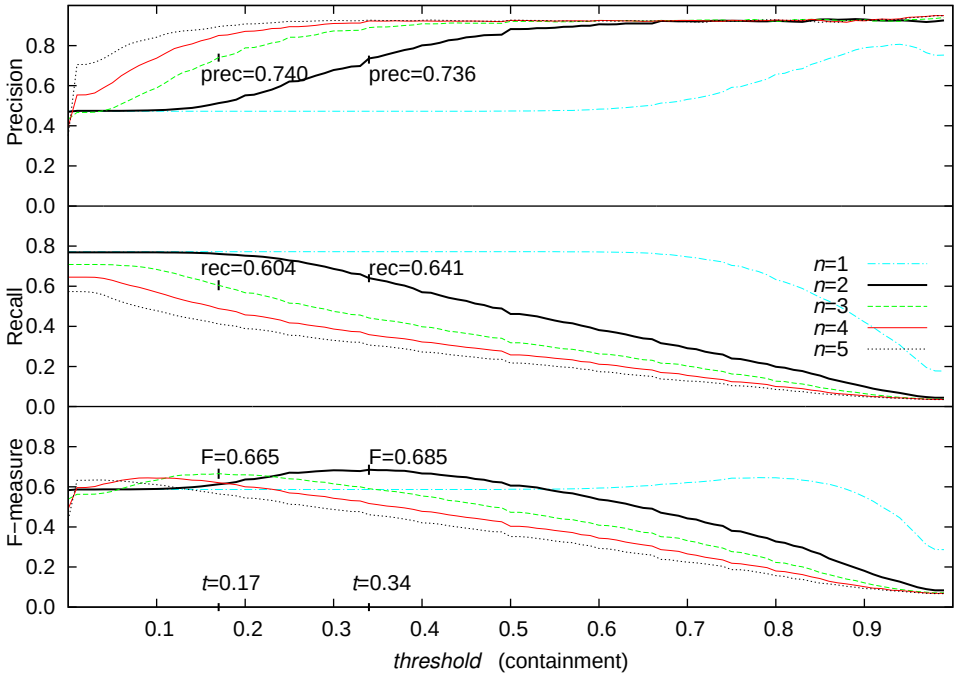


Figure 5.12: Containment experiments over the over the METER corpus. *prec*, *rec*, and F curves displayed for different levels of n-grams ($n = [1, \ldots, 5]$) and threshold values.

The experiment is evaluated on the basis of precision, recall and F-measure. We carried out a 5-fold cross validation process. The aim was tuning the containment threshold that decides whether a suspicious sentence is re-used or not. The threshold selected is that for which the best levels of *prec*, *rec* and F are obtained when considering 4 sets of suspicious documents. The threshold with the best F-measure t^* was considered when evaluating the retrieval quality with the fifth, previously unseen, set.

5.3.2 Results and Discussion

Figure 5.12 shows the obtained results by considering from 1-grams up to 5-grams (once again, we experimented up to $n = 10$, but the obtained results are worst). When considering $n = 1$ (i.e., BoW) a good recall is obtained (practically constant until $threshold = 0.7$). However, it is very likely that d would contain the entire vocabulary of a sentence s_q. As a result, the values of precision are among the lowest obtained. On the other side, considering 4-grams and 5-grams (and $n > 5$) produces a very rigid search strategy. Minor changes in a re-used sentence prevents its detection. As a result, the values of recall in these cases are among the lowest and, as expected, with a high precision.

The best results are obtained by considering $n = \{2, 3\}$ (best F-measures: 0.68 and 0.66, respectively). In both cases, the resulting terms are short enough to handle modifications (i.e., paraphrasing)) in the re-used sentences. Additionally, they are still long enough to compose shingles which are unlikely to appear in any (but the actual source) text: 3-gram based search is more rigid, resulting in a better precision; 2-gram

based search is more flexible, allowing for a better recall. The difference is reflected in the threshold where the best F-measure is obtained in both cases: 0.34 for 2-grams versus 0.17 for 3-grams. Selecting 2-grams or 3-grams depends on the interest of catching as most as possible re-used fragments or leaving out some of them with the aim of after reviewing less candidates.

5.4 The Impact of Heuristic Retrieval

As aforementioned, much of the research on automatic plagiarism detection assumes that D is a small collection of documents and that any detailed analysis strategy, regardless how costly it is, is enough. Two facts are against this assumption: (i) a detailed analysis of the contents of d_q and every $d \in D$ is very costly, and (ii) such a comparison might be carried out considering too many irrelevant documents respect to d_q. The first issue can be handled with efficient comparison strategies. However, the second issue increases the risk of incorrectly considering that two, unrelated documents, have a relationship of re-use. We consider that a preliminary filtering of good potential source documents for d_q, the known as heuristic retrieval stage (cf. Section 5.1.2.2), is necessary.

In this experiment our efforts are oriented to localise the subset of documents D', that are more related to d_q ($|D'| \ll |D|$). We expect that D' will contain the most likely source documents for the potentially re-used text fragments in d_q. After obtaining D', a detailed analysis between the contents of d_q and $d \in D'$ can be performed. Our retrieval method, is based on the Kullback-Leibler symmetric distance (Barrón-Cedeño *et al.*, 2009b).

5.4.1 Proposed Heuristic Retrieval Model

The proposed model is based on Bigi's (2003) version of the Kullback-Leibler distance (Kullback and Leibler, 1951) (cf. Section 3.3.2.1). Given a reference corpus D and a suspicious document q we calculate the KL_δ of the probability distribution P_d with respect to Q_s (one distance for each document $d \in D$), in order to define a reduced set of reference documents D'. These probability distributions are composed of a set of features characterising d and d_q. The detailed analysis follows considering d_q and D' only.

5.4.1.1 Features Selection

The features selection necessary to compose a probability distribution P_d was carried out as follows. The terms' relevance was assessed on the basis of three alternative techniques already used before: (i) tf, (ii) tf-idf, and (iii) tp. As discussed in Section 3.2.2, these are weighting models that aim at determining how relevant a term for a given document is. They do so on the basis of the documents contents only (tf and tp), or the term's frequency inside of the document and the entire considered corpus (tf-idf). The aim of the feature selection process is to create a list of terms ranked by their relevance. Each probability distribution P_d is composed of the top terms in such a rank, as they are

supposed to better characterise the document d.

5.4.1.2 Term Weighting

Term weighting is considered as the terms' probability, simply estimated by a maximum likelihood estimation (MLE): it is calculated by Eq. (3.2) (cf. page 61), i.e., $P(t_i, d) = tf_{i,d}$. These probability distributions are independent of any other document and require to be computed only once.

Given a suspicious document d_q, a preliminary probability distribution Q'_{d_q} is obtained by the same weighting schema, i.e., $Q'(t_i, d_q) = tf_{i,d_q}$. Nevertheless, when comparing d_q to each $d \in D$, in order to determine whether d is a source candidate of the potentially plagiarised sections in d_q, Q'_{d_q} must be adapted. The reason is that the vocabulary in d_q and d, and therefore the corresponding probability distributions, will be different in most cases. KL_δ becomes infinite if a t_i exists such that $t_i \in d$ and $t_i \notin d_q$. As a result, the probability distribution Q_{d_q} does depend on each P_d it is compared against.

If $t_i \in P_d \cap Q'_{d_q}$, $Q(t_i, d_q)$ is smoothed from $Q'(t_i, d_q)$; if $t_i \in P_d \setminus Q'_{d_q}$, $Q(t_i, d_q) = \epsilon$. This is a simple back-off smoothing of Q. In agreement with Bigi (2003), the probability $Q(t_i, d_q)$ is estimated as:

$$Q(t_i, d_q) = \begin{cases} \gamma \cdot Q'(t_i \mid d_q) & \text{if } t_i \text{ occurs in } P_d \cap Q'_{d_q} \\ \epsilon & \text{if } t_i \text{ occurs in } P_d \setminus Q'_{d_q} \end{cases} . \tag{5.2}$$

Note that those terms occurring in d_q but not in d become irrelevant; γ is a normalisation coefficient estimated by:

$$\gamma = 1 - \sum_{t_i \in d, t_i \notin s} \epsilon \ , \tag{5.3}$$

respecting the condition:

$$\sum_{t_i \in s} \gamma \cdot Q'(t_i, s) + \sum_{t_i \in d, t_i \notin s} \epsilon = 1 \ . \tag{5.4}$$

ϵ is smaller than $\min_i P(t_i, d)$, i.e., the minimum probability of a term in document d.

After computing $KL_\delta(P_d \mid\mid Q_{d_q})$ for all $d \in D$, it is possible to define a subset of source documents D' of the potentially re-used fragments in d_q. We define D' as the ten reference documents d with the lowest KL_δ with respect to d_q. Once d_q and D' are at hand, a re-use detection between suspicious sentences and potential source documents, as described in Section 5.3, can be carried out.[19]

[19]Note that in our experiments over the PAN-PC corpora (cf. Section 7.4) we do not apply any feature selection during the heuristic retrieval step. The reason is that in those corpora the plagiarised fragments and their contexts (i.e., the documents they are inserted in), are not necessarily on the same topic, and sub-sampling could mislead the retrieval of good potential source documents.

Given d_q and D:

// **Distance computations**
Compute P_d for all $d \in D$
Compute Q'_{d_q}
For each d in D
    Compute Q_{d_q} given P_d
    Compute $KL_\delta(P_d \mid\mid Q_{d_q})$

// **Heuristic retrieval**
$D' = \{d\}$ such that $KL_\delta(P_d \mid\mid Q_{d_q}) \in 10$ lowest distances

// **Detailed analysis**
// As described in Fig. 5.11, considering d_q and D'

Figure 5.13: Heuristic retrieval process. Firstly, the distance between the documents' probability distributions are computed. Secondly, the heuristic retrieval is performed: the ten least distant documents $d \in D$ compose D'; d_q and D' are the input for the detailed analysis. Note that this algorithm is designed with explanatory rather than implementation purposes.

5.4.2 Experimental Setup

As pointed out, the aim of the proposed method is to select a good set of candidate source documents to become the input of the detailed analysis stage, based on sentence to document comparison. Once P_d is obtained for every document $d \in D$, the entire search process is as the one sketched in Fig. 5.13. The distances between the probability distributions of d_q and every $d \in D$ are computed. The ten documents in D with lowest distance respect to d_q are retrieved. Finally, a detailed analysis is carried out on the basis of the containment measure. The corpus used in this experiment is the one described in Section 5.3.1.

We carried out three experiments of the plagiarism detection process with and without heuristic retrieval. We aimed at comparing speed and quality of the results (in terms of precision, recall and F-measure). The experiments explore the following four parameters:

1. Length of the terms composing the probability distributions: $l = \{1, 2, 3, 4\}$
2. Feature selection technique: tf, tf-idf, and tp
3. Percentage of terms in d considered in order to define P_d: $[10, \ldots, 90]\%$
4. Length of the n-grams for the detailed analysis process: $n = \{1, 2, \ldots, 5\}$

In order to explore these parameters, we designed a total of three experiments. In experiments 1 and 2, a 5-fold cross validation was carried out.

Experiment 1: Terms length in heuristic retrieval. It aims at defining the best values for the first two parameters of the heuristic retrieval process. Given a suspicious document d_q, we consider that D' has been correctly retrieved if it includes the source document of d_q's contents. We explored all the combinations of l and feature selection techniques.

Experiment 2: Terms amount in heuristic retrieval. It aims at exploring the retrieval quality when varying the third parameter of the process: the percentage of considered terms when composing the probability distributions of every document. Once again, we explored all the combinations of l and the different feature selection techniques.

Table 5.2: Retrieval + detailed analysis versus detailed analysis in the METER corpus. Results displayed for the case when a heuristic retrieval stage is applied or not before performing the detailed analysis. (P =Precision, R =Recall, F = F-measure, t = relative time).

Heuristic retrieval	threshold	prec	rec	F	time
NO	0.34	0.73	0.63	0.68	t
YES	0.25	0.77	0.74	0.75	$0.08 \cdot t$

Experiment 3: Impact of heuristic retrieval. It aims at analysing the impact of the heuristic retrieval stage in the detailed analysis process. It compares the speed and output quality either applying heuristic retrieval or not.

5.4.3 Results and Discussion

Figure 5.14 presents the obtained results in Experiment 1 (they are displayed with standard deviation for the cross-validation). It contains the percentage of sets correctly retrieved in the experiments carried out on the different development sets. The results obtained with the different feature selection techniques are similar for every value of l. In the three cases the best results are obtained when considering 1-grams[20]

Higher n-gram levels produce very flat probability distributions, i.e., the most of the terms have nearly the same probability. These distributions do not allow KL_δ to determine how close two documents actually are.

Regarding the comparison of the different feature selection techniques, considering tf does not give good results. In this case a good number of functional words (prepositions and articles, for example), which are unable to characterise a document, are considered in the corresponding probability distributions. The results obtained by considering tp are close to those with tf. Considering "mid-terms" (which tries to discard functional words), seems not to characterise either this kind of documents because they are too noisy. The results with this technique could be better with longer documents, though. The best results are obtained with tf-idf. Functional and other kinds of words that do not characterise the document are eliminated from the considered terms and the probability distributions correctly characterise d (and after d_q).

The results of Experiment 2 are also reflected in Fig. 5.14, regarding at the amount of terms that compose the probability distributions. The best results of the previous experiment are obtained with 1-grams, hence we concentrate our analysis on these terms only. For the cases of tf and tp, the retrieval quality is very low when considering small amounts of terms (around 70%). Percentages higher than 90% are only reached when practically the entire vocabularies are considered to compose the distributions.

The behaviour is different when considering tf-idf, though. The quality of the re-

[20]Note that this result is not in contradiction with those obtained in Section 5.2.2 because in those cases the suspicious documents were fully represented (i.e., all the terms in d_q were considered, whereas here a sub-sampling is applied.

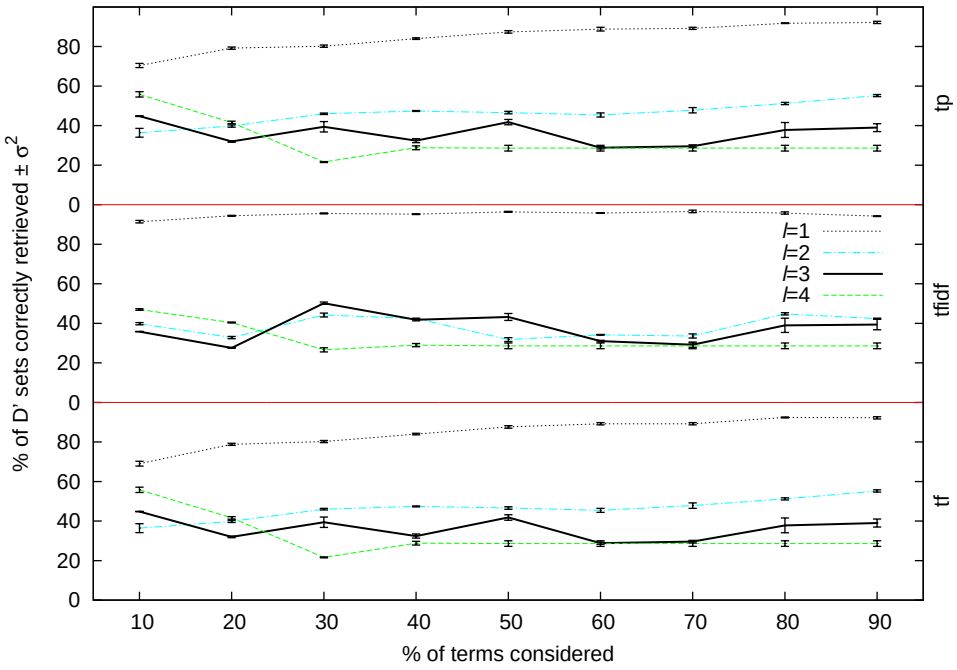


Figure 5.14: Evaluation of the heuristic retrieval process on the METER corpus. Percentage of sets correctly retrieved ($\{tf, tf\text{-}idf, tp\}$ = feature extraction techniques, l = term length).

trieval is practically constant since low amounts of terms are considered. The only real improvement is achieved with 20% of the document's vocabulary; the improvement from 10% to 20% of the considered vocabulary goes from 91% to 94%. As a result, composing the probability distribution with the 20% of the vocabulary is enough. In this way, we obtain a good percentage of correctly retrieved documents from D with a low dimension for the probability distributions. Once the best parameters were learnt during the training stage we applied them to the corresponding test partition. The obtained results did not variate significantly.

Finally, Experiment 3 aimed at comparing the output of the system in cases where the heuristic retrieval was either applied or not before the detailed analysis. Remember that, for the detailed analysis stage, we already explored considering terms of different lengths: both 2-grams and 3-grams offered the best balance between precision and recall (cf. Section 5.3). Hence, in this experiment our detailed analysis stage uses word 2-grams.

Table 5.2 shows the results of the detailed analysis when the contents of d_q are searched either over D or D'; i.e., the original and the previously filtered reference corpus. In the first case, we calculate the containment of $d_{q_i} \in d_q$ over the documents of the entire reference corpus D. Although this technique by itself obtains good results (cf. Fig. 5.12), considering too many reference documents that are unrelated to d_q produces noise in the output. This affects the quality in terms of precision and recall.

An important improvement is obtained when $s_q \in d_q$ is searched over D' only, after the heuristic retrieval process: precision and recall are clearly better. The reasons are twofold: (i) many documents in D that are not actually related to (the topic of) d_q are discarded in the first stage, and (ii) as a result of the filtering, lower similarity

thresholds can be considered to discriminate between an actual case of re-use and an incidental match.

With respect to the processing time, let us assume that the average time required to analyse a d_q in detail over the entire reference corpus D is t. The entire process of heuristic retrieval followed by a detailed analysis of d_q over the reduced corpus D' required only $0.08 \cdot t$, i.e., one tenth of the time. The reasons for such a big time difference are threefold: (*i*) P_d is pre-computed for every reference document only once, (*ii*) $Q'(d_q)$ is computed once and simply adapted to define each Q_{d_q} given P_d, and (*iii*) instead of searching the sentences of d_q in D, they are searched in D', which in this experiment contains only 10 documents.

5.5 Chapter Summary

In this chapter we defined the problem of automatic text re-use (and plagiarism) detection. Two approaches were described: intrinsic and external. In intrinsic detection unexpected variations within a document are looked for retrieving potential cases of re-use. In external detection the document's contents are compared to others, looking for unexpectedly high levels of similarity that might imply borrowing.

Afterwards, a review of the literature available on both approaches was presented. We discussed how the text similarity models and stylometric measures described in Chapter 3 can be exploited when trying to automatically uncover a potential case of text re-use. We noted that for intrinsic plagiarism detection a simple characterisation based on character n-grams seems to be the best available option. For external detection some strategies proposed during the last fifteen years were reviewed. The described strategies are different in nature, depending on whether they aim at detecting cases of exact or paraphrase copy.

We paid special attention to analyse how the terms representing a document, the units to compare, should be built. We considered the case of word n-grams and analysed the likelihood of an n-gram to occur in different documents, either when they are topically related or imply an actual case of borrowing. As expected, the longer a string, the more unlikely it will appear in two documents, unless one contains re-used fragments from the other.

In the second part of the chapter we presented a total of three experiments related to the external detection of text re-use. The first experiment had nothing to do with plagiarism detection, but was a query by example retrieval exercise. The aim was analysing how different representation, weighting, and similarity assessment models behave when aiming at retrieving related documents. Boolean and real weighting schemas (term frequency, term frequency-inverse document frequency, and transition point), the Jaccard coefficient and the cosine measure, and different n-grams levels were explored. The obtained results showed that, in general, using a simple Boolean weighting, together with the Jaccard coefficient is a better option. Moreover, it was seen that, beside tokenisation, no other pre-processing is strictly indispensable to get the best results using word 1-grams and 2-grams (as, for instance, stemming improves the results just slightly).

In the second experiment we aimed at detecting re-use in newspapers. The search strategy was based on the asymmetric search of suspicious sentences across a set of reference documents (both codified as word n-grams). Comparing sentences to entire documents became the search strategy even more flexible. The experimental results showed that 2-grams and 3-grams are the best comparison units for this task; 2-grams favour recall, whereas 3-grams favour precision.

The last experiment represented a more complete plagiarism detection process. We investigated the impact of applying a retrieval process as the first stage of external plagiarism detection. The retrieval method was based on the Kullback-Leibler symmetric distance, which measures how closed two probability distributions are. The probability distributions contained a set of terms from the reference and suspicious documents. We were interested in using the least possible amount of terms. In order to compose the distributions, term frequency, term frequency-inverse document frequency and transition point were considered. The best results were obtained when the probability distributions were composed of the 20% of the word 1-grams in a document, selected on the basis of term frequency-inverse document frequency.

A comparison of the obtained results was made by carrying out the detailed analysis over the entire and the reduced reference corpora. When the preliminary retrieval was carried out, the entire collection of potential source documents was reduced to only 10 candidate source documents. The quality of the obtained results was improved respect to the base model, showing the relevance of the preliminary retrieval in the detection process.

Related publications:

- Barrón-Cedeño and Rosso (2009a)
- Barrón-Cedeño, Rosso, and Benedí (2009b)
- Barrón-Cedeño and Rosso (2009b)

Chapter 6

Cross-Language Detection of Text Re-Use and Plagiarism

> I am translating the black and white impressions into another language —"that of colour".
>
> Vincent van Gogh

As previously discussed, cross-language text re-use implies translating some content and re-using it, even after further modification. Just as in the monolingual setting, if no reference is provided, plagiarism may be committed. This act is also known as *translated plagiarism* and, as Charles Reade defined it, *piratical translation* (Mallon, 2001, p. 54). Cross-language plagiarism can be defined as follows (Barrón-Cedeño *et al.*, 2008):

> **Cross-language plagiarism** A text fragment in one language is considered a plagiarism of a text in another language if their contents are considered re-used, no matter they are written in different languages, and the corresponding citation or credit is not included.

Contrary to what might be expected, cross-language plagiarism is far from being a new phenomenon. Samuel Taylor Coleridge, a lecturer from the early 1800s, has been claimed to commit, among his numerous cases of plagiarism, cross-language plagiarism from documents originally written by Scheling, in German (Mallon, 2001, p. 30). Yet another famous case is that of Charles Reade himself, a Victorian novelist with a frequent practice: borrowing his plays from French authors. In 1851 we wrote *The Ladies' Battles*, "adapted" from Augustin Eugène and Ernst Legouvé writings (Mallon, 2001, p. 43).[1]

Back in the 21st century, cross-language borrowing, as monolingual, seems to be on the rise. As mentioned by Corezola Pereira *et al.* (2010a, p. 15–16), of particular interest is the occurrence of translated plagiarism in two particular settings: (a) when $\mathcal{A}$ commits self-plagiarism across languages, aiming at increasing the amount of her publications, and (b) when $\mathcal{A}$ is a student, downloads a text, and translate it for an assignment. We noted

[1] As indicated by Mallon, the copyright agreement between France and England allowed this kind of borrowing by the time.

how likely the second case is in the survey we recently applied in Mexican universities. As seen in Table 2.9 (page 42), around 35% of students declares having plagiarised, at least once, from sources written in a language different than their native one. As the same survey shows, students are not sure whether translation may imply plagiarism, as it does. Chris Caren, iParadigms CEO, defines cross-language academic plagiarism as "[...] students taking existing source material in one language, translating it into the language used at their institution and misrepresenting it as their own work." (Turnitin, 2010).

Cross-language models have been identified as key factors in plagiarism detection since time ago (Clough, 2003; Maurer *et al.*, 2006). Indeed, Maurer *et al.* (2006, p. 1079) state that the "increased ease of access to global and multilingual contents makes detection of translated plagiarism a vital requirement for detection systems". This is supported by Chris Caren, that considers that "translated plagiarism is increasingly common at educational institutions around the world". Cross-language plagiarism detection can be formally defined as follows:

> **Cross-language plagiarism detection** Let d_q be a text written in language L. Let d' to be a text written in language L' ($L \neq L'$). Determine whether d_q contains fragments that have been borrowed, after translation, from $d' \in D'$).

The rest of this chapter is organised as follows. Section 6.1 gives an overview of the prototypical process of automatic cross-language plagiarism detection. The differences to the monolingual process are stressed. Section 6.2 reviews the literature available on the topic, including models originally designed for CLIR and MT, and those specifically proposed to approach the problem we are dealing with. Section 6.3 describes CL-ASA, a model we have proposed to detect plagiarism across languages. Section 6.4 includes the experiments we have carried out at *document level*. It includes comparisons among CL-ASA, the CL-ESA semantic model, and the CL-CNG syntactic model, representatives of the state of the art in cross-language plagiarism detection. Finally, Section 6.5 compares CL-ASA with a syntactic and an MT-based model when aiming at detecting cross-language plagiarism of sentences between distant languages.

Key contributions Cross-Language Alignment-based Analysis (CL-ASA), one of the few models proposed for cross-language plagiarism detection up to date is introduced (Section 6.3). CL-ASA is thoroughly compared to other —state-of-the-art— models on different steps and scenarios of cross-language plagiarism detection considering diverse languages, also less-resourced.

6.1 Cross-Language Plagiarism Detection Process

The prototypical process for cross-language plagiarism detection is the same as the represented in Figure 5.2 (page 114), with some modifications. At *heuristic retrieval*, methods to map the topic or the genre of d_q from L to L' are required. At *detailed analysis*, the model should measure the cross-language similarity between documents in L and L'

(optionally, d_q could be translated into L' for a monolingual process).[2]

6.1.1 Cross-Language Heuristic Retrieval

For the first stage, well-known methods from CLIR can be applied. We identify two similar options: (a) applying a keyword extraction method to obtain the set of representative terms T_q from d_q, map (or translate) them into L', and query them to the index that represents D'; and (b) translating d_q into L' to obtain d'_q, extracting the set of terms $T_{q'}$, and querying them to the index.

The output of this stage is the collection of source candidate documents D'^*, a reduced set with the most likely sources of the potential borrowings in d_q. Note that the exposed options do not necessarily require the use of MT, but other resources as well, such as multilingual thesauri, can be used.

6.1.2 Cross-Language Detailed Analysis

In this stage, the aim is measuring the cross-language similarity between sections of the suspicious document d_q and sections of the candidate documents in D'^*. Potthast *et al.* (2011a) identify four kinds of models that are further discussed in Section 6.2.2: (i) models based on language syntax, (ii) models based on dictionaries, gazetteers, rules, and thesauri, (iii) models based on comparable corpora, and (iv) models based on parallel corpora. Additionally, we add (v) models based on MT.

The alternatives imply a trade-off between retrieval quality and speed. Often more important, they depend on the availability of the necessary resources when dealing with a specific pair of languages.

6.2 Past Work

To the best of our knowledge, before 2008 no technology for cross-language plagiarism detection had been developed. Some efforts have been made in other research directions that could be useful, though. In this section we review: (a) models originally designed to approach other tasks but potentially useful in cross-language plagiarism detection and (b) models specifically designed to deal with this task.

6.2.1 Intrinsic Cross-Language Plagiarism Detection

In this chapter we refer to the problem of cross-language plagiarism detection from an external perspective (i.e., comparing a pair of documents, looking for similar fragments). Nevertheless, models for intrinsic plagiarism detection may be also used to detect cases

[2] If the first stage of the process implies translating every document into one common language, no modification to the monolingual schema is necessary at all.

of cross-language re-use. The main difference to the intrinsic models described in Section 5.1.1 is that here we are interested in determining whether a text fragment s_q was borrowed and translated. The most similar problem (not to say the same one) is that of identifying "translationese". Translationese (Gellerstam, 1985)[3] is identified as the set of "effects of the process of translation that are independent of source language and regard the collective product of this process in a given target language" (Koppel and Ordan, 2011). The automatic detection of this phenomenon is based upon the assumption that "translations from a mix of different languages are sufficiently distinct from texts originally written in the target language" (Koppel and Ordan, 2011); they "exhibit significant, measurable differences" (Lembersky, Ordan, and Wintner, 2011). Baroni and Bernardini (2006) applied a support vector machine (SVM) to detect original from translated documents in Italian. Their results suggest that function words, morphosyntactic categories, personal pronouns, and adverbs are the most relevant features in the discrimination process.[4] The relevance of function words for this kind of tasks is in agreement with the findings of Stamatatos (2011) when dealing with monolingual plagiarism.

Other experiments, this time with a Bayesian logistic regression classifier, found that animate pronouns (e.g. *I*, *we*, *he*) and cohesive markers (e.g. *therefore*, *thus*, *hence*) are particularly distinguishing features for the classification process when aiming at discriminating original from translated English (Koppel and Ordan, 2011, p. 1322). In particular, they found that cohesive markers are more frequent in translations. Baroni and Bernardini (2006, p. 264) suggested that this kind of technique could be applied for multilingual plagiarism detection. However, to the best of our knowledge, no research has been carried out with this purpose yet.

Lembersky *et al.* (2011) computed the language models for originally written and translated texts. They observed, among other interesting results, that the perplexity of the latter model was always the lowest when considering different languages. Whereas this research work was originally intended to generate better language models for MT, it can certainly be considered for cross-language plagiarism detection.[5]

A related phenomenon to cross-language re-use is misuse of machine translation. Somers *et al.* (2006) try to detect a particular kind of miss-conduct: the unauthorised use of machine translation in language students. In a pilot study, they requested students to translate documents from L into L'. They were instructed either to manually perform the translation (with the help of dictionaries and other tools), or make it with an automatic translator (having the chance to briefly modify the outcome). They found that *hapax legomenon* are the best discriminating features between "derived" and "honest translations". Their conclusion is that an amount above 50% of hapax legomenon in a text can trigger suspicion. This idea is extremely similar to those used for authorship attribution and plagiarism detection from a forensic linguistics point of view (cf. Section 2.3). Nevertheless, we do not aim at separating bad from good translations, but simply determining whether d_q was generated by translation.

[3] As seen in Koppel and Ordan (2011, p. 1318)

[4] Interestingly, they report that the SVM performed better than human translators in this task!

[5] Nevertheless, in a real scenario, enough amount of text that allow for the language models computation could not be always at hand.

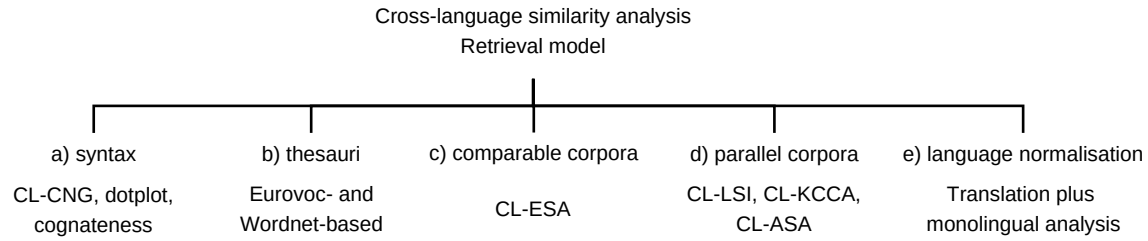


Figure 6.1: Taxonomy of retrieval models for cross-language similarity analysis. Some representative models are included within the family they belong to. Partially derived from Potthast *et al.* (2011a).

6.2.2 External Cross-Language Plagiarism Detection

The problem of external cross-language plagiarism detection has attracted special attention recently (Barrón-Cedeño *et al.*, 2008; Ceska, Toman, and Jezek, 2008; Lee, Wu, and Yang, 2008; Pinto *et al.*, 2009; Potthast *et al.*, 2008a). However, once again, problems previously approached provide with useful techniques for plagiarism detection. Here we review five families of models than can be applied to assess text similarity across languages. An overview, including some representative models is depicted in Fig. 6.1.

6.2.2.1 Models based on Syntax

Plenty of syntactically similar languages exist (e.g. English-French, Spanish-Catalan). Syntax-based models rely on this property and on the appearance of foreign words in a text. Such similarity may be easily reflected when using short terms, such as character n-grams, prefixes, or suffixes.

Character dot-plot is one of these models. As discussed already in Section 3.1.4, the dot-plot model, proposed by Church (1993) for bitexts alignment, considers character 4-grams. If the documents are not parallel and this model is applied, it looks for specific translated, borrowed, fragments.[6] The problem of detecting cross-language text re-use can be considered fairly equivalent to that of extracting text fragments with a high level of comparability (in particular parallel and highly comparable) from a multilingual corpus. That is, we could consider that cross-language text re-use and text alignment are indeed the same task, viewed from two different perspectives.

Character *n*-grams achieve remarkable performance in CLIR for languages with syntactical similarities (Mcnamee and Mayfield, 2004). In this model, a simplified alphabet $\Sigma = \{a, \ldots, z, 0, \ldots, 9\}$ is considered; i.e., any other symbol, space, and diacritic is discarded. The text is then codified into a vector of character n-grams ($n = \{3, 4\}$). And the resulting vectors can be compared by means of the cosine similarity measure. This is one of the most simple models at hand, and still offers very good results when comparing texts in different languages. Indeed, the potential of this model in monolingual comparison was already shown in the examples provided in Tables 2.2, 2.3, and 2.4.

[6] As in the monolingual plagiarism detection approaches of Basile *et al.* (2009) and Grozea *et al.* (2009).

Cognateness is another characterisation originally proposed for bitexts alignment (Simard *et al.*, 1992) that can be exploited in this task. Shingles from documents d_q and d can be extracted according to the criteria described in Section 3.1.4; e.g. selecting the first four characters of the alphabetic tokens, and complete tokens if they contain at least one number. Once again, the similarity between the resulting vectors can be computed as the cosine between their angles.

It is worth noting that these models require little linguistic resources. The former two models need to case fold the texts, eliminate diacritics, and, probably, discard punctuation marks. Additionally, the latter model requires a tokeniser. Nothing else is needed.

Later in this chapter we exploit the model of Mcnamee and Mayfield (2004) for cross-language detailed analysis in plagiarism detection. We call it CL-C3G because we use character 3-grams. The cognateness characterisation of Simard *et al.* (1992) is exploited in Chapter 9, when dealing with cross-language text re-use over Wikipedia. As aforementioned, Church (1993, p. 3) considers that this kind of model can be used with any language using the Latin alphabet. However, as seen in that chapter, existing technology allows for transliterating from one alphabet into another one (e.g. Greek to Latin) and still obtaining good results.

6.2.2.2 Models based on Thesauri

These models can be called cross-language vector space models as well. Their aim is bridging the language barrier by translating single words or concepts such as locations, dates, and number expressions, from L into L'. Thesauri are used in monolingual detection to enrich the documents representation. In the cross-language setting they are considered for mapping two documents written in different languages into a common comparison space. As the terms in a multilingual thesaurus are connected to their synonyms in the different languages —as multilingual synsets— texts can be compared within an "inter-lingual index" (Ceska *et al.*, 2008).

MLPlag (Multilingual plagiarism detector), a prototype developed by Ceska *et al.* (2008), applies this idea. Their model is built upon the EuroWordNet thesaurus[7], which is available in eight European languages (Czech, Dutch, English, Estonian, French, German, Italian, and Spanish). MPLag is aimed at detecting entirely plagiarised documents. As a result, additionally to the thesaurus-based relationships, the position of a word within d_q and d' is considered relevant as well (i.e., if two "equivalent" words appear at the beginning of the corresponding documents rather than one at the beginning and the other at the end, the similarity is considered higher).

The Eurovoc thesaurus has been exploited during the last ten years with a similar purpose by Pouliquen *et al.* (2003); Steinberger, Pouliquen, and Hagman (2002).[8] They aim at searching for document translations within a document's collection. As in the case of Baroni and Bernardini (2006), the authors pointed out this approach could be useful in plagiarism detection.

[7] `http://www.illc.uva.nl/EuroWordNet`
[8] `http://europa.eu/eurovoc`

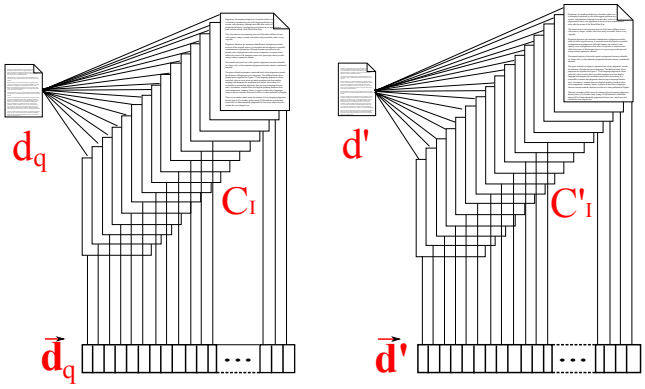


Figure 6.2: Cross-language explicit semantic analysis graphical explanation. d_q (d) is compared to every document $c \in C_I$ ($c' \in C'_I$) composing the similarities vector $\vec{d_q}$ ($\vec{d'}$). Afterwards, $\vec{d_q}$ and $\vec{d'}$ can be compared.

As these models can be mapped into the VSM, their speed is comparable to it. However, thesauri are not always easily found. Moreover, they require significant efforts with respect to disambiguation and domain-specific term translations. Indeed, the difficulty Ceska *et al.* (2008) identified when dealing with plagiarism detection through this kind of approach is the incompleteness of the thesaurus; document's terms may not appear in it.

6.2.2.3 Models based on Comparable Corpora

In this case, the models for similarity assessment are trained over comparable corpora, i.e., a collection of documents C, C' where $c_i \in C$ covers the same topic than $c'_i \in C'$.

Cross-language explicit semantic analysis (CL-ESA) (Potthast *et al.*, 2008a) is a cross-language extension of explicit semantic analysis (ESA) (Gabrilovich and Markovitch, 2007). In ESA, d is represented by its similarities to the documents of a so-called index collection C_I, which are computed with a monolingual similarity model, such as the cosine measure. Given (d_q, d) a similarities' vector is computed as:

$$\vec{d_q} = \{sim(d_q, c) \forall c \in C_I\} \quad (6.1)$$

$$\vec{d} = \{sim(d\ , c) \forall c \in C_I\}\ \text{, i.e.,} \quad (6.2)$$

$$\vec{d_q} = [sim(d_q, c_0), sim(d_q, c_1), \ldots, sim(d_q, c_C)] \qquad \vec{d} = [sim(d, c_0), sim(d, c_1), \ldots, sim(d, c_C)]\ .$$

The i-th element of $\vec{d_q}$ and $\vec{d}$ represent their similarity to a common text c_i. As a result, $sim(d_q, d)$ can be estimated by computing $sim(\vec{d_q}, \vec{d})$.

In the cross-language setting, a comparable corpus $\{C_I, C'_I\}$ ($C_I \in L$, $C'_I \in L'$) is used. Again, similarities' vectors can be computed as in Eqs. (6.1) and (6.2), considering the corresponding corpus C_I (C'_I). The development of Wikipedia as a multilingual comparable corpus during the last years has represented a valuable resource for this model. A graphical explanation of CL-ESA is included in Fig. 6.2.

6.2.2.4 Models based on Parallel Corpora

In this case, the models for similarity assessment are trained over parallel corpora, i.e., a collection of documents C, C' where $c \in C$ is a translation of $c' \in C'$.

Cross-language latent semantic indexing (CL-LSI) is an extension to latent semantic indexing (LSI) (Dumais, Letsche, Littman, and Landauer, 1997; Littman, Dumais, and Landauer, 1998). LSI aims at extending the retrieval model by considering not only the terms in a document, but their concept. It do so by analysing the co-occurrences of a word in a document. In a feature space reduced by means of singular value decomposition (SVD), "words that occur in similar contexts are near each other" (Littman *et al.*, 1998). As a result, the similarity between a pair of words is not computed between themselves, but respect to the set of words they appear with, stored in the reduced feature space.

When latent semantic indexing is applied to a parallel corpus, the "equivalent" terms $t \in c$, $t \in c'$ will have an identical representation. The reason is that being c and c' exact translations, the co-occurrences will be practically equal. This is how a "language independent representation" is generated, through which similarity can be assessed across languages, without translating them.

One of the main disadvantages of CL-LSI is that it requires generating a matrix of $m \times n$, where m is the number of terms and n is the number of documents in the collection. Moreover, the SVD process is computationally expensive.

Cross-language kernel canonical correlation analysis (CL-KCCA) is another model that aims at representing conceptual information across languages by obtaining the terms correlations in a parallel corpus (Vinokourov, Shawe-Taylor, and Cristianini, 2003).

Parliamentary proceedings and other official documents generated in multilingual settings (for instance, countries with multiple official languages such as Canada or Switzerland or the European Union) are probably the biggest sources of parallel documents which can be exploited by this kind of models.

In Section 6.3 we describe the CL-ASA parallel corpus-based model we have proposed for cross-language plagiarism detection.

6.2.2.5 Models based on Machine Translation

Whereas the previous models use the principles (and resources) of MT, they do not perform any actual translation of d_q or d'. Many models for cross-language text re-use detection do apply MT directly to the analysed documents, though. They are based upon the principle of simplifying the problem by making it monolingual. Indeed, this idea has gained popularity in recent years.

Language normalisation is one of the most common pre-processing strategies in CLIR and, in particular, in cross-language plagiarism detection. The idea is coming out

with a representation where every document, either d_q or $d \in D$, is written in the same language. Corezola Pereira *et al.* (2010a) proposes using English as the base language. The reasons they expose for this decision are twofold: (*i*) most of the contents on the Web are published in English and (*ii*) the amount of translation tools between any other language and English are the most commonly available.

Firstly, they apply a language detector to determine the most likely language d (or d_q) is written in. If a document is not written in English, it is translated into it. Secondly, the detection process is completely monolingual, and any of the models described in Chapter 5 could be applied. The same schema has been followed by Corezola Pereira, Moreira, and Galante (2010b), Gottron (2010), Nawab, Stevenson, and Clough (2010), and Vania and Adriani (2010) when dealing with the Spanish-English and German-English plagiarism cases in the PAN-PC-10 corpus (cf. Section 7.2). Indeed, this is a strategy already "predicted" by Maurer *et al.* (2006).

Web-based cross-language models have been proposed as well, built within the same principles of language normalisation. Kent and Salim (2009, 2010) propose using the Google translation and search APIs for this purpose.[9]

In a first stage, they translate d_q from Malay (their "suspicious" language) into English (their "reference" language"). Stopwords from the resulting text are discarded and the tokens are stemmed. The resulting text is queried to Google. Once a set of related documents D' is retrieved, the detailed analysis is carried out.

Multiple translations have been considered as well. Muhr *et al.* (2010) tried using just a part of the machine translation process when aiming at detecting plagiarism between Spanish-English and German-English at PAN (cf. Section 7.2.1.1). Instead of using the final translation, they consider the output of the translation model only, i.e., the one in charge of obtaining every likely translation for the elements in the text.

Their word-based model was built with the BerkeleyAligner software,[10] using the Europarl parallel corpus[11]. Every token $t \in L$ is substituted by (*a*) up to five translation candidates $t'_1, t'_2, \ldots, t'_5 \in L'$, or (*b*) t itself if no possible translation exists in the dictionary. This enriched representation is after used to query the collection of source documents D'.

The model of Muhr *et al.* (2010) may be considered as one based on parallel corpora as well. However as it actually performs a translation, we include it among those based on MT.

[9]The authors stress the advantage of using Google translation API, as it was free. Nevertheless, according to `http://code.google.com/apis/language/translate/overview.html`, as of December 1st, 2011 this is a paid service (last visited, December 2011).

[10]`http://code.google.com/p/berkeleyaligner/` (last visited, December 2011).

[11]`http://www.statmt.org/europarl`

6.3 Cross-Language Alignment-based Similarity Analysis

Here we describe the cross-language alignment-based similarity analysis model (CL-ASA), one of the few models based on parallel corpora that has been specifically proposed for automatic plagiarism detection (Barrón-Cedeño *et al.*, 2008; Pinto *et al.*, 2009). Its aim is estimating the likelihood of two texts of being valid translations of each other.[12] CL-ASA is inspired by the principles of statistical MT. It combines a model often used for corpora alignment with a model used in translation. We already described the principles CL-ASA is based on in Section 3.3.2.2, when discussing the statistical models for similarity assessment. Refer to that section to revise its two composing parts: the *length model* and the *translation model*.

In brief, the length model aims at determining whether the length of a document (fragment) d_q corresponds to the expected length of a valid translation from d'. The translation model, in turn, aims at estimating how likely is that the contents (words) in d_q are valid translations of those in d'.

We have learnt the parameters of the length models, known as *length factors* for diverse languages and from different corpora. In (Potthast *et al.*, 2011a) we estimated length factors from the JRC-Acquis corpus (Steinberger, Pouliquen, Widiger, Ignat, Erjavec, Tufis, and Varga, 2006) for five language pairs: English-{German, Spanish, French, Dutch, and Polish}. Later we estimated language models for two more pairs: Basque-Spanish and Basque-English (Barrón-Cedeño *et al.*, 2010c).[13]

The length factors are no other than the mean (μ) and the standard deviation (σ) of the character lengths between translations of texts from L into L'. Therefore, computing them is not a hard problem; μ and σ are substituted in Eq. (3.31) (page 70) to compose the length model, which approximates a normal distribution. The length factors for the afore mentioned language pairs are included in Table 6.1.[14] Figure 6.3 includes a graphical illustration of the distributions described by the different length models. Given d_q, if the length of d' is not the expected it will result in a point far from the Gaussian mean, resulting in a low probability.

The second element of CL-ASA is the translation model, which depends on a bilingual statistical dictionary. We learnt the corresponding dictionaries for the implied languages on the basis of the IBM model one (cf. Section 3.3.2.2 and Appendix A) (Barrón-Cedeño *et al.*, 2010c; Potthast *et al.*, 2011a). A few examples of the obtained dictionary entries, in particular for English-Basque, are included in Table 6.2.

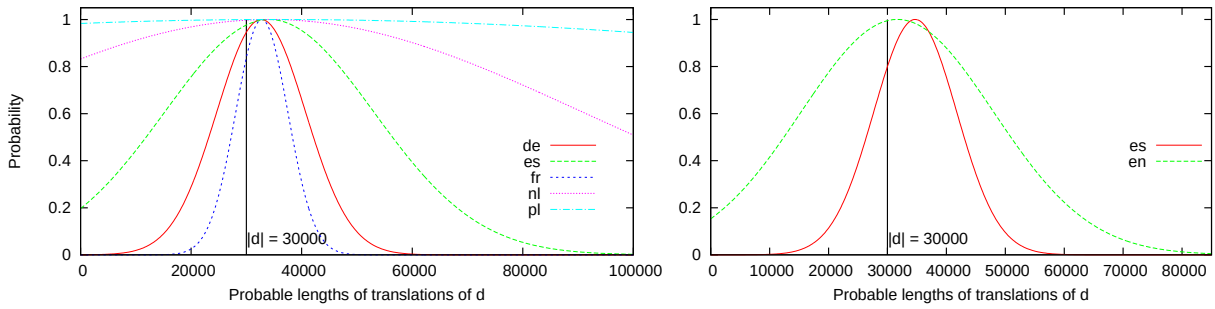
[12] Similar models have been proposed for extraction of parallel sentences from comparable corpora (Munteanu, Fraser, and Marcu, 2004).

[13] This time the corpora used are: *Software*, an English-Basque translation memory of software manuals generously supplied by Elhuyar Fundazioa (`http://www.elhuyar.org`), and *Consumer*, a corpus extracted from a consumer oriented magazine that includes articles written in Spanish along with their translations into Basque, Catalan, and Galician (`http://revista.consumer.es`).

[14] We use these values in our experiments described in Section 6.4 and 6.5 and also over the cross-language partition of the PAN-PC-11 corpus (cf. Section 7.5).

Table 6.1: Estimated length factors for the language pairs L–L', measured in characters. A value of $\mu > 1$ implies $|d| < |d'|$ for d and its translation d'. de=German, en=English, es=Spanish eu=Basque, fr=French nl=Dutch, pl=Polish.

Parameter	en-de	en-es	en-fr	en-nl	en-pl	en-eu	es-eu
μ	1.089	1.138	1.093	1.143	1.216	1.0560	1.1569
σ	0.268	0.631	0.157	1.885	6.399	0.5452	0.2351



(a) $d \in$ English (b) $d \in$ Basque

Figure 6.3: Length model distributions that quantify the likelihood whether the length of the translation of d into the considered languages is larger than $|d|$. In the examples d is an English (a) or Basque (b) document of 30 000 characters (vertical line), corresponding to 6,600 words. The normal distributions represent the expected lengths of its translations.

6.4 Document Level Cross-Language Similarity Experiments

For our document level cross-language experiments we selected three different cross-language similarity estimation models, namely: CL-ESA, CL-ASA, and CL-CNG. The reason is that all of them are reported to provide a reasonable retrieval quality, they require no manual fine-tuning, few cross-language resources, and they can be scaled to work in a real-world setting. A comparison of these models is also interesting since they represent different paradigms for cross-language similarity assessment.

x_{eu}	y_{en}	$p(x,y)$	x_{eu}	y_{en}	$p(x,y)$
beste	another	0.288	**beste**	**other**	**0.348**
dokumentu	**document**	**0.681**	batzu	some	0.422
makro	**macro**	**0.558**	**ezin**	**not**	**0.179**
ezin	cannot	0.279	izan	is	0.241
izan	the	0.162	atzi	access	0.591
.	**.**	**0.981**			

Table 6.2: Entries in a statistical bilingual dictionary (for Basque-English). We include the vocabulary of the sentence *"beste dokumentu batzuetako makroak ezin dira atzitu."* and its translation *"macros from other documents are not accessible."*. Relevant entries for the example are in bold.

6.4.1 Corpora for Model Training and Evaluation

To train the retrieval models and to test their performance we extracted large collections from the parallel corpus JRC-Acquis and the comparable corpus Wikipedia.[15] The JRC-Acquis Multilingual Parallel Corpus comprises legal documents from the European Union which have been translated and aligned with respect to 22 languages (Steinberger *et al.*, 2006). The Wikipedia encyclopedia is considered to be a comparable corpus since it comprises documents from more than 200 languages which are linked across languages in case they describe the same topic (Potthast *et al.*, 2008a).

From these corpora only those documents are considered for which aligned versions exist in all of the languages we include in these experiments: Dutch, English, French, German, Polish, and Spanish (nl, en, fr, de, pl, es). JRC-Acquis contains 23,564 such documents, and Wikipedia contains 45,984 documents, excluding those articles that are lists of things or which describe a date.[16]

The extracted documents from both corpora are divided into a training collection that is used to train the respective retrieval model, and a test collection that is used in the experiments (4 collections in total). The JRC-Acquis test collection and the Wikipedia test collection contain 10,000 aligned documents each, and the corresponding training collections contain the remainder. In total, the test collections comprise 120,000 documents: 10,000 documents per corpus $\times$ 2 corpora $\times$ 6 languages. CL-ESA requires the comparable Wikipedia training collection as index documents, whereas CL-ASA requires the parallel JRC-Acquis training collection to train bilingual dictionaries and length models. CL-C3G requires no training at all.[17]

6.4.2 Experimental Setup

The experiments are based on those of Potthast *et al.* (2008a): let $d_q \in L$ be a query document from a test collection D, let $D' \in L'$ be the documents aligned with those in D, and let d'_q denote the document that is aligned with d_q. The following experiments have been repeated for 1,000 randomly selected query documents with all three retrieval models on both test collections, averaging the results.

Experiment 1: Cross-language ranking Given d_q, all documents in D' are ranked according to their cross-language similarity to d_q; the retrieval rank of d'_q is recorded. Ideally, d'_q should be on the first or, at least, on one of the top ranks.

Experiment 2: Bilingual rank correlation Given a pair of aligned documents $d_q \in D$ and $d'_q \in D'$, the documents from D' are ranked twice: (i) with respect to their cross-language similarity to d_q using one of the cross-language retrieval models, and, (ii) with respect to their monolingual similarity to d'_q using the vector space model. The top 100 ranks of the two rankings are compared using Spearman's ρ, a rank correlation coefficient which measures the agreement of rankings as a value between -1 and 1. A value of -1

[15] These documents can be downloaded from `http://www.dsic.upv.es/grupos/nle/downloads.html`.

[16] If only pairs of languages are considered, many more aligned documents can be extracted from Wikipedia, e.g., currently more than 200,000 between English and German.

[17] We use character 3-gram, hence CL-CNG becomes CL-C3G.

implies a "perfect negative relationship" whereas a value of 1 implies a "perfect positive relationship"; 0 implies that no relationship exists (cf. Vaughan (2001, pp. 140–143)). This experiment relates to comparing a monolingual reference ranking to a cross-language test ranking.

Experiment 3: Cross-language similarity distribution This experiment contrasts the similarity distributions of comparable documents and parallel documents.

Our evaluation is of realistic scale: it relies on 120,000 test documents which are selected from the corpora JRC-Acquis and Wikipedia, so that for each test document highly similar documents are available in all of the six languages. More than 100 million similarities are computed with each model.

6.4.3 Results and Discussion

Experiment 1: Cross-language ranking This experiment resembles the situation of cross-language plagiarism in which a document (fragment) is given and its translation has to be retrieved from a collection of documents (fragments). The results of the experiment are shown in Fig. 6.4.

It follows that CL-ASA has in general a large variance in its performance, while CL-ESA and CL-C3G show a stable performance across the corpora. Remember that JRC-Acquis is a parallel corpus while Wikipedia is a comparable corpus, so that CL-ASA seems to be working much better on "exact" translations than on comparable documents. On the contrary, CL-ESA and CL-C3G work better on comparable documents than on translations. An explanation for these findings is that the JRC-Acquis corpus is biased to some extent; it contains only legislative texts from the European Union and hence is quite homogeneous. In this respect both CL-ESA and CL-C3G appear much less susceptible than CL-ASA, while the latter may perform better when trained on a more diverse parallel corpus. The Polish portion of JRC-Acquis seems to be a problem for both CL-ASA and CL-C3G, but less so for CL-ESA. However, CL-ASA still clearly outperforms the other two models when dealing with translations.

Experiment 2: Bilingual Rank Correlation This experiment can be considered as a standard ranking task where documents have to be ranked according to their similarity to a document written in another language. The results of the experiment are reported as averaged rank correlations in Table 6.3.

As in Experiment 1, CL-ASA performs well on JRC-Acquis and unsatisfactory on Wikipedia. In contrast to Experiment 1, CL-ESA performs similar to both CL-C3G and CL-ASA on JRC-Acquis with respect to different language pairs, and it outperforms CL-ASA on Wikipedia. Unlike in the first experiment, CL-C3G is outperformed by CL-ESA. With respect to the different language pairs, all models show weaknesses, e.g., CL-ASA on English-Polish and, CL-ESA as well as CL-C3G on English-Spanish and English-Dutch. It follows that CL-ESA is more applicable as a general purpose retrieval model than are CL-ASA or CL-C3G, while special care needs to be taken with respect to the involved languages. We argue that the reason for the varying performance is rooted in the varying quality of the employed language-specific indexing pipelines and not in

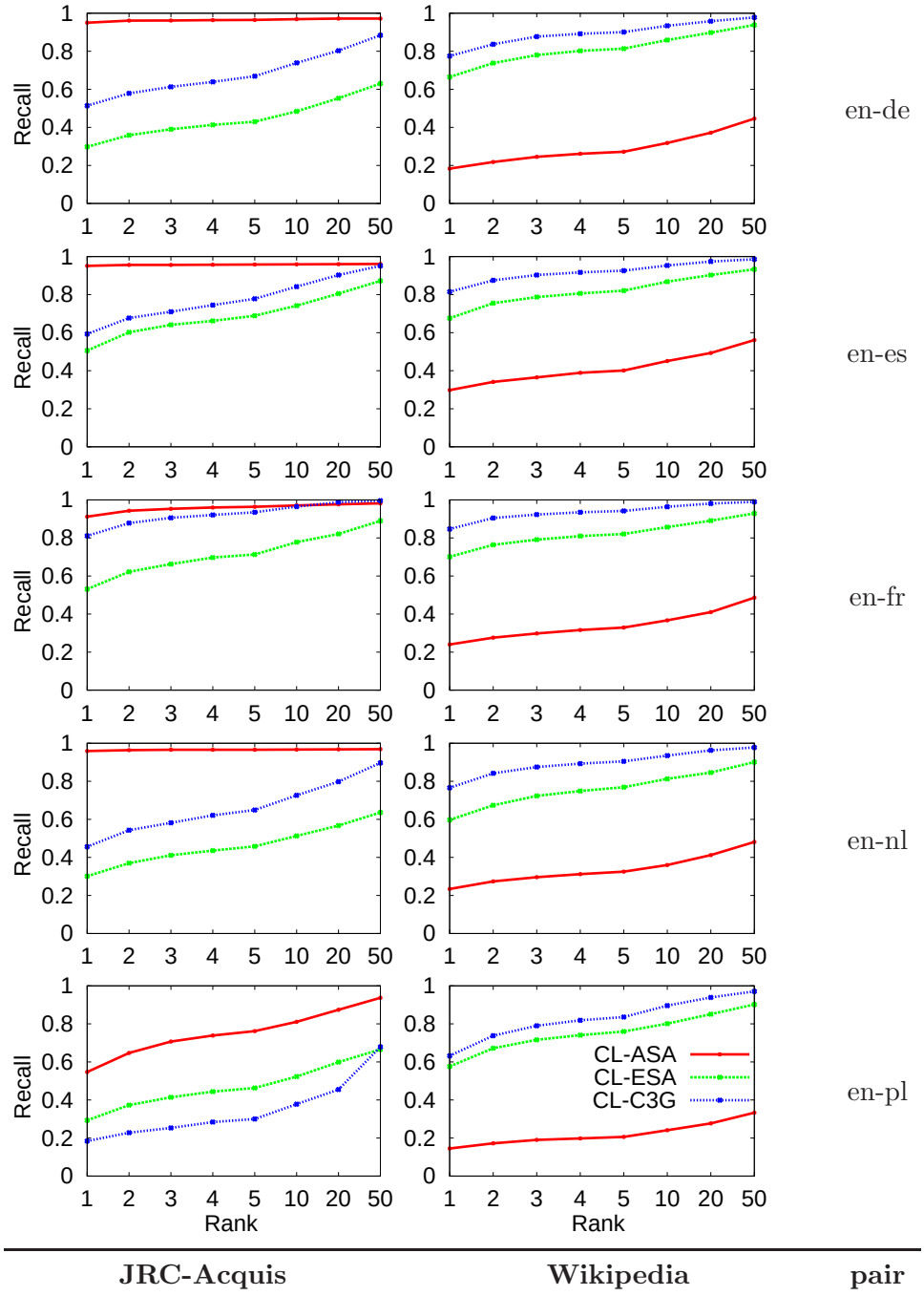


Figure 6.4: Results of Experiment 1 for the cross-language retrieval models. The curves represent values of recall at k.

the retrieval models themselves.

Experiment 3: Cross-Language Similarity Distribution This experiment shall give us an idea about what can be expected from each retrieval model; the experiment cannot directly be used to compare the models or to tell something about their quality. Rather, it tells us something about the range of cross-language similarity values one will measure when using the model, in particular, which values indicate a high similarity

Table 6.3: Results of Experiment 2 for the cross-language retrieval models. The values represent averaged rank correlations respect to the monolingual retrieval exercise.

Pair	JRC-Acquis			Wikipedia		
	CL-ASA	CL-ESA	CL-C3G	CL-ASA	CL-ESA	CL-C3G
en-de	0.47	0.31	0.28	0.14	0.58	0.37
en-es	0.66	0.51	0.42	0.18	0.17	0.10
en-fr	0.38	0.54	0.55	0.16	0.29	0.20
en-nl	0.58	0.33	0.31	0.14	0.17	0.11
en-pl	0.15	0.35	0.15	0.11	0.40	0.22

and which values indicate a low similarity. The results of the experiment are shown in Fig. 6.5 as plots of ratio of similarities-over-similarity intervals.

Observe that the similarity distributions of CL-ASA have been plotted on a different scale than those of CL-ESA and CL-C3G: the top x-axis of the plots shows the range of similarities measured with CL-ASA, the bottom x-axis shows the range of similarities measured with the other models. This is necessary since the similarities computed with CL-ASA are not normalised. It follows that the absolute values measured with the three retrieval models are not important, but the order they induce among the compared documents is. In fact, this holds for each of retrieval models, be it cross-language or not. This is also why the similarity values computed with two models cannot be compared to one another: e.g. the similarity distribution of CL-ESA looks "better" than that of CL-C3G because it is more to the right, but in fact, CL-C3G outperforms CL-ESA in Experiment 1.

The results of our evaluation indicate that CL-C3G, despite its simple approach, is the best choice to rank and compare texts across languages if they are syntactically related. CL-ESA almost matches the performance of CL-C3G, but on arbitrary pairs of languages. CL-ASA works best on "exact" translations but does not generalize well.

The results obtained with CL-ASA are not surprising, as the length based models are designed to detect exact translations (Manning and Schütze, 2002, p. 471-474). It is interesting to look at these results from the perspective of plagiarism detection. CL-ASA performs better with exact translations, a borrowing method that is very likely to be applied by a plagiarist. The rest of models perform better with a comparable corpora, however, at what extent the Wikipedia articles in different languages are co-derived from each other is unknown.[18]

CL-ESA seems to be robust with respect to different languages but tends to be a more topical similarity measure. For related languages, CL-C3G shows to be the option to consider. This led us to carry out another experiment considering less related languages, one of them under resourced.

[18]Cf. Section 9.4 to see our preliminary efforts on measuring this phenomenon.

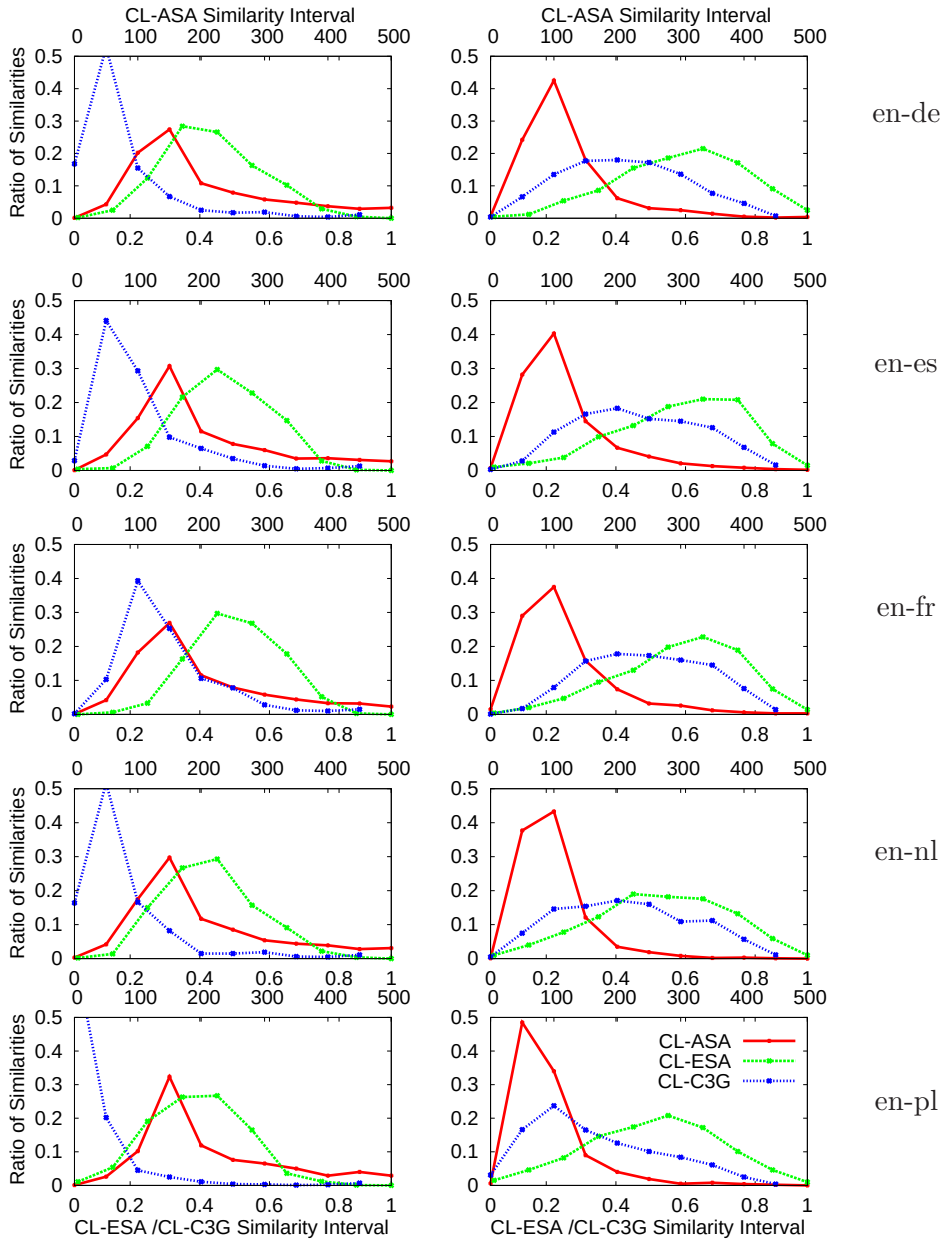


Figure 6.5: Results of Experiment 3 for the cross-language retrieval models. Plots represent the ratio of similarities-over-similarity intervals.

6.5 Sentence Level Detection across Distant Languages

The most of the language pairs used in the experiments of Section 6.4 are related, whether because they have common predecessors or because a large proportion of their vocabularies share common roots. In fact, the lower syntactical relation between the en-pl pair caused a performance degradation for CL-C3G, and for CL-ASA to a lesser

The Party of European Socialists (PES) is a European political party comprising thirty-two socialist, social democratic and labour parties from each European Union member state and Norway.

El Partido Socialista Europeo (PSE) es un partido político pan-europeo cuyos miembros son de partidos socialdemócratas, socialistas y laboristas de estados miembros de la Unión Europea, así como de Noruega.

Europako Alderdi Sozialista Europar Batasuneko herrialdeetako eta Norvegiako hogeita hamahiru alderdi sozialista, sozialdemokrata eta laborista biltzen dituen alderdia da.

Figure 6.6: First sentences from common Wikipedia articles in different languages. We consider the articles "Party of European Socialists" (*en*),"Partido Socialista Europeo" (*es*), and "Europako Alderdi Sozialista" (*eu*) Wikipedia (2010b).

extent. In order to confirm whether the closeness among languages is an important factor, in this section we work with more distant language pairs: English-Basque and Spanish-Basque.

Moreover, cross-language plagiarism may occur more often when the target language is a less resourced one[19], as is the case of Basque. Basque is a pre-indoeuropean language with less than a million speakers in the world and no known relatives in the language families (Wikipedia, 2010a). Still, Basque shares a portion of its vocabulary with its contact languages (Spanish and French). Therefore, we decided to work with two language pairs: Basque with Spanish, one of its contact languages, and with English, perhaps the language with major influence over the rest of languages in the world. Although the considered pairs share most of their alphabet, the vocabulary and language typologies are very different. For instance Basque is an agglutinative language.

In order to illustrate the relations among these languages, Fig. 6.6 includes extracts from the English (*en*), Spanish (*es*) and Basque (*eu*) versions of the same Wikipedia article. The fragments are a sample of the lexical and syntactic distance between Basque and the other two languages. In fact, these sentences are completely co-derived and the corresponding entire articles are a sample of the typical imbalance in text available in the different languages (around $2,000$, $1,300$, and only 100 words are contained in the *en*, *es*, and *eu* articles, respectively).

Here we compare three cross-language similarity analysis methods: (*i*) a machine translation-based model: translation followed by monolingual analysis (T+MA from now onwards); (*ii*) a syntax based model: CL-C3G, and a (*iii*) a parallel corpus-based model: CL-ASA. To the best of our knowledge, no work has been done in cross-language similarity analysis considering less resourced languages, nor comparing the selected models.

As we have not applied T+MA before, it deserves further precisions. As multiple translations from d_q into d'_q are possible, performing a monolingual similarity analy-

[19]Less resourced language is that with a low degree of representation on the Web (Alegria, Forcada, and Sarasola, 2009). Whereas the available text for German, French or Spanish is less than for English, the difference is more dramatic with other languages such as Basque.

sis based on "traditional" techniques, such as those based on word n-grams comparison (Broder, 1997; Schleimer *et al.*, 2003), is not an option. Instead, we take the approach of the bag-of-words, which has shown good results in the estimation of monolingual text similarity Barrón-Cedeño *et al.* (2009a). Words in d'_q and d' are weighted by the standard $tf\text{-}idf$, and the similarity between them is estimated by the cosine similarity measure.

This time CL-ESA is not included. The reasons are twofold. On the one hand, this experiment aims at detecting exact translations, and in our previous experiments CL-ASA showed to outperform it on language pairs whose alphabet or syntax are unrelated (cf. Section 6.4). This is precisely the case of *en-eu* and *es-eu* language pairs. On the other hand, we consider that the amount of Wikipedia articles in Basque available for the construction of the required comparable corpus is insufficient to build a proper corpus for CL-ESA.

6.5.1 Experimental Setup

In these experiments we use two parallel corpora: *Software*, an *en-eu* translation memory of software manuals generously supplied by Elhuyar Fundazioa[20]; and *Consumer*, a corpus extracted from a consumer oriented magazine that includes articles written in Spanish along with their Basque, Catalan, and Galician translations[21] (Alcázar, 2006). *Software* includes 288,000 parallel sentences; 8.66 (6.83) words per sentence in the English (Basque) section. *Consumer* contains 58,202 sentences; 19.77 (15.20) words per sentence in Spanish (Basque). These corpora also reflect the imbalance of text available in the different languages.

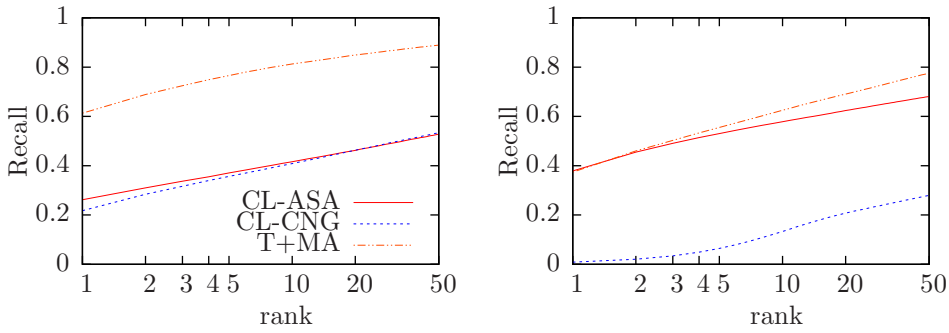
Of high relevance is that the two corpora were manually constructed by translating English and Spanish texts into Basque. This is different from the experiments of Section 6.4. The JCR-Acquis corpus (Steinberger *et al.*, 2006) is a multilingual corpus where, as far as we know, no clear definition of source and target languages exists. And in Wikipedia no specific relationship exists between the different languages in which a topic may be broached. In some cases (cf. Fig. 6.6) they are clearly co-derived, but in others they are completely independent.

We consider d_q and d' to be two entire documents from which plagiarised sentences and their source have to be detected. We work at this level of granularity, and not entire documents, for two main reasons: (i) here we focus on the detailed comparison stage of the plagiarism detection process; and (ii) even a single sentence could be considered a case of plagiarism, as it transmits a complete idea. Note that the task becomes computationally more expensive as, for every sentence, we are looking through thousands of topically-related sentences that are potential sources of d_q, and not only those of a specific document.

Let $d_q \in D_q$ be a plagiarism suspicion sentence and $d' \in D'$ be its source sentence. We consider that the result of the process is correct if, given d_q, d' is properly retrieved, on top of the ranking. A 5-fold cross validation for both *en-eu* and *es-eu* was performed.

[20] `http://www.elhuyar.org`

[21] `http://revista.consumer.es`



(a) es-eu (b) en-eu

Figure 6.7: Evaluation of the cross-language ranking. The curves represent values of recall at k for the three evaluated models and the two language pairs.

Bilingual dictionaries, language and length models were estimated with the (same) corresponding training partitions. The length model is used by CL-ASA, the language model is used by T+MA and the translation model is used by both.[22] The average values for μ and σ are those included in Table 6.1. On the basis of these estimated parameters, an example of length factor for a specific text is plotted in Fig. 6.3.

In the test partitions, for each suspicious sentence d_q, 11,640 source candidate sentences exist for *es-eu* and 57,290 for *en-eu*. This results in more than 135 million and 3 billion comparisons carried out for *es-eu* and *en-eu*, respectively.

6.5.2 Results and Discussion

For the evaluation we consider a standard measure: recall. More specifically recall at k (with $k = [1 \ldots, 50]$). Figure 6.7 plots the average recall value obtained in the 5-folds with respect to the rank position (k).

In both language pairs, CL-C3G obtained worse results than those obtained for English-Polish in Section 6.4: $R@50 = 0.68$ vs. $R@50 = 0.53$ for *es-eu* and 0.28 for *en-eu*. This is due to the fact that neither the vocabulary nor its corresponding roots keep important relations. Therefore, when language pairs have a low syntactical relationship, CL-C3G is not an option. Still, CL-C3G performs better with *es-eu* than with *en-eu* because the first pair is composed of contact languages.

About CL-ASA, the results obtained with *es-eu* and *en-eu* are quite different: $R@50 = 0.53$ for *es-eu* and $R@50 = 0.68$ for *en-eu*. Whereas in the the first case CL-ASA completely outperforms CL-CNG, in the second case the obtained results are comparable. The improvement of CL-ASA obtained for *en-eu* is due to the size of the training corpus available in this case (approximately five times the number of sentences available for *es-eu*). This shows the sensitivity of the model with respect to the size of the available training resources.

Lastly, although T+MA is a simple approach that reduces the cross-language simi-

[22]For the case of T+MA, the training is carried out on the basis of Giza++ (Och and Ney, 2003), Moses (Koehn et al., 2007) and SRILM (Stolcke, 2002).

larity estimation to a translation followed by a monolingual process, it obtained a good performance ($R@50 = 0.77$ for *en-eu* and $R@50 = 0.89$ for es-eu). Moreover, this method proved to be less sensitive than CL-ASA to the lack of resources. This could be due to the fact that it considers both directions of the translation model (*e[n|s]-eu* and *eu-e[n|s]*). Additionally, the language model, applied in order to compose syntactically correct translations, reduces the amount of wrong translations and, indirectly, includes more syntactic information in the process. On the contrary, CL-ASA considers only one direction of the translation model *eu-e[n|s]* and completely disregards syntactical relations between the texts.

Note that better results come at the cost of higher computational demand. CL-C3G only requires easy to compute string comparisons. CL-ASA requires translation probabilities from aligned corpora, but once the probabilities are estimated, cross-language similarity can be computed very fast. T+MA requires the previous translation of all the texts, which can be very costly for large collections and potentially infeasible for "on the fly" applications.

6.6 Chapter Summary

In this chapter we discussed cross-language plagiarism detection, a problem nearly approached that has received recent interest, particularly during the last three years. The chapter started exposing some reasons why translated plagiarism represents nowadays an real problem in academia and science. Afterwards, we described the prototypical process of cross-language plagiarism detection. Although similar to the one followed for external monolingual plagiarism detection, we stressed the differences and extra difficulty that the cross-language issue implies. A review of literature available was included as well. Special attention was paid to what we consider will be addressed in future research: intrinsic cross-language plagiarism detection. A total of five families of similarity models that can be applied to compare texts in different languages, to perform cross-language external plagiarism detection, were reviewed and discussed.

The second part of the chapter starts with the definition of a new model proposed for cross-language plagiarism detection known as CL-ASA. The process for estimating the features requires to assess similarity between documents in different languages without actually translating. In the last two sections, in order to evaluate the model, several experiments are presented, ranging from the detection of translated and comparable documents to the identification of translated sentences. Texts in eight languages were used to analyse how four recently proposed models for automatic plagiarism detection perform.

In the first case, three experiments at document level were carried out, considering both comparable and parallel corpora. CL-ASA was compared to models based on syntax (CL-CNG) and comparable corpora (CL-ESA) in different tasks related to cross-language ranking. Our findings include that CL-CNG and CL-ESA are in general better suited for the task of retrieving topically-similar documents. CL-ASA achieves much better results with professional and automatic translations, which are closer phenomena to cross-language text re-use and plagiarism. Moreover, CL-ASA (and CL-ESA), can be

used with language pairs whose alphabet or syntax are unrelated. CL-CNG can be used with languages with different alphabets only if a transliterator is at hand.

These findings are further supported by the second set of experiments, which considered parallel corpora composed of texts written in distant languages. This time CL-ASA was compared against CL-CNG and a model based on language normalisation followed by a monolingual comparison (T+MA). Our findings were that T+MA obtains the best results; however, its performance depends on a machine translator, which is not always at hand for every language pair. On the other side CL-CNG, that with more related languages offered remarkable results, obtained the worst performance. Better results come at the cost of more expensive processing.

The capabilities of CL-ASA when aiming at detecting the cross-language cases in the PAN-PC-11 corpus are discussed in Chapter 7, after analysing the results obtained by the PAN competition participants. Additionally, the results obtained with other models when dealing with unrelated languages with different alphabets (Latin and Devangari) are analysed in Chapter 9, were we analyse the results of the PAN@FIRE Cross-Language !ndian Text Re-Use Detection competition.

Related publications:

- Potthast, Barrón-Cedeño, Stein, and Rosso (2011a)
- Barrón-Cedeño, Rosso, Agirre, and Labaka (2010c)
- Pinto, Civera, Barrón-Cedeño, Juan, and Rosso (2009)
- Barrón-Cedeño, Rosso, Pinto, and Juan (2008)

Chapter 7

PAN International Competition on Plagiarism Detection

> If you want to do something really big not only you have to have a good technical idea but you have to learn how to sell other people a wanting to do the same thing so that everyone will help.
>
> Vinton Cerf

As noted by Clough (2003), since the development of the first automatic plagiarism detectors, the most of the research work has focussed on detecting borrowing cases within a closed-set of documents (e.g. Bernstein and Zobel (2004), Kang *et al.* (2006), and Barrón-Cedeño and Rosso (2009a)). Some efforts have addressed the problem as a Web-scale process, though (Malcolm and Lane, 2008), (Kent and Salim, 2009). However, in most cases no objective comparison was made among different models in most cases. As a result, a question remained unanswered: *what plagiarism detection model performs best?* The answer to such a question was not clear due to the lack of a standard evaluation framework for automatic plagiarism detection.

As a response, the *International Competition on Plagiarism Detection* was organised within the framework of the workshop *PAN: Uncovering Plagiarism, Authorship and Social Software Misuse*. The main aim of such a competition is setting an evaluation framework in which different models can be objectively compared. The first one was held in 2009 as part of the *Spanish Conference for Natural Language Processing* (SEPLN) (Potthast *et al.*, 2009), whereas the second and third editions were held as one of the *Cross-Language European Forum* (CLEF, now Conference and Labs of the Evaluation Forum) labs (Potthast *et al.*, 2010d, 2011b).[1] All the editions of the competition have been sponsored by Yahoo! Research.

[1]Challenges on automatic Wikipedia vandalism detection in 2010 and 2011 (Potthast and Holfeld, 2011; Potthast, Stein, and Holfeld, 2010c) and authorship identification in 2011 (Argamon and Juola, 2011) have been organised in PAN as well. In 2012 the plagiarism task is focussed on external detection, including two sub-tasks: (*i*) candidate document retrieval (i.e., retrieving a set of candidate source documents from a Web search engine) and (*ii*) detailed comparison (i.e., detect the plagiarised frag-

The three PAN-PC, available up to date, corpora were discussed already in Section 4.2.3. The evaluation measures designed for the competition —granularity, F-measure and special versions of precision and recall— were discussed in Section 4.3.3. Therefore, in this chapter we centre our analysis on participants' approaches and results in the three competitions. Both PAN-PC-10 and PAN-PC-11 included more diverse obfuscation strategies when generating the different re-use cases, including also manual paraphrases (cf. Section 4.2.3.4 and 4.2.3.5). This will allow to investigate what are the major difficulties that plagiarism detection systems would have to face in a real scenario. The results obtained with a plagiarism detector based on word n-grams comparison (such as the one described in Section 5.2) are discussed in Section 7.4. Finally, Section 7.5 approaches the detection of translated cases in the PAN-PC-11 corpus. We apply our CL-ASA model and compare its performance to the ones obtained by the systems that participated in the competition.

When looking at the results obtained in the PAN evaluation framework, it must be observed that PAN is not intended to evaluate models that only uncover unexpectedly similar text fragments. It also intends to determine whether the offsets of the re-used text (and its source) are properly identified, which from the PAN viewpoint represents the entire detection work-flow. As a result, counting with an optimal similarity estimation model is not enough to get a good result; heuristics must be applied to set "plagiarism-original borders". The impact of this finishing step is particularly clear in Section 7.5, where a model is tested on different scenarios of cross-language plagiarism detection.

One of the tracks of the Forum of Information Retrieval Evaluation, PAN@FIRE, focussed on cross-language text re-use detection only. As the corpus we generated for such a challenge was composed of Wikipedia articles (cf. Section 4.2.5), we discuss the results in Chapter 9.

Key contributions The contributions of the author of this dissertation regarding the PAN competition, as partially described in Chapter 4, have been: (*i*) the selection of text sources for the generation of simulated cases of plagiarism and (*ii*) the design of some of the strategies for the cases generation. The rest of the contribution was more focussed on the logistics of the competition. Taking advantage of the PAN framework, CL-ASA, our model proposed for detecting cross-language plagiarism, was tested on different plagiarism detection scenarios, showing to be competitive with models that do not depend on the same translation mechanisms than those with which the cases were generated.

7.1 PAN @ SEPLN 2009

In its first edition, a total of thirteen worldwide research teams took part in the competition. This participation showed that not only plagiarism, but the efforts for its automatic detection have raised in recent years. Most of the teams approached the external detec-

ments from a suspicious document together with their corresponding source fragments). Now intrinsic plagiarism detection is a sub-task of the authorship attribution competition, which includes a new task: sexual predator identification. The last task is quality flaw prediction in Wikipedia, which aims at automatically detecting poor writing style, bad referencing, and other flaws in the encyclopedia.

Table 7.1: Pre-processing, heuristic retrieval, detailed analysis, and post-processing generalisation at the 1st International Competition on Plagiarism Detection. Notation in Table 7.2.

Participant	Step (*0*) Pre-processing				Step (*1*) Heuristic Retrieval		Step (*2*) Detailed Analysis			Step (*3*) Post-processing	
										S discarded if	S_1, S_2 merged if
	case folding	sw removal	diacritics removal	doc. splitting	word n-grams	char. n-grams	word n-grams	char. n-grams	dot-plot	$\lvert s_q \rvert < thres_1$	$\delta(s_1, s_2) < thres_2$
Grozea						16		16	■	■	■
Kasprzak		□	■		5		5				■
Basile					8			7	■		■
Zechner	■	■		■	1		1				■
Vallés							6				
Malcolm					3		3			■	■

tion task only. Moreover, no team tried to detect cases of cross-language plagiarism. In fact, as already discussed, cross-language plagiarism detection has drawn attention just recently (cf. Chapter 6).

7.1.1 Tasks Overview

7.1.1.1 External Detection

For the external analysis, the schema depicted in Fig. 5.2 was followed in most cases, i.e., (*1*) heuristic retrieval: for a suspicious document d_q, the most related documents $D^* \subset D$ are retrieved; (*2*) detailed analysis: d_q and $d \in D^*$ are compared in order to identify specific plagiarism-source candidate fragment pairs; and (*3*) post-processing: bad candidates (very short fragments or not similar enough) are discarded and neighbour text fragments are combined. We consider a preliminary step (*0*) pre-processing: standard (and not so standard) IR pre-processing operations are applied to the texts.[2] This step gathers all shallow linguistic processes. A summary of the parameters considered in the four steps in 2009 is included in Table 7.1.[3] The notation used in this and the rest of tables that overview the intrinsic and external approaches' settings is described in Table 7.2. If terms are composed of words, tokenisation is applied; if characters are considered, it is not (therefore, we do not include this pre-processing in the tables).

(*0*) Pre-processing Not all the participants apply the "standard" pre-processing strategies (e.g. case folding, stemming, stopwords removal). This is somehow justified

[2]We use this schema to discuss the external approaches in Sections 7.2.1.1 and 7.3.1.1.

[3]Participants that did not report on their approaches operation are omitted. Note that such table (as well as the rest presenting this kind of summary), represent only a generalisation of the employed operations.

Table 7.2: Summary of notation for the detection approaches.

Symbol	Description
Participant	Surname of the first member of the participating team.
■	The parameter is applied in the approach.
□	The parameter is applied in a particular —non-standard— way.
number	The value of n.
	Acronyms
sw	Stopword.
!αnum	Non-alphanumeric.
S	Pair of plagiarism suspicion-source detected fragments $\{s_q, s\}$.
Word classes+	Different kinds of words and other features.
Syn. normalisation	Synonymic expansion (with Wordnet).
Lang. normalisation	Language normalisation (in most cases, translation).
$thres_k$	A given threshold.
	Operations
sim	Similarity.
δ	Distance.
$\lvert \cdot \rvert =$	Length of $\cdot$.
	Intrinsic comparison strategy
chunk vs. doc	The chunks are compared to the entire document's representation.
chunk vs. chunk	The chunks are compared to the rest of chunks' representation.

by the results shown in Tables 5.8 to 5.10; pre-processing seems not to represent a very important factor in text re-use detection. For instance, Grozea and Popescu (2011), do not report performing any pre-processing.

The pre-processing that Kasprzak, Brandejs, and Kriřpač (2009) report is diacritics removal and short words removal. The reason behind the first operation is that they aim at creating a multilingual (not cross-language) detection model. The second operation is performed because they are mainly focussed on working with documents in Czech, and short words are mostly prepositions. Zechner *et al.* (2009) use more pre-processing operations, including case folding, stopword removal and even document splitting, at sentence level.

Perhaps one of the most original pre-processing strategies is the one of Basile *et al.* (2009). For heuristic retrieval, they substitute every word in d (d_q) by its length in characters.[4] As a result, they compose "length" rather than word n-grams. This can be considered as a "pseudo-hashing" operation. Whereas the number of collisions for low levels of n is very high, the higher the n the more unique the hashes become. See Fig. 7.1 for a graphical representation of this fact. At detailed analysis, they use a codification inspired by mobile phones `T9` technology.[5]

(*1*) Heuristic Retrieval Kasprzak *et al.* (2009) considered hashed word 5-grams as terms and identify a pair of documents $\{d_q, d\}$ as similar if they share at least twenty terms (the general schema is similar to that of COPS; cf. Section 5.1.2.1). Basile *et al.*

[4]For instance, "this is an example" becomes 4227.

[5]For instance, "example" becomes 3926753.

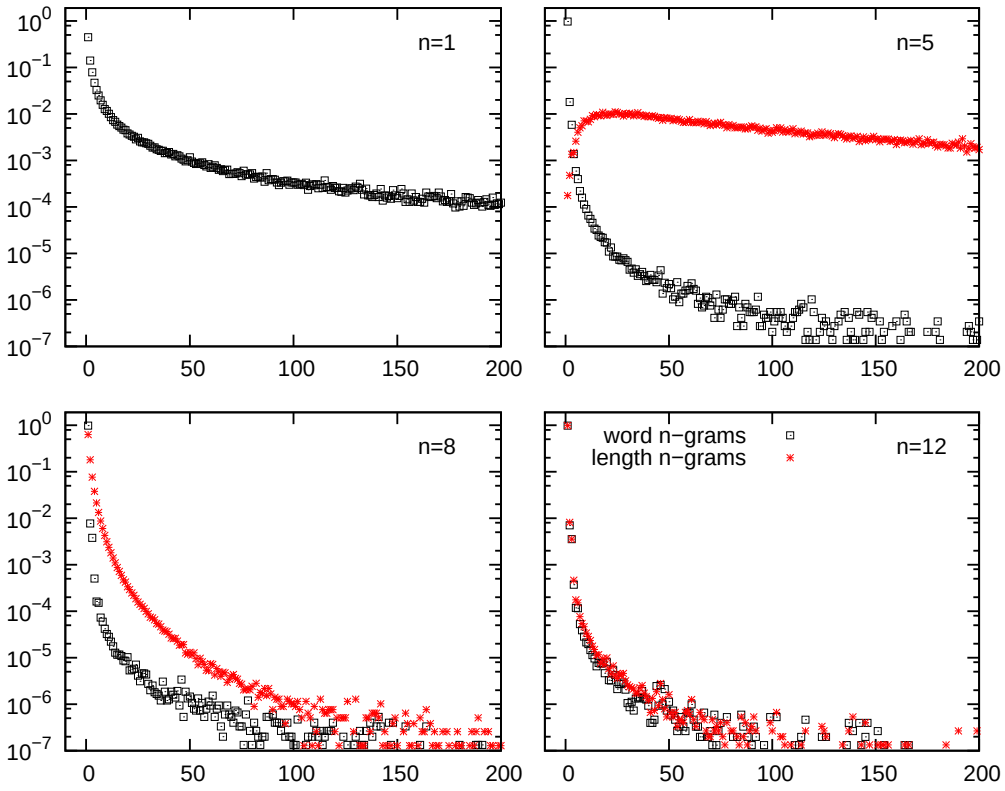


Figure 7.1: Frequency distributions for word n-grams and length n-grams, for $n = \{1, 5, 8, 12\}$ in the PAN-PC-09 (in Barrón-Cedeño *et al.* (2010d)). The number of occurrences lies on the $x-$axis, with the corresponding percentage of n-grams on the $y-$axis. The length $n-$gram distribution converges to the one of word $n-$grams as n grows. No stars appear in the first plot because we show up to 200 occurrences only, which is lower than the frequency of any possible 1-gram of length encoded text in a representative corpus.

(2009) used number 8-grams, and compared the resulting vectors with a real valued similarity model. They decided to compose D^* by the ten documents in D with highest similarities respect to d_q. Scherbinin and Butakov (2009) applied Winnowing fingerprinting (cf. Section 5.1.2.1). They retrieved every document $d \in D$ that shared at least one shingle with d_q.

The approach of Zechner *et al.* (2009) is very different. Instead of performing a comparison of d_q to the entire corpus D, they performed a clustering-based organisation of D's sentences (cf. Section 5.1.2.2, in particular page 127). Afterwards, a large amount of sentences $s \in D$ are represented by the centroid they belong to. A sentence $s_q \in d_q$ is compared to the centroids of D's clusters and D^* is composed of the documents whose sentences are in the two most similar clusters.

Grozea and Popescu (2011) approached the problem from a different perspective. Instead of looking for the potential source of d_q, they looked for the potential plagiarism of d. Intuitively this makes sense as they are somehow imitating the process followed by the borrower when plagiarising (i.e., from the source to the plagiarism). The similarities between every $d \in D$ and $d_q \in D_q$ are computed and only the fifty-one most similar documents to d_q are considered for further analysis.

(*2*) Detailed analysis For this step, very diverse strategies were applied as well. In particular, Grozea and Popescu (2011) proposed the so called Encoplot. This is a variation of dot-plot technique (cf. Section 3.3.1.2), where the terms are character 16-grams (nearly equivalent to word 3-grams), and the vectors $\vec{d_q}$ and $\vec{d}$ are sorted for an efficient comparison. Basile *et al.* (2009) applied a dot-plot technique as well, but they considered a smaller representation: character 7-grams.

Kasprzak *et al.* (2009) identified the plagiarised-source chunks as those that: (*i*) share at least twenty terms and (*ii*) the first and last term of a chunk are present in the other one. Once again, Zechner *et al.* (2009) performed a sentence-level analysis to identify specific fragments. They considered a sentence s_q plagiarised from s if $sim(s_q, s) > threshold$. Finally, Scherbinin and Butakov (2009) extended the shingles identified by means of their Winnowing algorithm to left and right as far as the character strings' Levenshtein distance (Levenshtein, 1966) accomplished a given threshold.

(*3*) Post-processing Once the set of candidates $\{s_q, s\}$ were identified, heuristics were applied to improve the quality of the output. For instance, Grozea *et al.* (2009) discarded those pairs that were too short to be considered relevant and iteratively merged contiguous pairs. Basile *et al.* (2009), Kasprzak *et al.* (2009), Scherbinin and Butakov (2009), and Zechner *et al.* (2009) joined adjacent matches if the gaps between them were short enough.

Some well established plagiarism detection systems participated to the competition as well. Malcolm and Lane (2009) used an adaptation of the well-known Ferret system, considering word 3-grams (cf. Section 5.1.2.1). Vallés Balaguer (2009) used WCopyfind[6] considering word 6-grams. Finally, Palkovskii (2009) used a commercial plagiarism detector (no further details were provided).

7.1.1.2 Intrinsic Detection

This approach was tried by fewer teams.[7] The schema depicted in Fig. 5.1 was followed in most cases; i.e., (*1*) document chunking, (*2*) retrieval, (*3*) outlier detection, and (*4*) post-processing. In order to discuss the approaches, we include, once again, a preliminary step: (*0*) pre-processing. A summary of the parameters considered during pre-processing, chunking and outlier detection is included in Table 7.3.[8]

(*0*) Pre-processing The pre-processing operations for intrinsic analysis are slightly different than for external. Stamatatos (2009b) applies case folding only and extracts character 3-grams to characterise d_q, discarding those in which no single letter occurs. Zechner *et al.* (2009) use other kinds of features, in particular average word frequency class, text statistics, part of speech, and closed-class features (cf. Sections 5.1.1.1, 3.4.1, and 3.4.3 to 3.4.4). Therefore, their pre-processing consists of the identification of these features over the text. Finally, Seaward and Matwin (2009) consider the amount of different word categories in the chunks. As a result, they apply POS tagging.

[6] `http://plagiarism.phys.virginia.edu`

[7] This trend remained during the three competitions.

[8] Participants that did not report on their approaches operation are omitted.

Participant	Pre-processing				Chunking		Outlier det.
	case folding	!alnum removal	word classes +	char n-grams	$\|chunk\|$	$\|step\|$	chunk vs. doc.
Stamatatos	■	□		3	1,000 c	200	■
Zechner			■		12 s		■
Seaward			■				■

Table 7.3: Pre-processing, chunking, and outlier detection for intrinsic analysis at the 1st International Competition on Plagiarism Detection. Three stages considered: pre-processing, chunking (lengths of windows and steps in terms of characters c or sentences s), and outlier detection. Notation in Table 7.2.

***(1)* Chunking** In this step d_q is split into fragments s_q in order to further determine whether its contents fit with the entire document. Stamatatos (2009b) opts for generating fixed length chunks, in particular of 1,000 characters. Zechner *et al.* (2009), in turn, consider chunks of twelve sentences.

***(2)* and *(3)* Retrieval and outlier detection** All the approaches consider comparing the fragment s_q to the representation of the entire document d_q. If the profile of a chunk is particularly different to the rest, borrowing is suspected. The characterisation of Stamatatos (2009b) is based on character n-gram profiles, which are compared with the nd_1 measure (cf. Section 5.1.1.2). Zechner *et al.* (2009) base their approach on a classification process considering the features obtained during pre-processing. Finally, Seaward and Matwin (2009) apply the model described in Section 5.1.1.3, based on Kolmogorov complexity measures.

7.1.2 Results and Discussion

The corpus used in this edition of the competition is the PAN-PC-09 (cf. Section 4.2.3.3). The evaluation results for the external approaches are summarised in Fig. 7.2. The most successful approach is that of Grozea *et al.* (2009), closely followed by Kasprzak *et al.* (2009) and Basile *et al.* (2009). For the analysis, we first focus on F-measure. Four teams manage to obtain $F > 0.6$, either using dot-plot, indexing or Winnowing. This is remarkable if we consider that 10% of cases were translated and the approaches did not aspire to detect them.

The highest precision is obtained by Scherbinin and Butakov (2009) and their Winnowing-based approach ($prec = 0.75$). This result was expected as their shingles are as long as 50 characters. This rigid comparison strategy does not affect recall so dramatically, letting them to still obtain a competitive $rec = 0.53$. On the other side, the word 5-grams shingling of Kasprzak *et al.* (2009) obtained the highest recall ($rec = 0.70$) at the cost of the lowest precision among the competitive approaches ($prec = 0.56$). This is precisely the behaviour expected when considering a more flexible representation strategy.

The heuristics applied by the top three participants for post-processing allow them for obtaining good levels of granularity and precision. Discarding short plagiarism-source candidates and merging neighbouring cases results in a cleaner output for the user. As

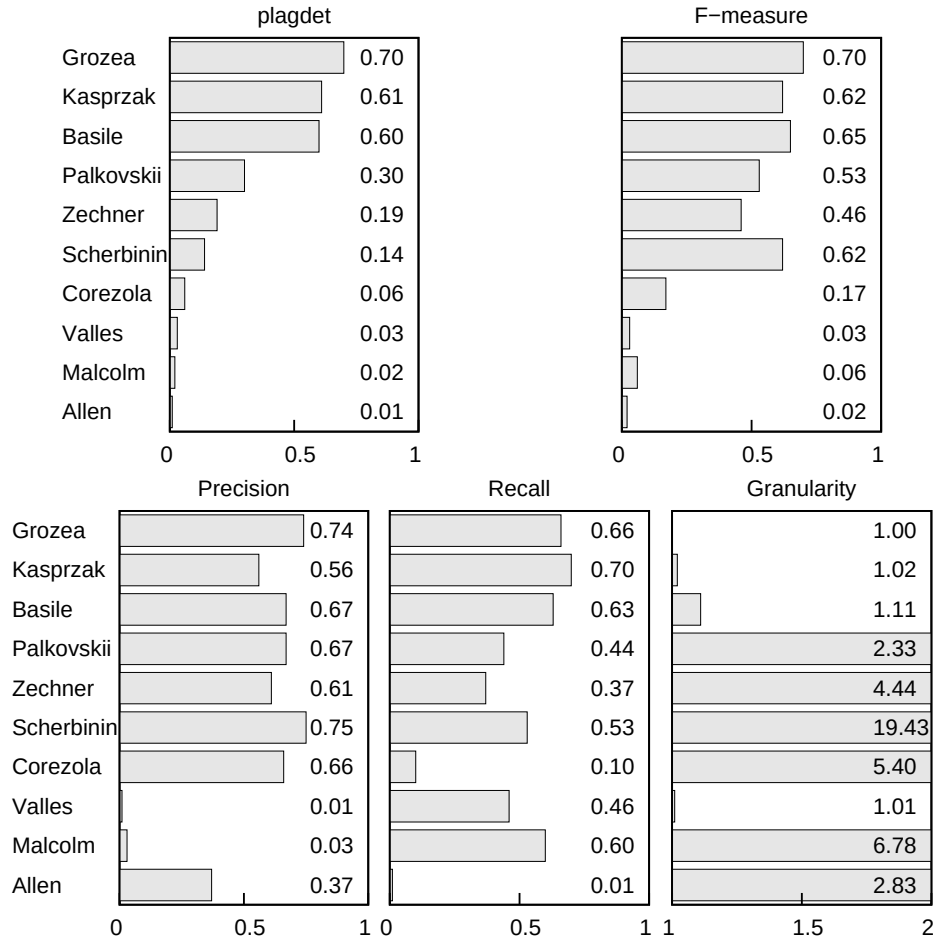


Figure 7.2: Results of *external* plagiarism detection at the 1st International Competition on Plagiarism Detection.

seen in the granularity figure, not many participants paid attention to this issue.[9] For instance, Scherbinin and Butakov (2009) obtained an F-measure as good as the second best in the overall ranking, but a high granularity causes them to fall to the mid-rank zone.

The results for the intrinsic detection task are summarised in Fig. 7.3. Hagbi and Koppel[10] practically considered that every text fragment had been plagiarised, so establishing the baseline for this task (Potthast *et al.*, 2009, p. 8). The character n-gram profiles of Stamatatos (2009b) showed to be the best in this case (indeed, variations of this approach have been applied in the following competitions by other participants). The multi-feature classification model of Zechner *et al.* (2009) is still competitive in terms of F-measure, but a slightly higher granularity causes it to become the third best.

[9]This was probably caused by the fact that the evaluation schema was being defined while the competition was already running.

[10]As no paper was submitted, no reference about the model they applied is available.

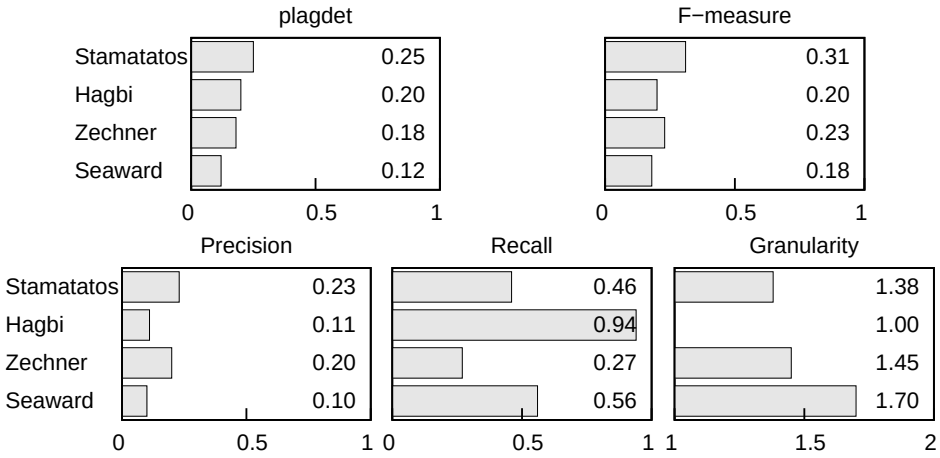


Figure 7.3: Results of *intrinsic* plagiarism detection at the 1st International Competition on Plagiarism Detection.

7.2 PAN @ CLEF 2010

Eighteen participants took part of the 2nd International Competition on Plagiarism Detection (Potthast *et al.*, 2010d). This time, some teams turned their attention to the translated plagiarism cases. In 2009 one of the biggest difficulties the participants pointed out was the size of the corpus. They considered it too large and hard to deal with. This year more efficient strategies were proposed, some of which were able to process the entire corpus in less than one hour (Rodríguez Torrejón and Martín Ramos, 2010).

7.2.1 Tasks Overview

7.2.1.1 External Detection

This task received even more attention this year compared to the intrinsic one. We identify 2010 as the time when the knowledge-based post-processing —which aimed at discarding cases of re-use including a proper citation (cf. Section Chapter 5, in particular page 114)—, was conceptually switched into a heuristic post-processing, which aims at providing a cleaner output to the user (Potthast *et al.*, 2010d). The parameters applied at the four different strategies is summarised in Table 7.4.

(*0*) Pre-processing The most interesting pre-processing strategy applied this year is that of alphabetically ordering the words within the n-grams that compose the document's fingerprint (Gottron, 2010; Kasprzak and Brandejs, 2010; Rodríguez Torrejón and Martín Ramos, 2010). This operation aims at "defuscating" the operations performed during the paraphrasing simulation process on the corpus. Another way of defuscating is substituting the vocabulary of d by its synonyms, by means of Wordnet, as Alzahrani and Salim (2010) did (cf. Section 5.1.2.1).

A different pre-processing is that of document's splitting for retrieval (Corezola

Table 7.4: Pre-processing, heuristic retrieval, detailed analysis, and post-processing generalisations at the 2nd International Competition on Plagiarism Detection. Notation in Table 7.2.

Participant	Step (*0*) Pre-processing								Step (*1*) Heuristic Retrieval		Step (*2*) Detailed Analysis				Step (*3*) Post-processing		
															discard S if		merge S_1, S_2 if
	case folding	sw removal	!anum removal	stemming	doc. splitting	n-grams ordering	syn. normalisation	lang. normalisation	word n-grams	char. n-grams	word n-grams	char. n-grams	dot-plot	greedy str. tiling	$\lvert s_q \rvert < thres_1$	$sim(s_q, s) < thres_2$	$\delta(s_1, s_2) < thres_3$
Kasprzak						■			5		5						■
Zou	■			■				■	5				■			■	
Muhr					■			■	1		3				■	■	■
Grozea										16		16	■				
Oberreuter		■	■		■				3		3						
Rodriguez	■	■		■		■			3		3				■		
Corezola		■		■	■			■	1		1						■
Palkovskii									5		5						
Sobha									4		4						
Gottron					■	■		■	1		5		■		■		■
Micol		■	■						1			30			■		■
Costa-jussà	■	■		■	■				1				■				■
Nawab	■		■					■	5					■			■
Gupta									9		7						■
Vania		■						■	1		6					■	
Alzahrani		■		■			■		3		1						■

Pereira *et al.*, 2010b; Gottron, 2010; Muhr *et al.*, 2010; Oberreuter, L'Huillier, Ríos, and Velásquez, 2010; R. Costa-jussà *et al.*, 2010). The reason behind this heuristic may be in the same nature of the PAN-PC series; as the topics of $s_q \in d_q$ —a plagiarised fragment— and d_q —the suspicious document s_q is inserted in— are in general unrelated, the use of standard document level strategies may be skewed.

Aiming at detecting cross-language cases, language normalisation was carried out as well (cf. Section 6.2.2.5); i.e., non-English documents were translated into English during pre-processing. The process is as follows: (a) the most likely language L^* of the text in d is detected, (b) if $L^* \neq$ English, d's contents are mapped into this language. Some participants used "traditional" machine translation (Corezola Pereira *et al.*, 2010b; Gottron, 2010; Nawab *et al.*, 2010; Vania and Adriani, 2010). Some others ran built-in translation (mapping) models, considering multiple translations per word (Muhr *et al.*, 2010) (cf. Section 6.2.2.5).

Some participants did not report any pre-processing (Grozea and Popescu, 2010a; Gupta *et al.*, 2010; Kasprzak and Brandejs, 2010; Lalitha Devi, Rao, Sundar Ram, and Akilandeswari, 2010; Palkovskii *et al.*, 2010).[11]

[11]There are participants that used standard information retrieval engines, such as Lucene (`http://lucene.apache.org`) or Indri (`http://www.lemurproject.org/`). These engines include dif-

(*1*) Heuristic retrieval During this step, most of the participants performed a comparison between d_q and $d \in D$ on the basis of word n-grams (with $n = \{1, 3, 4, 5\}$) or character 16-grams. As aforementioned, some of them sorted these n-grams alphabetically. In order to compose their queries from d_q, Gupta *et al.* (2010) considered only those non-overlapping word 9-grams with at least one named entity within them. The most related documents D^* are retrieved for further comparison. Zou *et al.* (2010) used Winnowing (cf. Section 5.1.2.1).

(*2*) Detailed analysis Various participants considered sorted n-grams (Gottron, 2010; Kasprzak and Brandejs, 2010; Rodríguez Torrejón and Martín Ramos, 2010).[12] Others, such as Corezola Pereira *et al.* (2010b), apply a machine learning approach considering different features: bag-of-words cosine similarity, the score assigned by the IR system, and length deviation between s_q and s, among others. Dotplot-based strategies were used by many participants (Gottron, 2010; Grozea and Popescu, 2010a; R. Costa-jussà *et al.*, 2010; Zou *et al.*, 2010) and only one team applied GST (Nawab *et al.*, 2010). Alzahrani and Salim (2010) are the only team that, on the basis of WordNet synsets, semantically expanded the documents' vocabulary.

(*3*) Post-processing At this final step, two different heuristics are applied: (*i*) discarding plagiarism candidates shorter than a given threshold or not similar enough to be considered relevant and (*ii*) merging detected discontinuous fragments that are particularly close to each other. Probably the most interesting operation is merging. The maximum merging threshold is 5,000 characters (R. Costa-jussà *et al.*, 2010).

7.2.1.2 Intrinsic Detection

This approach received even less attention this year. Probably the main reason was that, aiming at composing a more realistic challenge, the intrinsic and external corpora were mixed together. Research teams that had models for intrinsic analysis only, had to face a huge corpus where the source of most plagiarism cases was included (i.e., they were intended to be detected with external approaches). These factors discouraged participation. As the parameters used by the two participant teams that approached intrinsic analysis are different in nature, no summary of them is included.

(*0*) Pre-processing The only approach that reported performing pre-processing was that of Muhr *et al.* (2010). During this step, they extracted: (*a*) stopwords and (*b*) stem-suffixes, i.e., the suffix that a stemmer would remove from a word to obtain its stem.

(*1*) Chunking Muhr *et al.* (2010) divided d_q into coherent blocks of multiple sentences.[13] Suárez, González, and Villena-Román (2010) divided d_q into paragraphs.

(*2*) and (*3*) Retrieval and outlier detection The features used by Muhr *et al.* (2010) are closed-class words (cf. Section 3.4.4); in particular stopwords, and word suffixes are considered to try to characterise an author's style. The feature vectors of Suárez

ferent pre-processing modules, but it is not reported whether they have been applied.

[12] Rodríguez Torrejón and Martín Ramos (2010) call them *contextual n-grams*.

[13] This strategy could be influenced by the nature of the corpus as s_q is in general unrelated to d_q.

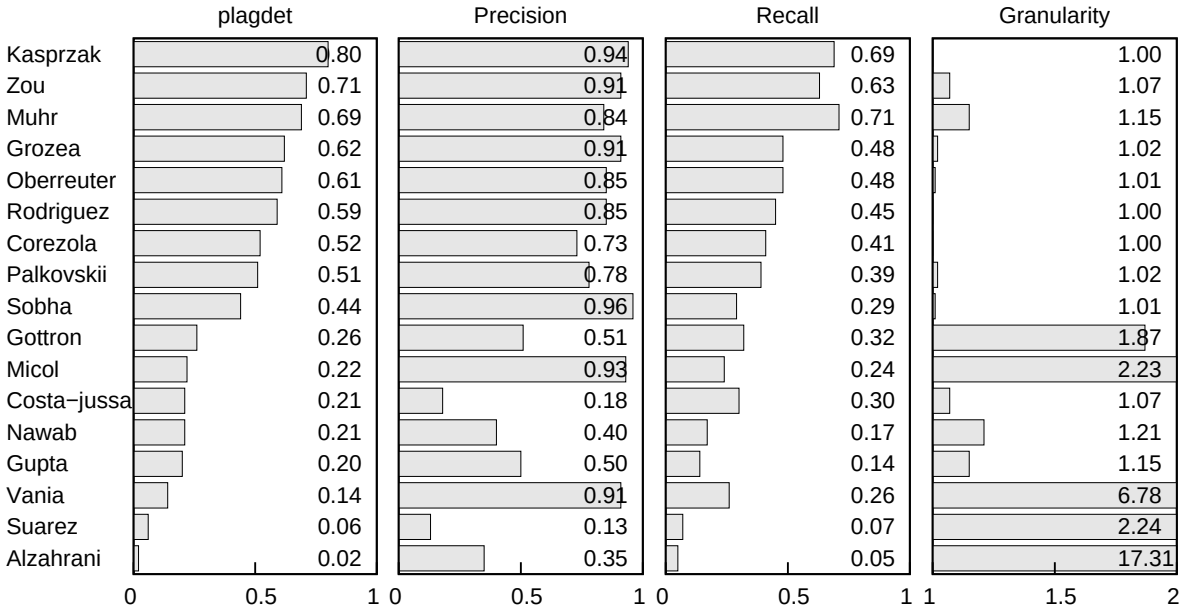


Figure 7.4: Overall results of automatic plagiarism detection at the Second International Competition on Plagiarism Detection.

et al. (2010) seem to be composed of d_q's vocabulary (this information is not fully provided in their report). As in 2009, the analysis in both approaches is made considering chunks vs. documents comparison. Muhr *et al.* (2010) computed the cosine similarity measure, whereas Suárez *et al.* (2010) opted for the LempelZiv distance (Ziv and Lempel, 1977).

7.2.2 Results and Discussion

The corpus used in this edition of the competition was the PAN-PC-10 (cf. Section 4.2.3.4). The overall results are summarised in Fig. 7.4.[14] Disregarding granularity, the performance of the top three approaches is very similar. The models of Kasprzak and Brandejs (2010) and Muhr *et al.* (2010) were based on word n-grams, but the former one included tokens ordering. Zou *et al.* (2010) applied a combination of Winnowing, clustering, and dot-plot. Granularity shows to be an important factor for the final rank. The next three positions correspond to the second block of systems, which reach a similar recall (around 0.47).

7.2.2.1 External Detection

The results when looking at the external cases are included in Fig. 7.5. As described in Section 4.2.3.4, the biggest difference between the PAN-PC-09 and PAN-PC-10 is the inclusion of manually simulated cases of re-use. Figures 7.6 and 7.7 include the evaluation

[14] In the rest of histograms, those participations with $rec \leq 0.1$ are not included. The complete figures can be consulted in Potthast *et al.* (2010d).

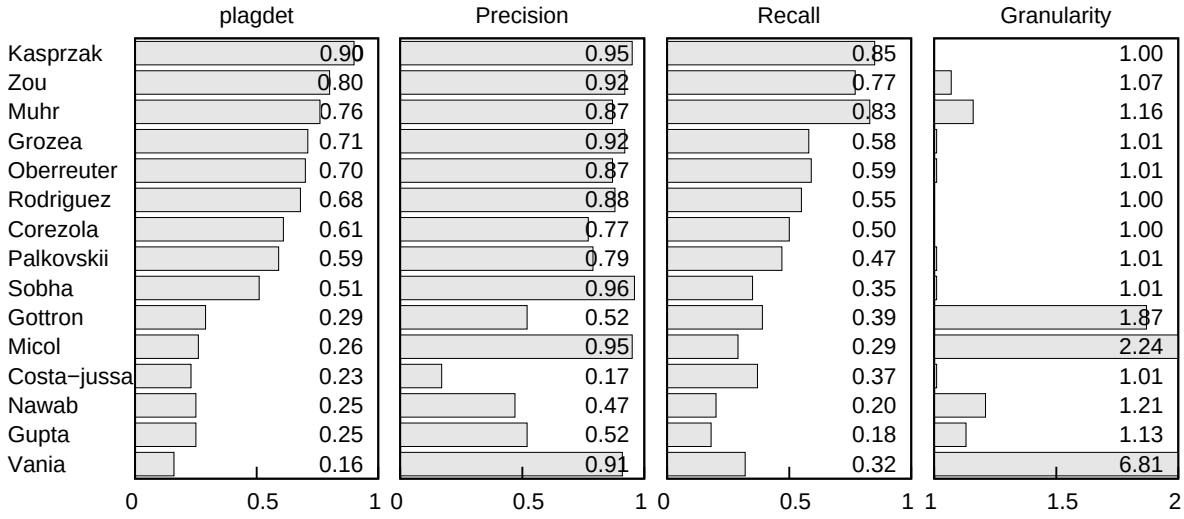


Figure 7.5: Overall results of *external* plagiarism detection at the Second International Competition on Plagiarism Detection.

results for different paraphrasing strategies, namely manual and automatically created. For comparison purposes, those cases of exact copy (none: verbatim) are included as well.

A dramatic decrease of both precision and recall is experienced by all the models (nearly one third) with the manually paraphrased text respect to verbatim copies (Figure 7.6). The reasons are twofold: (*i*) the manually created cases are among the shortest in the collection, and models have shown to face difficulties when dealing with short cases of plagiarism, and (*ii*) manually created cases were generated with the explicit instruction of strongly paraphrase the borrowed fragments. As a result, these cases are the hardest to detect and remain an open issue in plagiarism detection (Stein *et al.*, 2011a). The plagdet values of Grozea and Popescu (2010a) and Nawab *et al.* (2010) are slightly higher than 0.25. The models behind their approaches, GST and dot-plot, show to be able to detect many cases of text re-use, even after high levels of modification. The impact is not so high when looking at automatically paraphrased cases with low and high obfuscation levels (Figure 7.7). A deeper analysis of the manually generated cases and why the different detectors manage (or not) to detect them is provided in Chapter 8.

Figure 7.8 contains the evaluation of cross-language plagiarism detection. The best system obtained a $plagdet = 0.80$. However, we have to consider that the language normalisation of many participants, including Kasprzak and Brandejs (2010), is indeed exactly the same of the corpus generation, i.e., translating a text fragment $s \in L$ (German or Spanish) into $s_q \in L'$ (English) with Google Translator. Therefore, it is very likely that the resulting text will be exactly s_q. As different models for monolingual detection are very good in detecting exact copies, they do not have big difficulties with these cases. The results obtained by Muhr *et al.* (2010) for translated text are remarkable. On the basis of a cross-language mapping that does not depend on Google, but a built-in dictionary, they obtained $rec = 0.52$, with a very competitive $prec = 0.77$. However, they still have to deal with a high granularity. Note that nearly ten participants obtained

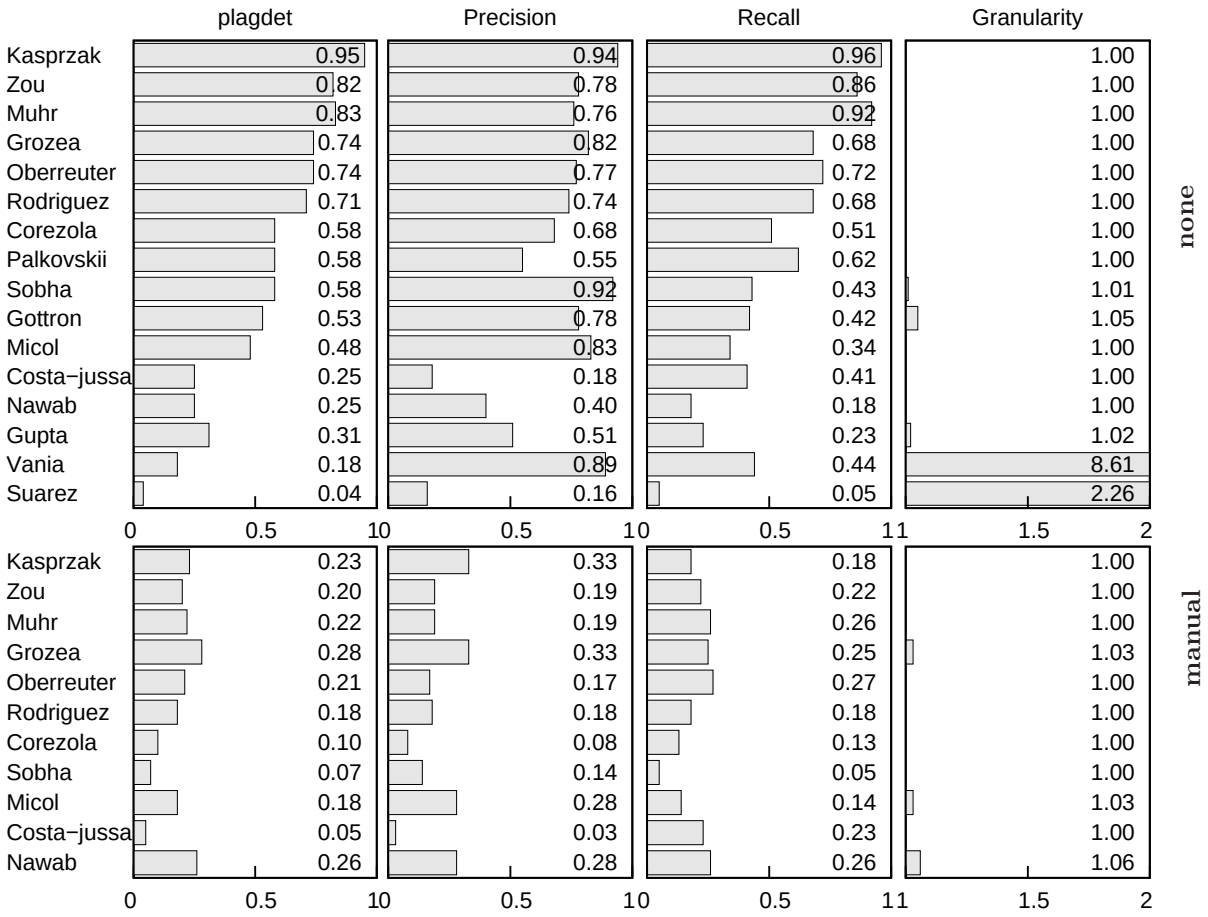


Figure 7.6: Results of external detection for *paraphrase* plagiarism at the Second International Competition on Plagiarism Detection (1 of 2). Kind of paraphrase on the right hand side.

a recall lower than 0.1 with these cases.

Figures 7.9 and 7.10 display the results when considering different lengths of suspicious documents and lengths of borrowed fragments. Detecting re-use in longer documents seems to be easier (the same could be said for long borrowings). However, as pointed out by Potthast *et al.* (2010d), the level of automatic obfuscation applied when producing the corpus was lower for longer documents and cases. This behaviour aimed at emulating the plagiarist that modifies short, but prefers cut & pasting longer fragments. As a result, whether detecting text re-use on longer or shorter cases (and documents) implies an easier or harder task cannot be evaluated with certainty.

Finally, the second big difference of PAN-PC-10 with respect to PAN-PC-09 is in terms of cases generation: whether the contexts of the plagiarised and source documents (i.e., the corresponding d_q and d), were on related topics. Figure 7.11 compares the results obtained in both cases. No important difference can be observed respect to intra-topic (d_q and d are on similar topics) and inter-topic (d_q and d are on random topics). The main reason is that at heuristic retrieval most of the models use a full representation

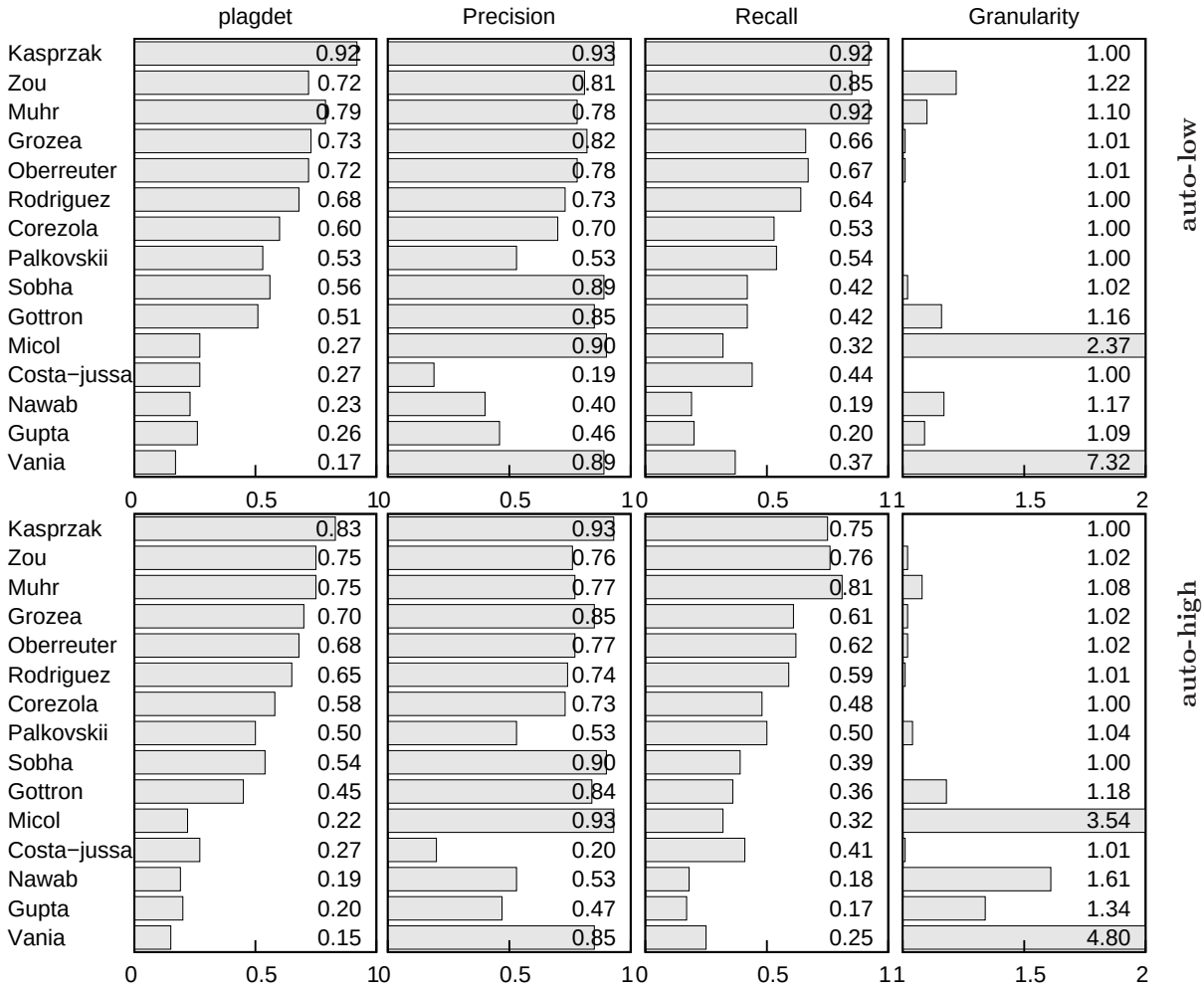


Figure 7.7: Results of external detection for *paraphrase* plagiarism at the Second International Competition on Plagiarism Detection (2 of 2). Kind of paraphrase on the right hand side.

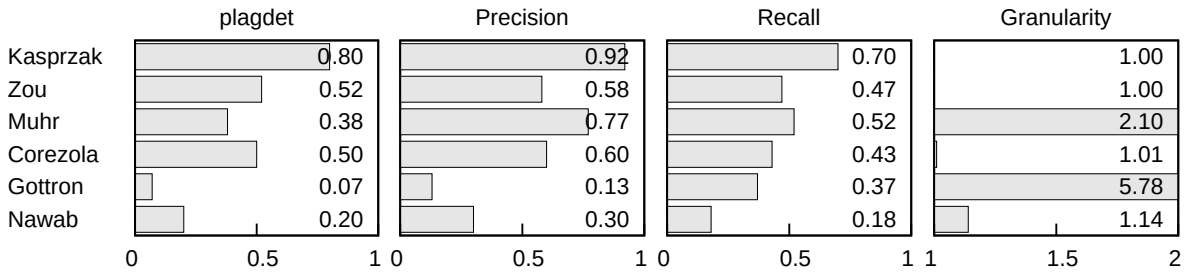


Figure 7.8: Results of external detection for *translated* plagiarism at the Second International Competition on Plagiarism Detection.

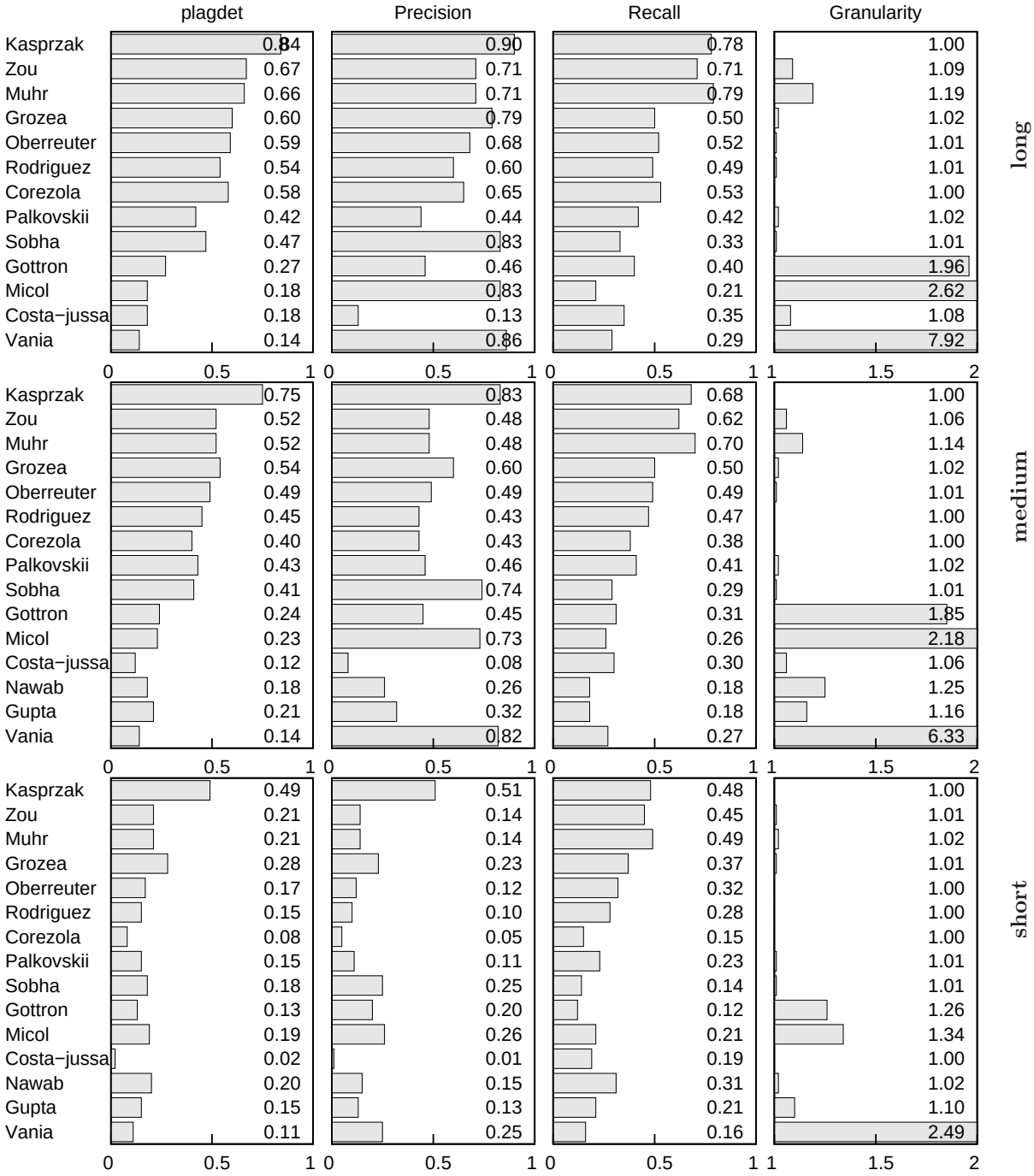


Figure 7.9: Results of external detection for documents with different lengths at the Second International Competition on Plagiarism Detection. Document's length on the right hand side.

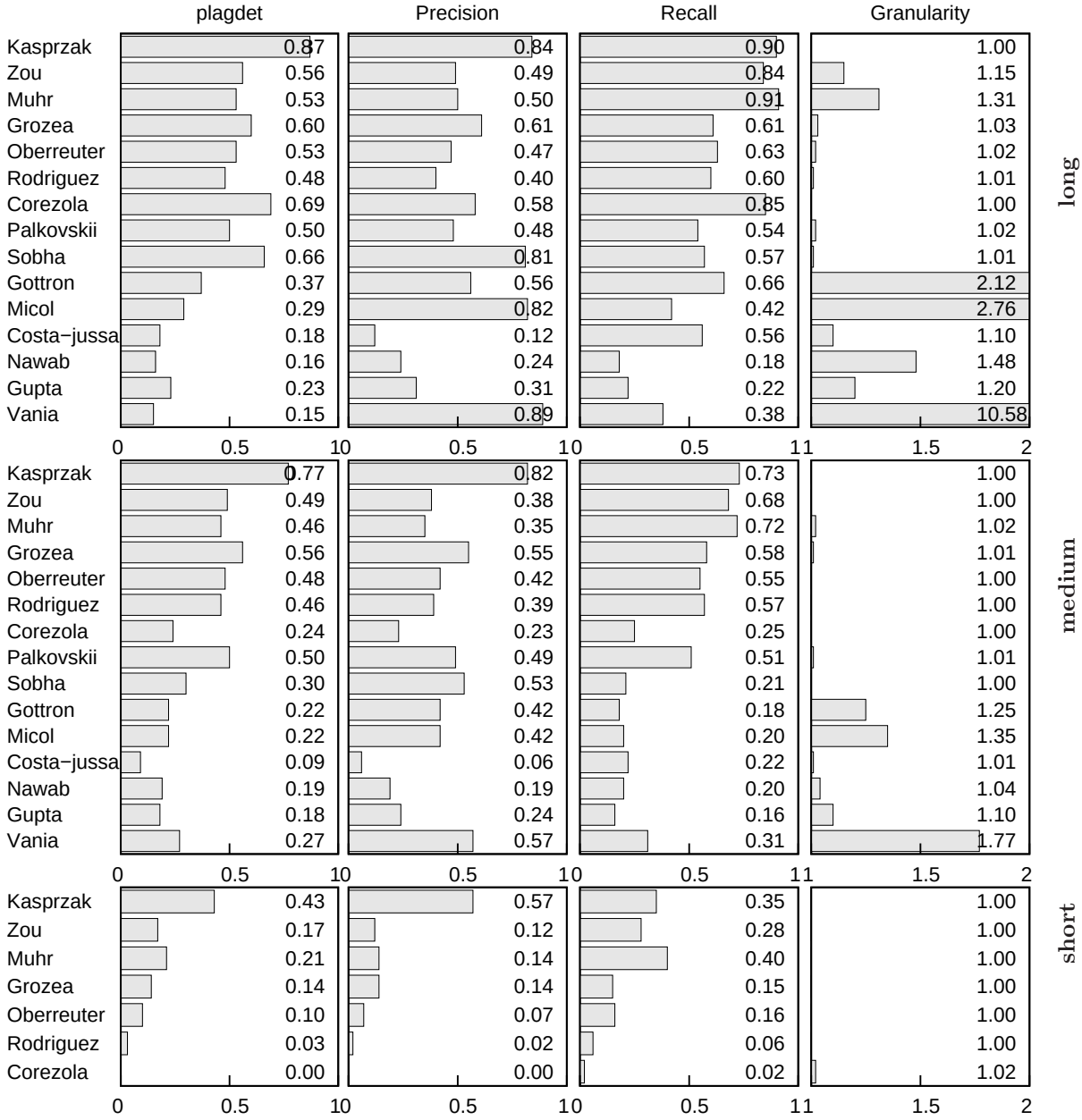


Figure 7.10: Results of external detection for case lengths at the Second International Competition on Plagiarism Detection. Cases' length on the right hand side.

of documents and, therefore, the topic similarity between the documents in D_q and D is not relevant.

7.2.2.2 Intrinsic Detection

Regarding intrinsic detection, the overall evaluation is included in Fig. 7.12. As aforementioned, only two participants applied this approach. When comparing the outcome of Muhr *et al.* (2010) to that of 2009 (by the same group, there leaded by Zechner *et al.*

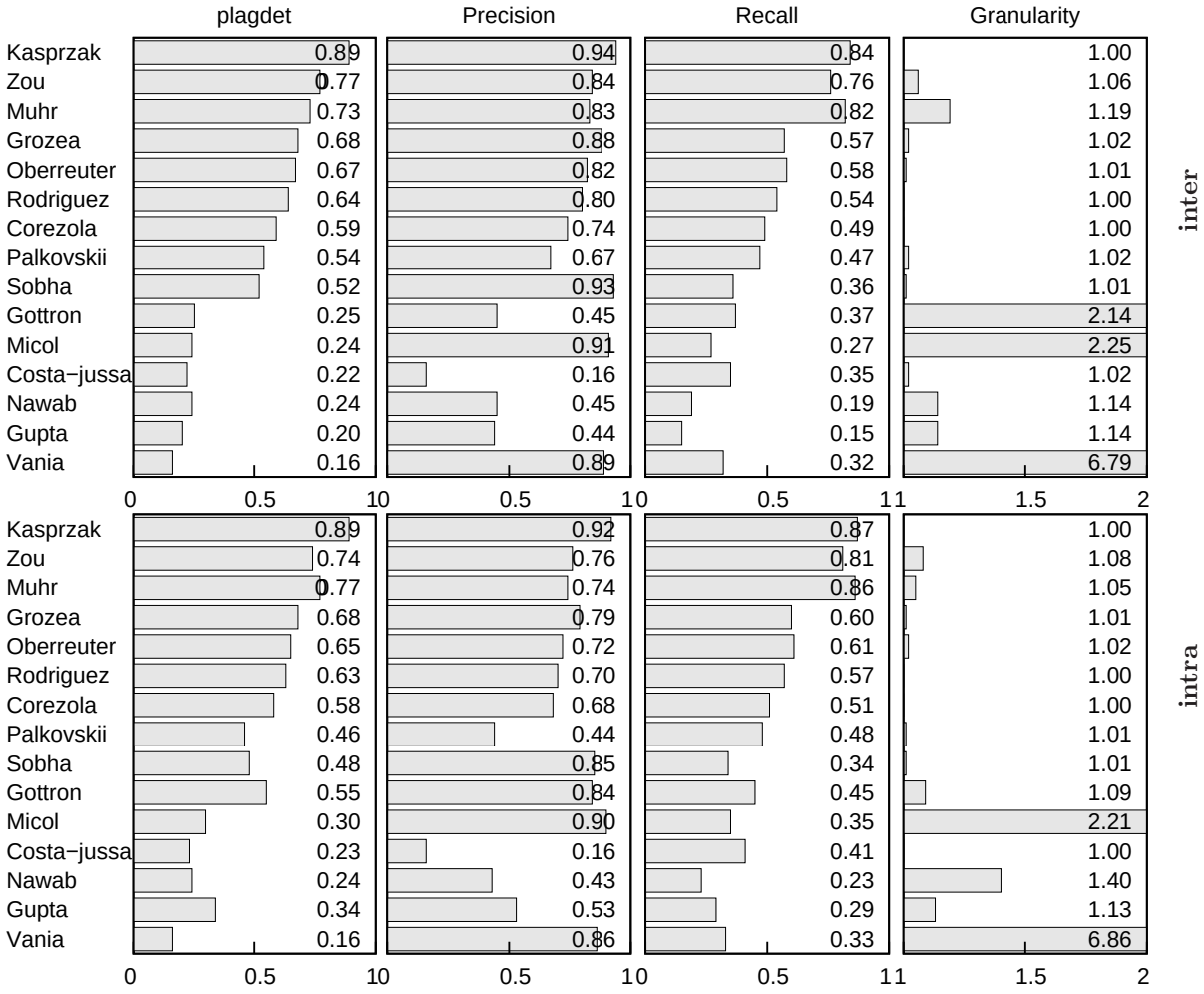


Figure 7.11: Results of external detection for inter- and intra-document at at the Second International Competition on Plagiarism Detection.

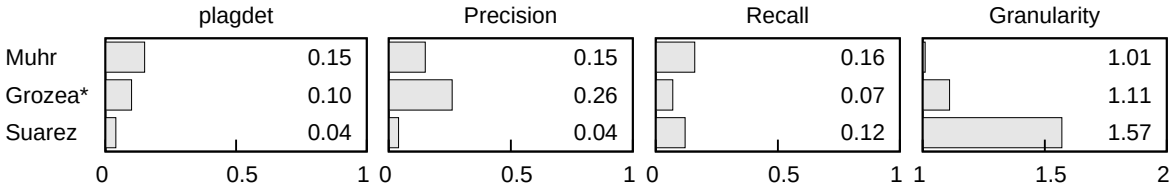


Figure 7.12: Overall results of *intrinsic* plagiarism detection at the Second International Competition on Plagiarism Detection. Note that Grozea and Popescu (2010a) did not perform an intrinsic analysis.

(2009)), a slight decrease in the quality can be observed. This is caused, in part, by the nature of this year's corpus. Intrinsic detection is still far to be (partially) solved.

On the one side, a few approaches for external detection "found" some fragments, aimed at being intrinsically detected, as plagiarised (Corezola Pereira *et al.*, 2010b; Grozea and Popescu, 2010a; Nawab *et al.*, 2010; Oberreuter *et al.*, 2010; R. Costa-jussà *et al.*, 2010). For instance, the approach of Grozea and Popescu (2010a) obtained $rec = 0.07$, with $prec = 0.26$ (a precision even higher than the best performing intrinsic approach (Muhr *et al.*, 2010)). However, as they detected fragments s_q together with their claimed source s (which was not in the corpus D), these detections cannot be considered as correct. On the other side, the intrinsic approach of Suárez *et al.* (2010) obtained $rec = 0.06$ in the subset of cases for which source texts were included in D. We consider that these few cases can be considered as properly detected (regardless their source was not identified).

7.3 PAN @ CLEF 2011

This is the last competition run up to date. A total of eleven plagiarism detectors took part in this edition. Interestingly, the winners of this year's competition (as in 2010), are a team that run a plagiarism detection system "in real life". That is, they have a real product, in service in their institutions.[15] Therefore, the PAN competition on plagiarism detection not only has fostered the research in the plagiarism detection topic, but has motivated the development of systems that are used in real academic (and other) scenarios.[16]

In 2010, the results obtained by the participants approaching both external and intrinsic analysis were importantly downgraded when aiming at detecting those cases for which source was not provided (Muhr *et al.*, 2010). Intrinsic analysis had been discouraged. In 2011 the corpus was divided in a similar fashion to that of the 2009 edition (i.e., specific partitions for intrinsic and external analysis, cf. Section 4.2.3.5). As a result, intrinsic analysis attracted more participants this year.

7.3.1 Tasks Overview

7.3.1.1 External Detection

The corpus generated this year has been the most challenging of all. Only 2% of the cases implied no paraphrasing at all. The rest included both automatic or manual obfuscation. Indeed, cases of translated plagiarism were manually obfuscated as well (cf. Table 4.12,

[15]The system of the winner participants, Oberreuter *et al.* (2011), is being used in Chile: `http://www.docode.cl/`

[16]For instance, Kasprzak and Brandejs (2010) work on the analysis of the Czech National Archive of Graduate Theses (`http://theses.cz/`), Grman and Ravas (2011) work on the plagiarism detection system for dissertations in 33 universities in the Slovak Republic, and Palkovskii, Belov, and Muzika (2011) offer a commercial plagiarism detection system (`http://plagiarism-detector.com`).

Table 7.5: Pre-processing, heuristic retrieval, detailed analysis, and post-processing generalisations at the 3rd International Competition on Plagiarism Detection. Notation in Table 7.2.

Participant	Step (*0*) Pre-processing								Step (*1*) Heuristic Retrieval		Step (*2*) Detailed Analysis				Step (*3*) Post-processing			
															discard S if		merge S_1, S_2 if	
	case folding	sw removal	!anum removal	stemming	doc. splitting	n-grams ordering	syn. normalisation	lang. normalisation	word n-grams	char. n-grams	word n-grams	char. n-grams	dot-plot	greedy str. tiling	$\lvert s_q\rvert < thres_1$	$sim(s_q, s) < thres_2$	$\delta(S_1, S_2) < thres_3$	
Grman				■			■	■	1		1				■	■	■	
Grozea						■		■		16		16	■					
Oberreuter		□							4		3							
Rodriguez	■	■		■		■		■	3		3						■	
Rao								■	1		7					■	■	
Palkovskii				■			■											
Nawab	■		■		■									■				
Ghosh		■		■	■		■											

page 96). A summary of the parameters used by the participants at the different external detection steps is described in Table 7.5.

(*0*) Pre-processing The pre-processing operations are again very standard this year. Grozea and Popescu (2011) joined Rodríguez Torrejón and Martín Ramos (2011) on the strategy of sorting the tokens within each considered n-gram. Additionally, more teams decided to use thesauri in order to consider terms' synonyms, hypernyms and other semantically related words through Wordnet (Ghosh, Bhaskar, Pal, and Bandyopadhyay, 2011b; Grman and Ravas, 2011; Palkovskii *et al.*, 2011). Oberreuter *et al.* (2011) decided to eliminate stopwords for the retrieval step, but take them back for the detailed analysis.

Respect to language normalisation, this year the novelty can be found in the approach of Rodríguez Torrejón and Martín Ramos (2011). Whereas their monolingual strategy was as in the 2010 competition (Rodríguez Torrejón and Martín Ramos, 2010), they proposed a new cross-language strategy, without appealing any translation service to translate the non-English documents. Instead, they used a dictionary extracted both from the Wiktionary and Wikipedia langlinks. Rather than a complete "traditional" translation process, they performed a one-to-one mapping of words from L into L'. In order to reduce the amount of potential noise inserted by this process, the mapping strategy is divided in two steps: (i) a term $t \in d$ is substituted by the most likely stem translation into L', and (ii) if no entry exists for t in the dictionary, a dictionary of stems is used, following a similar strategy. As they used ordered n-grams, no syntactically correct translations were necessary.

(*1*) Heuristic retrieval The heuristic retrieval strategies were again, in general, standard. For instance, Rao, Gupta, Singhal, and Majumdar (2010) retrieved the most similar documents $d \in D$ respect to d_q on the basis of the cosine measure. Nawab,

Stevenson, and Clough (2011) used Terrier as IR system and composed their queries of d_q's sentences. The ten documents to be retrieved were decided by considering the similarities to the top-retrieved documents $d \in D$. The strategy of Oberreuter *et al.* (2011) was slightly different, and based on the scarcity of common n-grams in two independent documents, for high levels of n (cf. Fig. 2.2). If d_q and d share at least two 4-grams g_1, g_2, such that $\delta(g_1, g_2) < thres$ (i.e., they are close enough to potentially belong to the same paragraph), d is retrieved.

(*2*) Detailed analysis In the approach of Grman and Ravas (2011), d_q and d were divided into non-overlapping fragments, which were compared on the basis of Eq. (3.7) of page 64; i.e., they simply consider the intersection of the vocabularies in s_q and s. This simple approach confirms what we have signalled before: in text re-use detection, the frequency of the terms is not so important, but their simple occurrence is.

Ghosh *et al.* (2011b) performed a sentence-wise comparison in which the final similarity assessment is indeed a combination: the similarity between s_q and s is computed as $sim(s_q, s) - dissim(s_s, s)$; i.e., both the similarity and dissimilarity between the text fragments is considered. Rao *et al.* (2010), in turn, look for common matches of word 7-grams between d_q and d. If one is found, windows of length 25 are extended both in d_q and d iteratively while a minimum similarity threshold is maintained. Nawab *et al.* (2011) use GST again, as in the previous competition (Nawab *et al.*, 2010).

(*3*) Post-processing The post-processing operations are as well as in the previous edition. Grman and Ravas (2011) reported discarding candidates which are not similar enough or too short. Both Grman and Ravas (2011) and Rao *et al.* (2010) merged two fragments if they consider them to appear close enough within the text.

7.3.1.2 Intrinsic Detection

A summary of the parameters applied is included in Table 7.6. The only participants that report performing pre-processing operations are Oberreuter *et al.* (2011). All of the approaches follow a similar strategy where the chunks in d_q are characterised and compared to the global document representation. Only Kestemont *et al.* (2011) opts for a section-wise comparison.

(*0*) Pre-processing Two of the teams opted for characterising d_q over the BoW model (Akiva, 2011; Oberreuter *et al.*, 2011). The former one reports applying case folding and removal of non-alphanumeric characters.

(*1*) Chunking All of the participants compose chunks of fixed length, either considering the number of characters (Akiva, 2011; Kestemont *et al.*, 2011; Rao *et al.*, 2010) or the number of tokens (Oberreuter *et al.*, 2011).

(*2*) and (*3*) Retrieval and outlier detection The core idea of Oberreuter *et al.* (2011) is that "if some of the words used on the document are author-specific, one can think that those words could be concentrated on the ... [fragments] ... that the mentioned author wrote." The model considered is extremely similar to that proposed by Stamatatos (2009b), with slight differences: (*i*) word 1-grams (i.e., BoW) are used

Table 7.6: Pre-processing, chunking, and outlier detection for intrinsic analysis at the 3rd International Competition on Plagiarism Detection. Lengths of windows and steps in terms of tokens t or characters c. Notation in Table 7.2.

Participant	Pre-processing				Chunking		Outlier detection	
	case folding	!αnum removal	word n-grams	char n-grams	$\|chunk\|$	$\|step\|$	chunk vs. doc.	chunk vs. chunk
Oberreuter	■	■	1		400 t		■	
Kestemont				3	5,000 c	2,500 c		■
Akiva			1		1,000 c		■	
Rao			multiple		2,000 c	200 c	■	

instead of character 3-grams, (*ii*) the considered frequency is absolute rather than normalised, and (*iii*) a different similarity model is used, which aims at determining the deviation respect to the entire d_q's fingerprint.

Kestemont *et al.* (2011) designed a number of modifications to the model of Stamatatos (2009b), justified by the fact that "stylometric comparison of two samples of so different size (the single chunk vs. the entire document) is hard to justify from a theoretical perspective". As a result, the chunks' profiles are not compared to that of the entire document d_q. Instead, a $k \times k$ covariance matrix is computed for the chunks $c_1, c_2, \ldots, c_k$. Moreover, the features considered to compute the dissimilarity, on the basis of Eq. (5.1) of page 116, are pre-filtered. Only those n-grams belonging to a pre-defined set of "high frequency n-grams" over the entire collection D_q are used. As a result, nd_1 becomes symmetric. The outlier detection is then carried out over the covariance matrix, where each row describes one chunk.

Akiva (2011) used a completely different approach. First of all, the chunks fingerprints are composed of a fixed-length binary vector. Such a vector contains the 100 rarest words appearing in at least 5% of the chunks. Once the vectors are composed, a clustering process is carried out, assuming that one of the resulting two clusters will contain the original chunks, and the other those plagiarised. Rao *et al.* (2010) included discourse markers additionally to the features considered by Muhr *et al.* (2010); Stamatatos (2009b); and Zechner *et al.* (2009).

7.3.2 Results and Discussion

7.3.2.1 External Detection

It is evident that the cases of exact copy are particularly scarce this year. Figure 7.13 reflects this fact. The success in terms of recall has an important decrease respect to last year: around 0.66 for the best performing approaches in 2010 versus less than 0.40 this time. The most successful approach is that of Grman and Ravas (2011), which obtained $plagdet = 0.56$. Their precision, as well as that of Oberreuter *et al.* (2011) is higher than

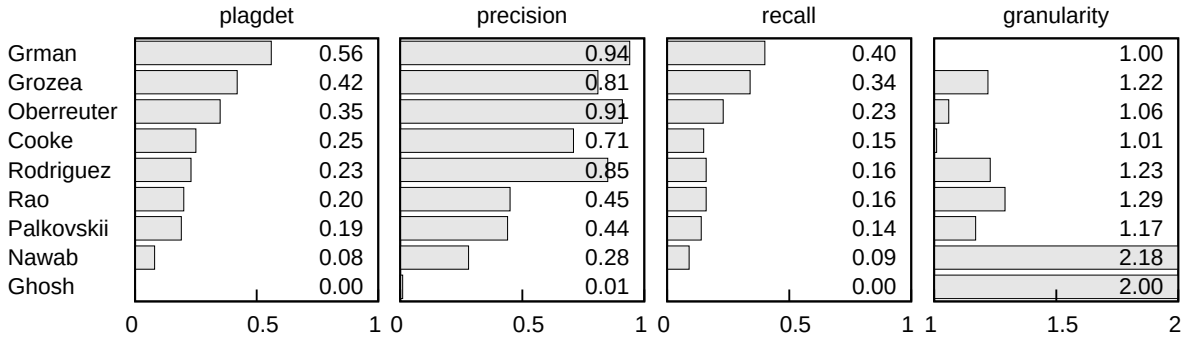


Figure 7.13: Overall results of *external* detection at the Third International Competition on Plagiarism Detection.

0.90. These values imply that both approaches miss some cases, but if they report a text fragment as re-used, it is certainly worth looking at it.

Figure 7.14 includes the results for different types of paraphrasing. As expected, when no paraphrase is applied at all, i.e., cut & paste copy, detectors are close to perfection (with both precision and recall around 0.90 and even 0.97 for the best one). The problem of detecting verbatim copies seems to be nearly solved. However, when cases are generated by manual paraphrasing, once again, the quality decreases to nearly one third. Indeed, the performance is still lower respect to the cases with a high level of automatic obfuscation. This is definitively an open issue in text re-use detection as it is indeed closer to plagiarism of ideas (where the source and plagiarised text are not very related, whereas the ideas they express are. As most of state-of-the-art systems are principally based on syntactic features (in part because of the difficulty of considering other kinds of knowledge), strong reformulation makes re-use extremely hard to be detected.

Figure 7.15 contains the figures for translated plagiarism: first for MT and second for MT plus manual obfuscation. As aforementioned, the use of automatic MT for generating and detecting cross-language plagiarism cases has caused a bias in the real estimation of the detectors achievements within this kind of borrowing. Probably the closer estimation to the expected quality when dealing with real cases of cross language plagiarism is that of Rodríguez Torrejón and Martín Ramos (2011). The reason is that they do not use a commercial translator, but a built-in mapping model. In the automatically translated cases of plagiarism this approach obtained $rec = 0.24$ with $prec = 0.69$, but these competitive values were not so high in the further obfuscated cases. We discuss further on this issue in Section 7.5.

The next two parameters have to do with both, documents' and plagiarism cases' length. The results regarding the former aspect are in Fig. 7.16, whereas those regarding the latter aspect are in Fig. 7.17. It is clear that the length of the document the plagiarism and its source are within is irrelevant for text re-use detection, at least when looking at recall and precision. However, some approaches seem to have problems respect to the granularity, which tends to be higher for bigger documents. The granularity issue becomes more relevant when looking at the length of the cases: in general, the length of the plagiarism cases is correlated to the level of granularity. Interestingly, medium-sized cases seem to be the most likely to be detected.

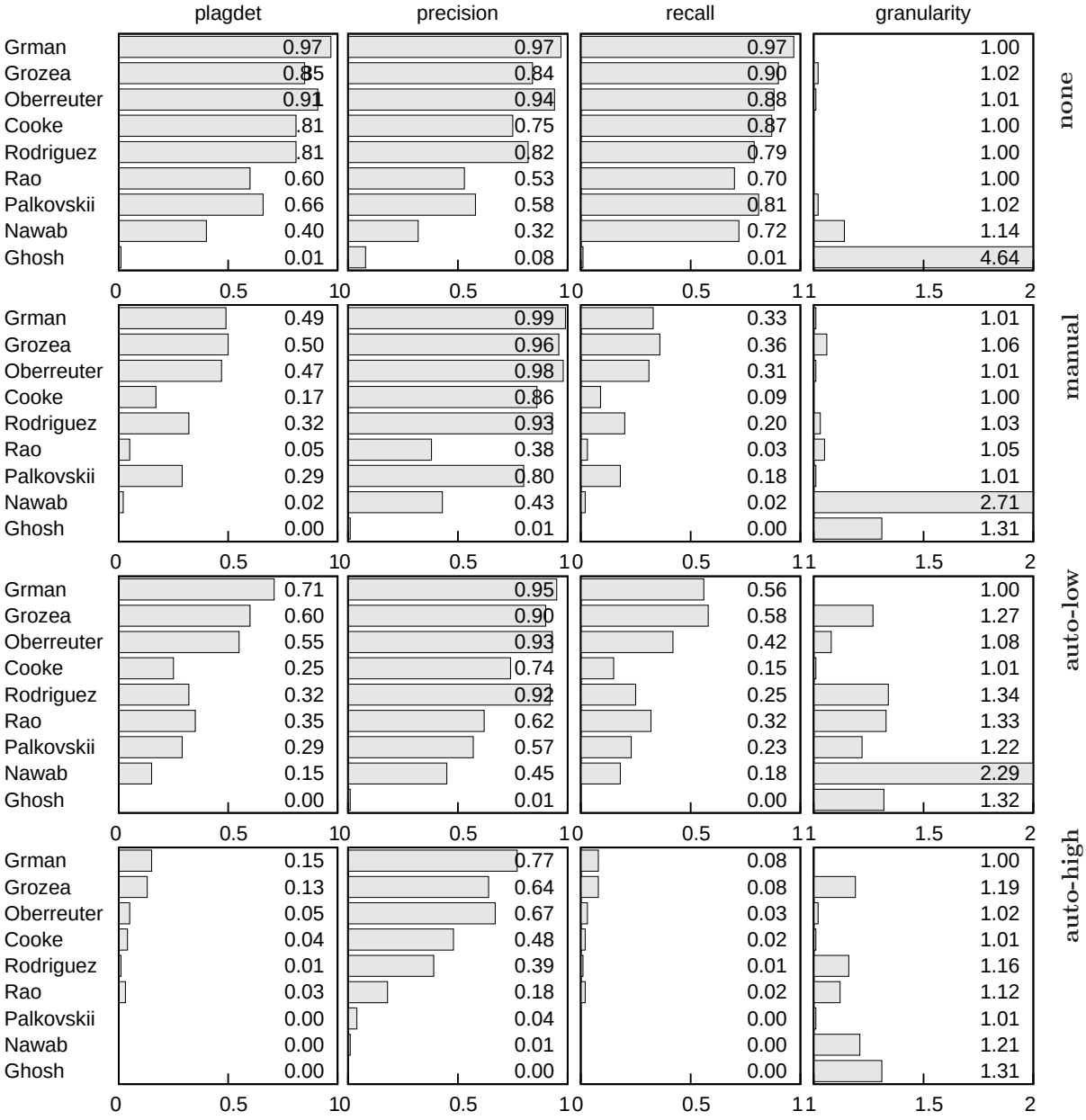


Figure 7.14: Results of external detection for *paraphrase* plagiarism at the Third International Competition on Plagiarism Detection. The kind of paraphrase is defined on the right hand side.

The last aspect is the amount of borrowed text per document. The results are included in Fig. 7.18. As with the size of the document, this factor seems not to be so relevant, as the obtained results are similar, with a drop between those documents that contain "medium" and "much" amounts of plagiarism.

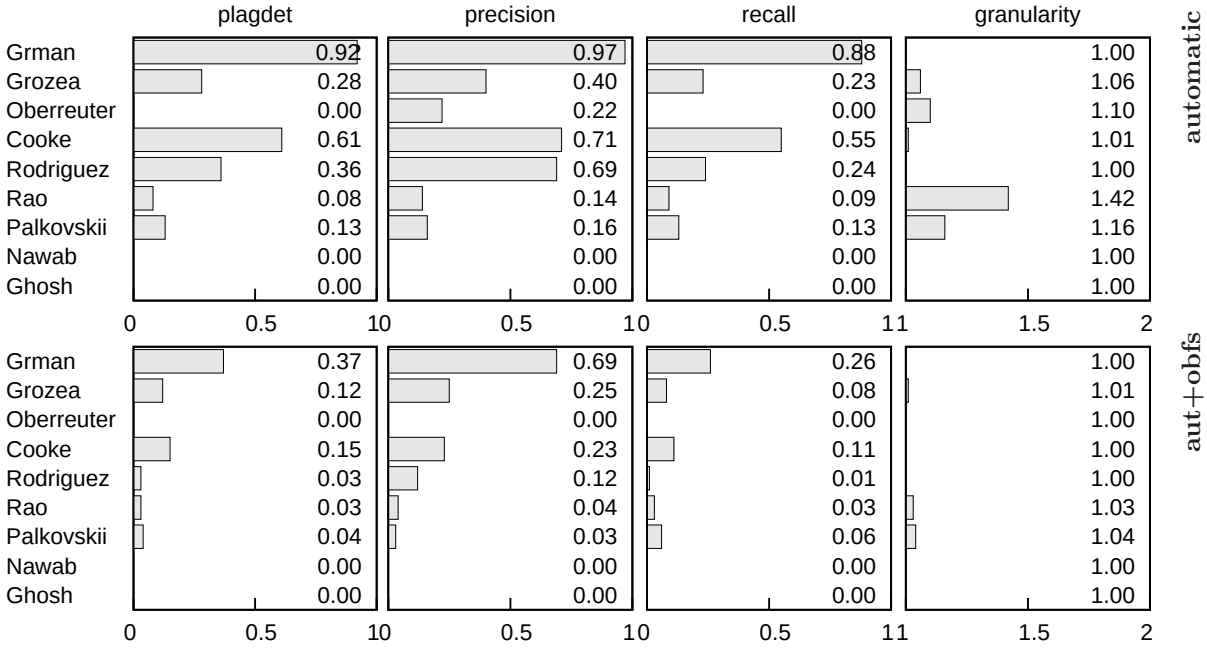


Figure 7.15: Results of external detection for *translated* plagiarism at the Third International Competition on Plagiarism Detection. The kind of translation is defined on the right hand side.

7.3.2.2 Intrinsic Detection

Regarding intrinsic analysis, the results are included in Fig. 7.19. The most successful approach, that of Oberreuter *et al.* (2011), has received some criticism due to the fact that the features it uses are tokens, which can be topic dependent and not very expressive in terms of complexity or style. All in all, it shows a good balance between precision and recall. The approach of Kestemont *et al.* (2011) seems promising as well. The comparison of chunks versus chunks rather that chunks versus document seems reasonable, also in those cases where a document contains re-use from different sources. Perhaps a combination of this comparison strategy with other features could offer better results.

7.3.2.3 Temporal Insights

Time is a relevant issue in external analysis. Though up to now the framework of PAN has not considered this aspect as an evaluation factor, some insights can be obtained from the same participants' reports. When looking at this factor, big differences can be found.

For instance, Grozea and Popescu (2011) report that their entire external process, running on high-performance hardware, took them around fifty hours (this without considering translation, which time they consider "prohibitive"). Cooke, Gillam, Wrobel, Cooke, and Al-Obaidli (2011) do not report how their model works, but claim that they need only 12 minutes, again on high-performance hardware. However, this time does not

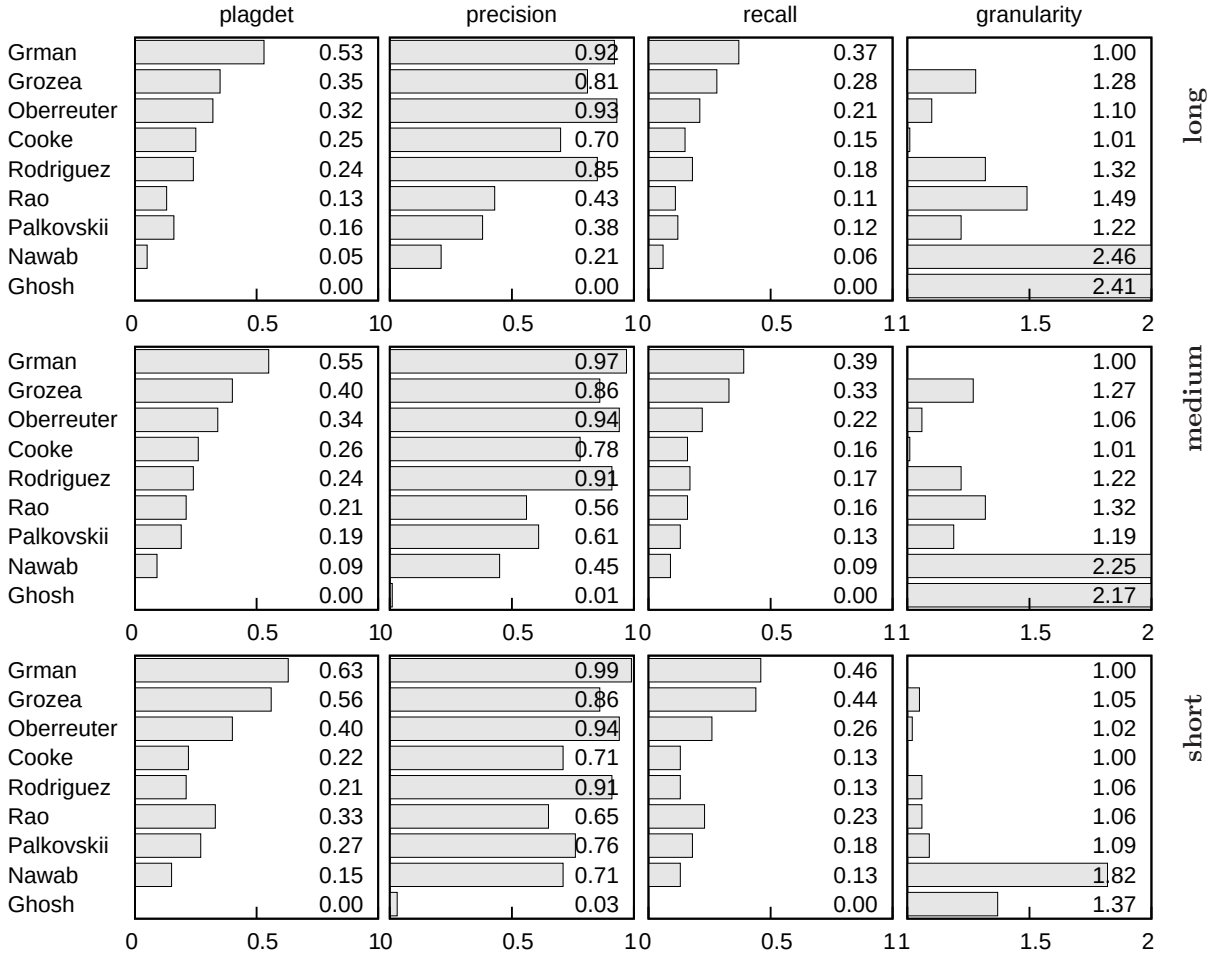


Figure 7.16: Results of external detection for document lengths at the Third International Competition on Plagiarism Detection. The length of the document is defined on the right hand side.

include system setup, language normalisation, and output generation; the overall process takes more than ten hours. Rodríguez Torrejón and Martín Ramos (2011) report processing the entire PAN-PC-11, including a built-in language normalisation process, in thirty minutes only.

7.4 Detection of Monolingual Plagiarism @ PAN

In this section we aim at determining the results obtained by means of a word n-gram text re-use detection model, as the one described in Section 5.3, on the PAN-PC-10 and PAN-PC-11. That is, d_q is split into small text fragments and compared against the entire d by considering word n-grams, with a Boolean similarity measure.[17]

[17]We would like to thank Diego Rodriguez Torrejón and José M. Martín Ramos for sharing their software implementation for running these experiments.

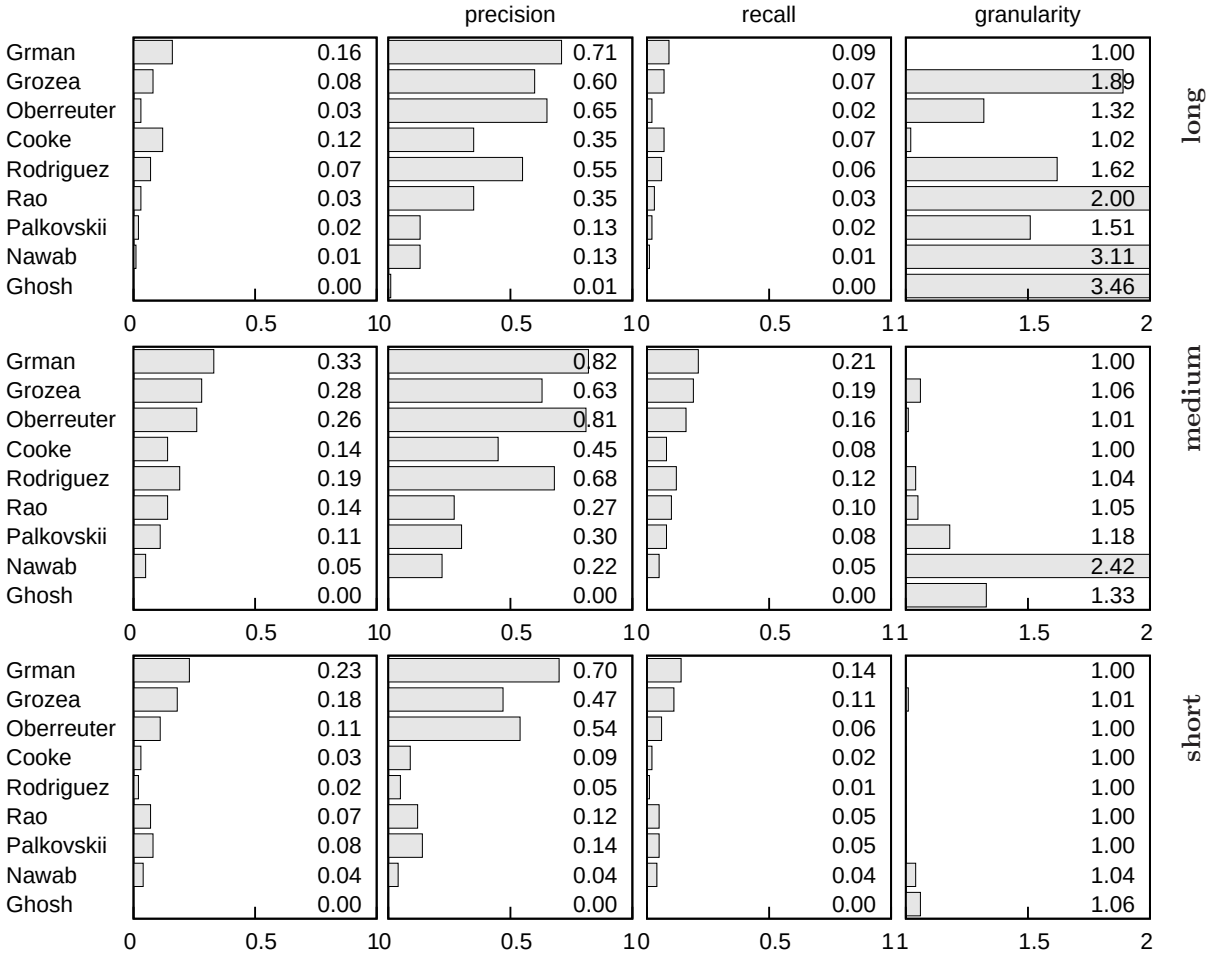


Figure 7.17: Results of external detection for case lengths at the Third International Competition on Plagiarism Detection. The length of the case is defined on the right hand side.

The retrieval strategy we had proposed in Section 5.4 cannot be used here. The reason is that in the PAN-PC series, a text fragment $s \in d$ is inserted into an arbitrary d_q. In general, the contents of d and d_q may be completely unrelated, causing the heuristic retrieval process, based on a reduced part of the vocabulary in d_q, to be useless. As a result, different modifications to the strategy are made:

1. The fragments s_q are not sentences, but chunks of a given length, measured as the amount of n-grams;
2. The similarity $sim(s_q, d)$ is computed on the basis of a containment-like measure; and
3. An additional strategy is included in order to delimit the specific plagiarised and source fragments in d_q and d.

The latter modification is necessary because here we are not interested in detecting a plagiarised fragment and its source document only. We want to identify the specific bor-

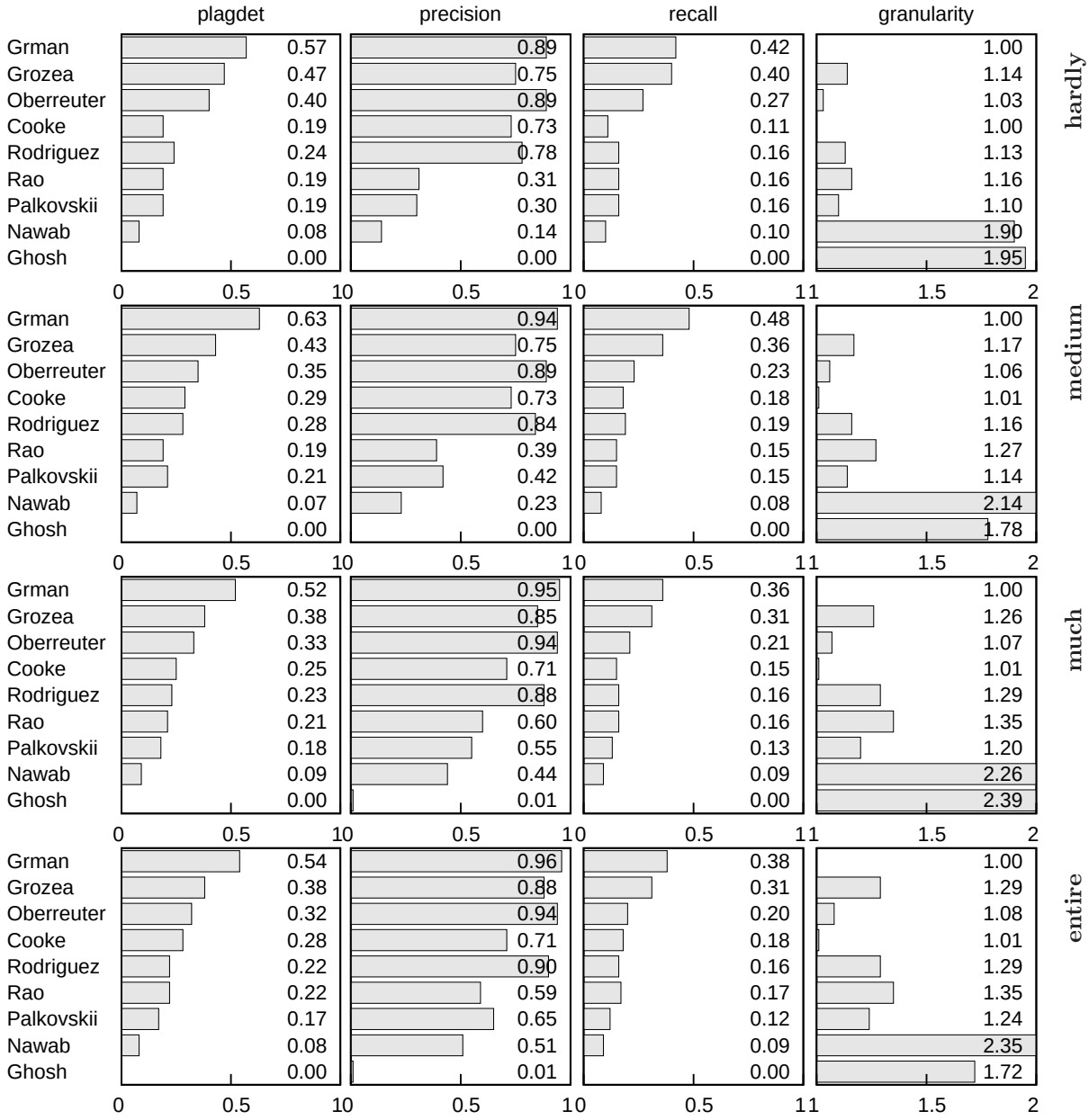


Figure 7.18: Results of external detection for documents different amounts of plagiarism per document at the Third International Competition on Plagiarism Detection. The amount of plagiarism per document is defined on the right hand side.

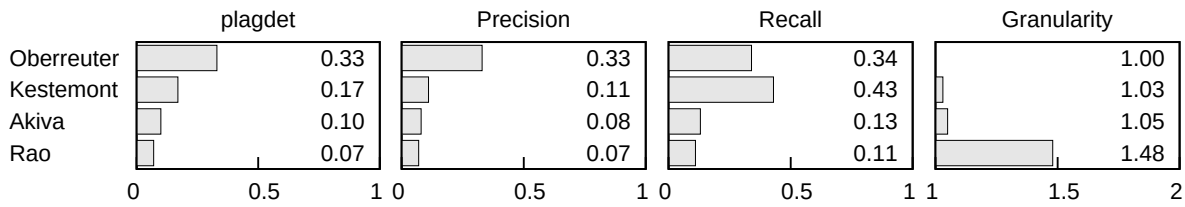


Figure 7.19: Overall results of *intrinsic* plagiarism detection at the Third International Competition on Plagiarism Detection.

Given D_q and D:

// **Pre-processing**
discard $d \in D$, $d_q \in D$ if d (d_q) $\notin$ English
for every $d \in D \cup D_q$:
    $case\_fold(d)$; $discard\_sw(d)$
    $discard\_single\_chars(d)$; $stem(d)$

// **Indexing**
$index(D, n)$

// **Heuristic retrieval**
for every $d_q \in D_q$:
    $S_q = split(d_q, k)$
    for each chunk s_q in S_q:
        $D^* \leftarrow \text{argmax}_{d \in D}\, sim(s_q, d)$

// **Detailed analysis**
    until no changes:
        if $source(s_{q,i}) == source(s_{q,i+1})$:
            $merge(s_{q,i}, s_{q,i+1})$
        if $length(s_q) > threshold$:
            $define\_offset\_length(s_q, d_q)$
            $define\_offset\_length(s, d)$

Figure 7.20: Monolingual detection algorithm at PAN. d_q (d) is a plagiarism suspicion (potential source) document; $case\_fold$ performs case folding over d, $discard\_sw$ applies stopwording, $discard\_single\_chars$ discards those tokens with one character, $stem$ applies stemming; $index$ generates an index of D with word n-grams as terms; $split$ splits d_q into chunks of word n-grams of length k, argmax returns the document $d \in D$ for which $sim(d_q, d)$ is maximum; $source$ returns the potential source document for s_q, $merge$ merges two contiguous (or overlapping) detections, $define\_offset\_length$ defines the offset and length of the plagiarism suspicious (source) chunks in d_q (d).

rowed and source chunks. The entire algorithm is depicted in Fig. 7.20. First, documents written in a language different than English are discarded. At pre-processing, standard operations are applied to both suspicious and source documents: case folding, stopwords deletion, and stemming. Afterwards, the n-grams in D are indexed. At retrieval, the chunks in d_q, composed of sequences of n-grams of length k, are queried to the index, retrieving only the most similar document $d \in D$. During detailed analysis, those $s_{q,i}$ for which the most similar document is the same d are combined in order to compose a plagiarism candidate. When the resulting candidate chunk $s_q = \{s_{q,i} \cup \ldots s_{q,i+k}\}$ reaches a given length (e.g. 4 contiguous chunks), it becomes a candidate of plagiarised chunk. The final step is refining the most likely offset and lengths for the suspicious and source chunks. In order to do that, the common beginning and ending n-grams in the resulting s_q and s are compared, looking for the best matches. This is indeed the procedure that Rodríguez Torrejón and Martín Ramos (2010, 2011) call *referential monotony*.

7.4.1 Results and Discussion

Firstly we concentrate on the results obtained for the cases of text re-use in the PAN-PC-10. On the basis of the results we previously obtained with other text re-use corpora (cf. Chapter 5), we decided to use word 3-grams. Two parameters are left to explore: (a) l the length of the chunks s_q and (b) m the minimum number of chunks that may compose a valid case of re-use (and which makes the filtering possible). The confusion matrix $l \times m$ with the *plagdet* values obtained when combining these parameters are included in Table 7.7.

Table 7.7: Confusion matrix for the results obtained at PAN-PC-10 in terms of *plagdet*, on the basis of the word n-grams model. On the left hand side we have the minimum number of chunks that are considered relevant. On top the length of the chunk, in number of word 3-grams.

		Length of the chunk (l)							
		10	12	14	16	18	20	22	24
min chunks (m)	3	0.543	0.561	0.572	0.576	0.580	0.582	0.582	0.580
	4	0.593	0.602	0.605	0.602	0.601	0.598	0.596	0.592
	5	0.602	0.607	**0.607**	0.604	0.601	0.598	0.596	0.593

As the values show, the top *plagdet* value is around 0.60 when considering $m = \{3, 4, 5\}$. As expected, the higher the value of m, the shorter the chunks can be. As mentioned by Clough (2001), the average sentence length in English, for instance, considering the British National Corpus, is between eleven and twenty words. Kornai (2007, p. 188) estimates that the median sentence length in journalistic text is above fifteen words. Considering chunks of 12 n-grams, after stopword deletion, would mean that a few more than twenty tokens are being considered, a value that approximates the average length of long sentences.

Figure 7.21 includes the results of this best configuration for every partition of the PAN-PC-10, according to different kinds of plagiarism. In the first row, the evaluation considering the overall corpus as well as the external partition is represented. Obviously the recall —and therefore plagdet— increases when the cases of re-use without available source are discarded. The rest of values were computed considering the external partition only. Respect to the paraphrase partitions, similar values of precision are obtained for the verbatim and automatically obfuscated cases; but a drop occurs with the manually generated cases. Recall decreases more gradually as the level of paraphrasing increases. The behaviour of the document and cases length is just as discussed already in Section 7.2.2. Because shorter plagiarism cases have a higher level of paraphrasing, they are the hardest cases to detect. This fact stands for shorter documents as well. The amount of plagiarism per document is clearly not a factor.

By considering these results, we tried the best parameters on the PAN-PC-11 corpus, just like a participant would have made in the 2011 competition (i.e., training with the PAN-PC-10, testing with the PAN-PC-11). With $m = 5$ and $l = 14$, the results over the PAN-PC-11 are as follows:

$$plagdet = 0.20 \qquad prec = 0.88 \qquad rec = 0.14 \qquad gran = 1.31$$

i.e., it would have been sixth in the competition (cf. Fig. 7.13), with a very competitive precision (note that characterising the texts with ordered n-grams and applying the same model results in an increase of *plagdet* to 0.23 only (Rodríguez Torrejón and Martín Ramos, 2011); however, this value was obtained by intending to detect cases of translated plagiarism as well, whereas we are concentrated on monolingual cases only).

We further tuned the word n-grams model considering the PAN-PC-11 corpus. The best parameters obtain $plagdet = 0.26$ ($prec = 0.83$, $rec = 0.17$, $gran = 1.10$). These values are obtained with a high $l = 40$ and $m = 4$.

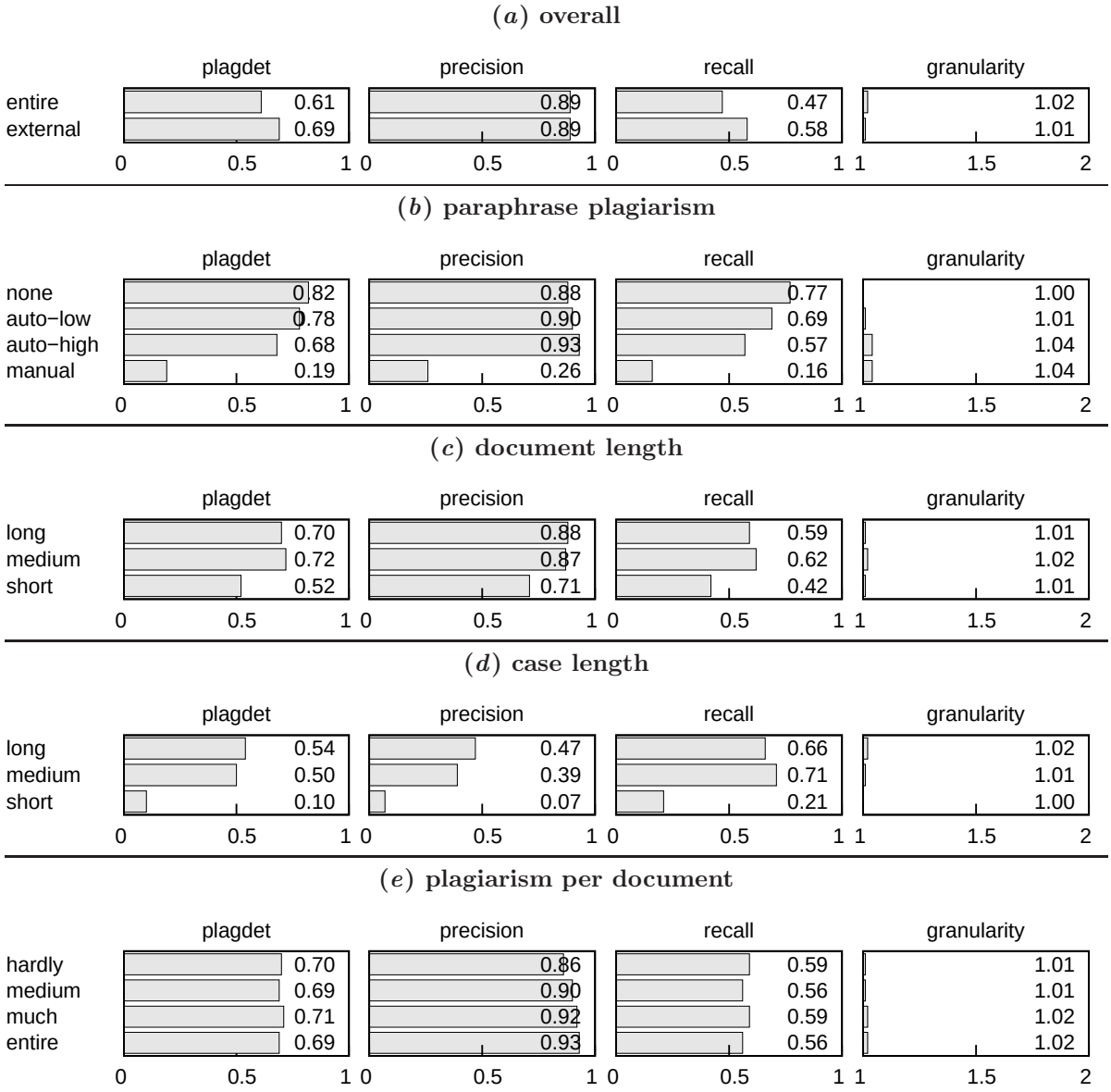


Figure 7.21: Overall results of word n-grams on the PAN-PC-10. From top to bottom: (a) overall: entire corpus and external partition only; (b) paraphrase plagiarism: none (verbatim copy), automatic with low and high obfuscation levels, and manually created; (c) document length (in pages): long (100–1,000), medium (10–100) and short documents (1–10); (d) case length (in tokens): long (3,000–5,000), medium (300–500), and short (50–150) re-used cases; (e) plagiarism per document (in percentage): hardly (5%–20%), medium (20%–50%), much, (50%–80%) and entire (>80%).

7.5 Detection of Cross-Language Plagiarism @ PAN

As discussed already, the competition participants were focussed on the development of better monolingual models. The cross-language cases deserved, in most cases, just an extra pre-processing operation: the translation of Spanish and German documents into English. The rest of the process is just as in the monolingual settings. However, we argue that such a language normalisation process is not the best approach. As seen in

Sections 7.2.2 and 7.3.2, this strategy has obtained encouraging results in the second and third competitions. Nevertheless, we identify an important weakness.

For explanatory purposes, let us remind the process followed to generate cases of translated re-use in the PAN-PC-09 and PAN-PC-10: (*i*) a text fragment s in Spanish or German is selected from d; (*ii*) s is translated into English, using Google translator, to generate s_q; and (*iii*) s_q is inserted into d_q. Now, let us remind the detection process followed by most participants: (*a*) the language of d is detected, (*b*) if d is written in Spanish or German, it is translated into English, with Google translator; (*c*) a monolingual detection is further performed. Steps (*ii*) and (*b*) are identical, using the same resources. Hence the problem becomes close to that of detecting verbatim copies. As a result, the detection quality was very high at PAN 2010 (cf. Fig. 7.8). Nevertheless, the inclusion of further obfuscated cases in the PAN-PC-11 showed that they were overestimated. When the cross-language case does not consist of an exact translation, detectors face big problems to detect it (cf. Fig. 7.15).

What if no manual obfuscation follows translation, but a manual translation is performed? Even more simple, what if a different translation service is used during the plagiarism generation and during the detection process? To understand this phenomenon we carried out a simple experiment. A text fragment s_{es} taken from the source partition of the PAN-PC-11 corpus was randomly selected and translated into English both manually and automatically. The outcome is in Table 7.8. An important difference between our four translations and the plagiarised text in the corpus occurs both at lexical (e.g. *Can* versus *Dog* (sic), *He* versus *It* versus *She*) and syntax level (e.g. *He went his way slowly* versus *Continued slowly his way* versus *She slowly continued her way*), sometimes resulting in a semantic drift.[18] As a result, the similarity when considering word n-grams (a common monolingual similarity strategy) is high between t_{pan} and t_g, but not so high with respect to other translations. As those cases could be considered highly obfuscated, the detection is much harder.

In Chapter 6 we offered an overview of the entire process necessary for cross-language plagiarism detection. Moreover, we described the proposed CL-ASA model (cf. Section 6.3) and we compared to CL-ESA and CL-CNG models. However, we did not further analyse its performance and the experiments reported were done at document and sentence level, independently.

[18]Note that t_m, whereas manually translated, is grammatically incorrect (it was translated by a Spanish native speaker and reviewed by two other people which native language was not English, and none of them detected it!). It is worth considering that the genre of most of the documents used at PAN for simulating the cases of plagiarism is literature (most of it old enough to be copyright-free material). The translation (and in general paraphrasing) of this kind of text is difficult. Moreover, when translating this fragment the purpose was not plagiarising it, but simply translating it: without time constraints and without trying to hide any fault. This is a common concern for the "plagiarism" cases in this family of corpora: whether it contains realistic cases of plagiarism. This is an interesting topic which could better focus the future research on this topic.

Table 7.8: Example of manual and automatic translations from Spanish (s_{es}) into English. Automatic translations: t_g Google (`http://translate.google.com`), t_a Apertium (`http://www.apertium.org`), t_b Babelfish (`http://babelfish.yahoo.com`), and manual translation (t_m) were applied. Similarities computed respect to t_{pan}, using the cosine measure with $\{1,2,3\}$-grams, after standard pre-processing. t_{pan} is the plagiarised fragment of s_{es} in the PAN-PC-11 corpus (from documents `suspicious-document00427.txt` and `source-document10990.txt`).

id	Text	$sim_n(t_g, t_x)$		
s_{es}	*Continuó lentamente su camino, para no alcanzar a la familia de Can Mallorquı. Margalida se había reunido con su madre y su hermano. Los vio desde una altura, cuando el grupo caminaba ya por el valle con dirección a la alquería.*	$n=1$	$n=2$	$n=3$
t_{pan}	He went his way slowly, not reaching the Mallorquí Can family. Margalida he joined his mother and brother. He watched from a height, when the group walked Now for the valley towards the farm.	1.00	1.00	1.00
t_g	He went his way slowly, not reaching to the family of Can Mallorquí. Margalida had met with his mother and brother. He watched from a height, when the group walked through the valley and towards the farmstead.	0.87	0.55	0.43
t_a	Continued slowly his way, not to achieve to the family of Can Mallorquí. Margalida Had gathered with his mother and his brother. It saw them from an altura, when the group walked already by the valley with direction to the alquería.	0.68	0.20	0.09
t_b	It slowly continued its way, not to reach to the family of Dog Mallorquí. Margalida had met with its mother and her brother. It saw them from a height, when the group walked already by the valley in the direction of the farmhouse.	0.63	0.20	0.12
t_m	She slowly continued on her way, trying not to reach the Can Mallorquí's family. Margalida had joined her mother and brother. She saw them from a high, while she was already walking on the valley, heading to the farm.	0.52	0.15	0.03

7.5.1 Cross-Language Detection Strategy

Here we propose an integral model for cross-language plagiarism detection including the steps of Fig. 5.2 (page 114), but without relying on MT systems. We designed a number of experiments to compare CL-ASA with CL-CNG (cf. Section 6.2.2.1) which in our experiments with English-[German, Spanish] showed good results (cf. Section 6.4).

The process is as follows. For cross-language heuristic retrieval, we select the top 50 $d' \in D'$ for each d_q according to $sim(d_q, d')$. The steps of cross-language detailed analysis, and heuristic post-processing are performed as explained in Fig. 7.22. We opted for representing the documents by means of sentences. The length of the chunk is of 5 sentences, with a step of 2. We decided to use five sentences aiming at considering text fragments that resemble paragraphs.[19] During the pairs identification step, we select the 5 most similar source fragments s for every s_q. $sim(s_q, s)$ is either computed with CL-ASA or CL-CNG. If the distance in characters between two (highly similar) candidate

[19]Optionally, a fixed length could have been chosen, for instance, on the basis of characters. However, such a decision would have caused the information provided by the length model in CL-ASA to be constant (cf. Section 6.3).

Given d_q and D':

// **Detailed analysis**
$S_q \leftarrow \{split(d_q, w, l)\}$ $S' \leftarrow \{split(d', w, l)\}$
for every $s_q \in S_q$:
  $P_{s_q,s'} \leftarrow \text{argmax}^5_{s' \in S'}\, sim(s_q, s')$
// **Post-processing**
until no change:
  for every combination of pairs $p \in P_{s_q,s'}$:
    if $\delta(p_i, p_j) < thres_1$:
      $merge\_fragments(p_i, p_j)$
// **Output**
return $\{p \in P_{s_q,s'} \mid |p| > thres_2\}$

Figure 7.22: Cross-language detailed analysis and post-processing. $split(d_q, w, l)$ splits d_q (d) into chunks of length w with step l; $\text{argmax}^5_{s \in S}\, sim(s_q, s)$ retrieves the 5 most similar fragments $s \in S$ respect to s_q; $\delta(p_i, p_j)$ measures the distance, in characters, between the suspicious and source fragments in p_i, p_j; $merge\_fragments(p_i, p_j)$ merge the plagiarism-source text fragments in the pairs p_i, p_j. $thres_1$ represents the maximum distance allowed between p_i, p_j to be merged; $thres_2$ is the minimum number of chunks p has to be composed of to be considered a plagiarism candidate.

pairs $\delta(p_i, p_j)$ is lower than a given threshold (in our case $thres_1 = 1,500$), p_i and p_j are merged. Only those candidates that are composed of at least three of the identified fragments ($thres_2$) are returned (these thresholds were defined empirically). In order to compute $sim(d_q, d')$ and $sim(s_q, s')$ we consider either CL-ASA and CL-CNG.

As described in Section 6.3, CL-ASA is composed of two models: length and translation models. The translation model relies on a statistical bilingual dictionary. Different approaches to cross-language plagiarism detection (e.g. Ceska *et al.* (2008)), stress that the incompleteness of this kind of resources harms the quality of the detections. Therefore, we have opted for experimenting with three dictionaries: (*i*) a dictionary empirically estimated from a parallel corpus, the same one we used in the experiments of Section 6.4; (*ii*) a dictionary of inflectional forms (INF), produced from a "traditional" bilingual dictionary, where all the possible inflectional forms of a word are generated, and the weights are estimated from large corpora distributions (Sidorov *et al.*, 2010); and (*iii*) a stemmed version of the previous dictionary (STEM), where the weights are accumulated and distributed over the entries' stems (cf. Appendix A). In the three cases we explore the impact of considering, for each word in a language, only the k best translations, those with the highest probabilities up to a minimum probability mass of 0.20. We call these dictionaries [JRC|INF|STEM].xx, where xx defines the considered probability mass.

For heuristic retrieval, $sim(d_q, d')$ CL-ASA neglects the length model. The reason is that the overall lengths of d_q and d' are completely independent from those of the specific borrowed fragments. CL-CNG, is used with character 3-grams in both steps (therefore, from now on we refer to the model as CL-C3G).

7.5.2 Experimental Setup

The corpus we use is the PAN-PC-11 (cf. Section 4.2.3.5). We focus on the Spanish-English translated plagiarism cases. Such partition comprises of 304 suspicious and 202 potential source documents, with two types of borrowing: automatic translation (*auto*) and further manually obfuscated automatic translation (*manual*). It is worth noting that, $s_q \in d_q$, the borrowed fragment, is in general on a different topic to that of d_q. For experimental purposes, we use three partitions of the corpus (the subscript x represents the experiment C_x the partition is considered in):

(*i*) C_1 is composed of the specific $\{s_q, s'\}$ pairs, which are considered as entire documents. This partition is composed of $2,920$ source and $2,920$ plagiarised documents (fragments).

(*ii*) C_2 includes the entire set of 304 suspicious and 202 potential source documents, with plagiarised fragments within them. The document d from which d_q's borrowings come from, is identified.

(*iii*) C_3 is as C_2, but no preliminary information about the source documents exist.

We designed a set of three experiments to investigate the performance of CL-ASA and compare it with CL-C3G, on the different plagiarism detection steps and scenarios.

Experiment 1: Cross-language ranking. This experiment resembles Experiment 1 of Section 6.4.2. We are given d_q and D' where d_q is entirely plagiarised from $d' \in D'$. The task is finding the source of d_q. This depicts the scenario where almost the whole document is plagiarised from one source. Moreover, it is an approximation to the scenario where fragment $s_q \in d_q$ and d_q are actually on the same topic (something that does not occur in the PAN-PC-11 corpus) and we are at the heuristic retrieval step. This experiment is also used to tune the parameters of CL-ASA by exploring different dictionaries, probability masses and the inclusion, or not, of the language model.

Experiment 2: Cross-language fragment-level re-use identification. This experiment is a simplification of the problem faced during the plagiarism detection competition. We are given d_q and d' and the task is finding $s_q \in d_q$ and $s' \in d'$ such that s_q is a plagiarised fragment from s'. This experiment depicts the scenario where d_q and d' are already identified and we aim at locating the borrowed text fragments, i.e., the detailed analysis step (the heuristic retrieval process is assumed to be solved).

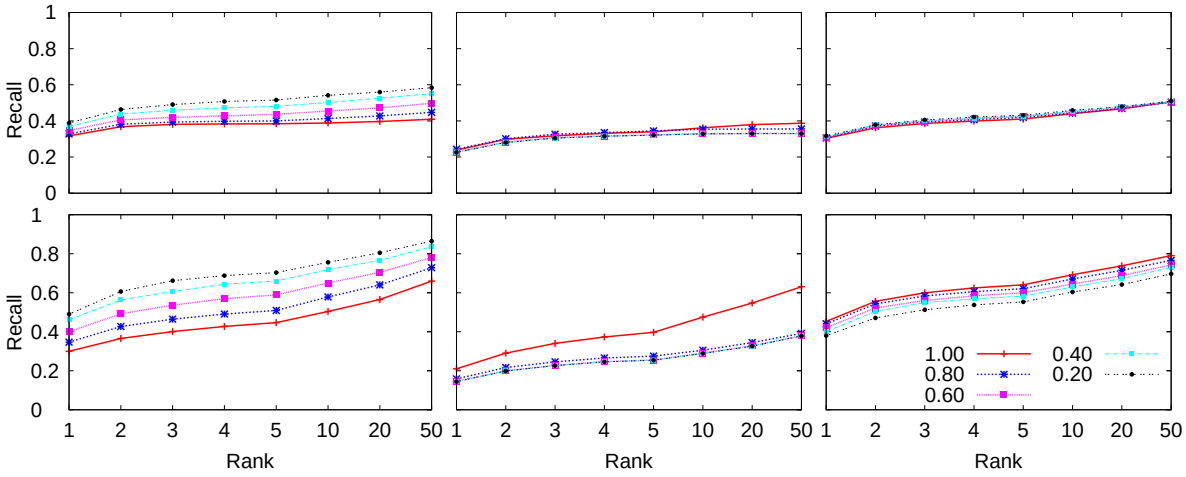
Experiment 3: Cross-language plagiarism detection. We face the actual cross-language external plagiarism detection challenge, as defined in the competition. We are given d_q and D' and the task is finding $s_q \in d_q$ and $s' \in d'$ ($d' \in D'$) where s_q is plagiarised from s'.

7.5.3 Results and Discussion

Experiment 1. Within this experiment we tune the parameters of CL-ASA. We tried with different probability masses for the dictionaries: $1.00, 0.80, \ldots, 0.20$.[20] Moreover, we aimed at determining how influential the length model actually is. The obtained results are represented in Fig. 7.23.

First, we analyse the results obtained with the translation model only, neglecting the length model. The best results are obtained with the JRC dictionary, using a probability mass= 0.20. The best results with INF and STEM come with mass= 1.0; i.e., the entire dictionary. On the one hand, JRC is empirically generated from a parallel corpus; noisy entries (with low probabilities) are included. Reducing the probability mass is roughly equivalent to discarding such noisy entries. We had noted already this behaviour (Barrón-Cedeño *et al.*, 2008; Pinto *et al.*, 2009). On the other hand, INF and STEM are generated

[20]That is, in the first case, every possible translation for a word is considered, for 0.80 only the most likely translations up to reaching a probability mass of 0.80 and so on.



(*a*) JRC (*b*) Inflectional (*c*) Stemmed

Figure 7.23: Comparison of dictionaries with and without length model for CL retrieval. Three dictionaries are considered: (*a*) JRC, (*b*) inflectional, and (*c*) stemmed. In the top plots the translation model is applied. In the bottom plots the translation and length models are applied. Evaluation in terms of $rec@k$, $k = \{1, 2, 3, 4, 5, 10, 20, 50\}$.

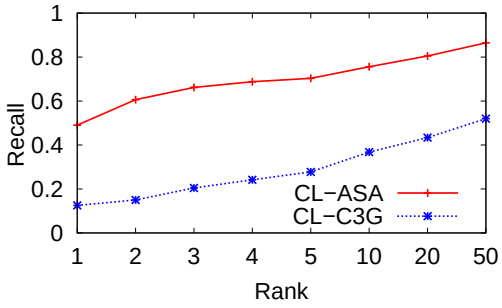


Figure 7.24: Comparison of CL-ASA and CL-C3G for cross-language retrieval. CL-ASA applied with length model and translation model (probability mass = 0.20). Evaluation in terms of $rec@k$, $k = \{1, 2, 3, 4, 5, 10, 20, 50\}$.

from traditional dictionaries, and every entry is presumably a correct translation. Nearly the same result can be obtained when considering the different amounts of entries (a probability mass of 1.0 offers slightly better results). The stemmed dictionary performs better, as the probability dispersed among all the inflections of a word are now better concentrated.

A common, highly relevant, phenomenon occurs with the three dictionaries: it is not necessary to consider every potential translation for every single word. Good results are obtained already with 0.20 probability mass. As expected, in the three cases, the length model empowers the ranking capabilities of the translation model, but the order remains; i.e., 0.20 (1.0) probability mass is the best option for the JRC (inflectional and stemmed) dictionary. From these results, it is evident that even when a dictionary is obtained from documents on topics different to those of the analysed texts, CL-ASA can still assess properly the similarity between some texts.

Figure 7.24 compares the best CL-ASA configuration (JRC translation model with 0.20 probability mass and length model) to CL-C3G. Back in Chapter 6, we had seen that CL-ASA outperformed CL-C3G when retrieving documents' translations (cf. Table 6.4, page 156). However, we had taken those results with caution, as they had been obtained

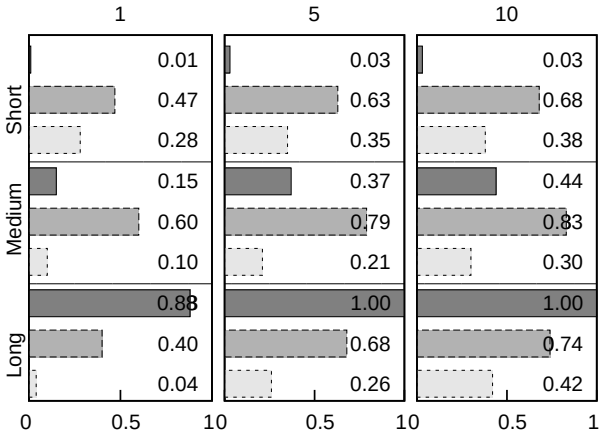


Figure 7.25: Comparison of CL-ASA and CL-C3G cross-language ranking over documents of different lengths. Evaluation in terms of rec@rank k, $k = \{1, 5, 10\}$. Darkest bars correspond to experiments with CL-ASA's translation model. Dark bars correspond to CL-ASA as well, this time including the length model. Light bars correspond to CL-C3G.

by considering a different partition of the same JRC-acquis corpus (i.e., the contents of the training and test partition were related). Now, CL-ASA clearly outperforms CL-C3G again, but this time the corpus has nothing to do with JRC-related documents, showing the robustness of the model.

We further analyse the results based on the length and type of re-use. Note that the length of the documents determines the amount of information available for the model. The obtained results for short, medium, and long texts are displayed in Fig. 7.25. As expected, the small amount of information available for CL-ASA when considering short texts harms its performance. However, the language model manages to improve the result, significantly overcoming CL-C3G again. For medium documents, even applying the translation model alone obtains better results than CL-C3G. However, an interesting phenomenon occurs when dealing with long documents: the length model causes the accuracy of CL-ASA to decrease. This is in agreement with our previous observations (Barrón-Cedeño *et al.*, 2010c): CL-ASA is sensitive to the amount of information it can exploit. The more text, the better the translation model performs, but not so for the language model. Obviously this behaviour can be used as a parameter: neglect the language model for long documents. Interestingly, the longer the document, the worst CL-C3G performs. This may be caused by the dispersion of data for longer texts.

Our last comparison regards to determining how the models perform when dealing with exact and further obfuscated —i.e., further paraphrased— translations. The results are displayed in Fig. 7.26. As expected, further obfuscated cases (i.e., translation plus paraphrasing) are harder to detect. However, still 37% of them are located at rank 1 by CL-ASA (respect to only 11% for CL-C3G). As the best CL-ASA results are in general obtained with JRC.20 including the language model, we use this version in Experiments 2 and 3.

Experiment 2. The results of the detailed analysis experiment are presented in Fig. 7.27, considering different cases' aspects. The precision of CL-ASA and CL-C3G are comparable, regardless of the length or nature of the case. Still, as in experiment 1, CL-ASA clearly outperforms CL-C3G in terms of recall. Differently to the previous experiments, the best results are not obtained with medium, but with long plagiarism cases. However, note that the chunks considered for comparison are fixed length: five sentences. Therefore, this behaviour does not contradict the previous results. The reason to fail

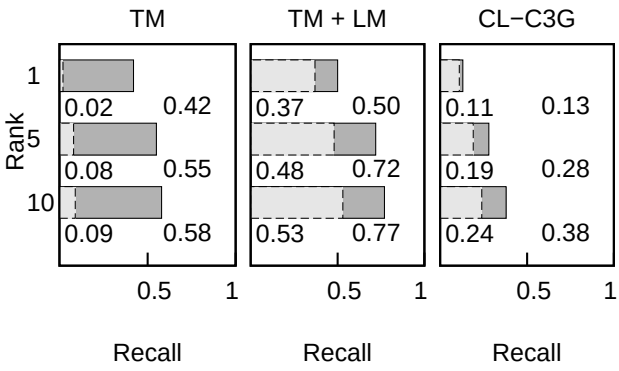


Figure 7.26: Comparison of CL-ASA and CL-C3G over documents with different reuse types. Evaluation in terms of *rec*@rank k, $k = \{1, 5, 10\}$. We include CL-ASA with the translation model alone (TM) and including the length model (TM + LM). Recall values below each bar for further obfuscated (light bars, left) and exact translations (dark bars, right).

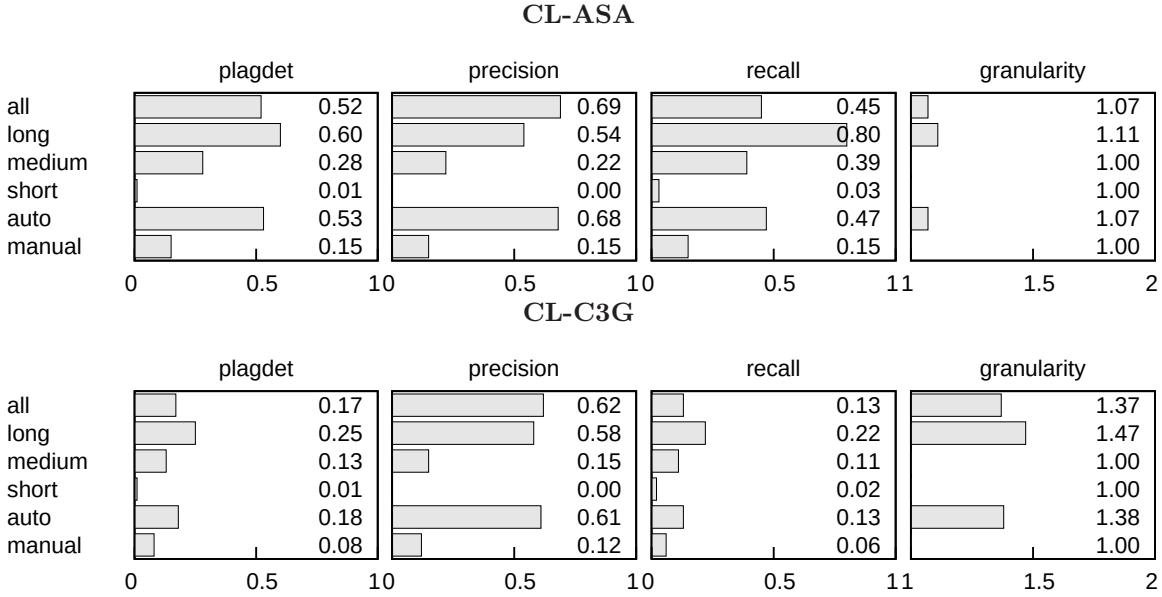


Figure 7.27: Cross-language plagiarism detection considering documents' pairs (Experiment 2). Plagdet, precision, recall and granularity computed for the entire collection C_2 and the different plagiarism types within it.

with the short cases is precisely on the heuristic we use to determine that an actual case of plagiarism is at hand: the algorithm needs evidence in terms of matching consecutive chunks, causing the short plagiarism cases to go unnoticed. Moreover, the majority of further paraphrased cases of the PAN-PC-11 corpus are short and hence harder to detect. For long cases F-measure= 0.64 ($plagdet = 0.60$) is obtained, but there is a drop in detection of medium and, especially, short cases.

It is remarkable to note that our selection and discrimination strategy (post-processing) obtains low values of granularity when measuring similarities by means of CL-ASA. The inaccurate similarity measures of CL-C3G cause it to miss more re-used fragments and get higher granularity values. The performance of both models is as expected: better for longer and for non-obfuscated cases. This behaviour had been identified by Corezola Pereira *et al.* (2010a) already when dealing with cross-language plagiarism.

Experiment 3. This experiment depicts the overall detection process. The performance of the heuristic retrieval process, i.e., properly including the source document of a case

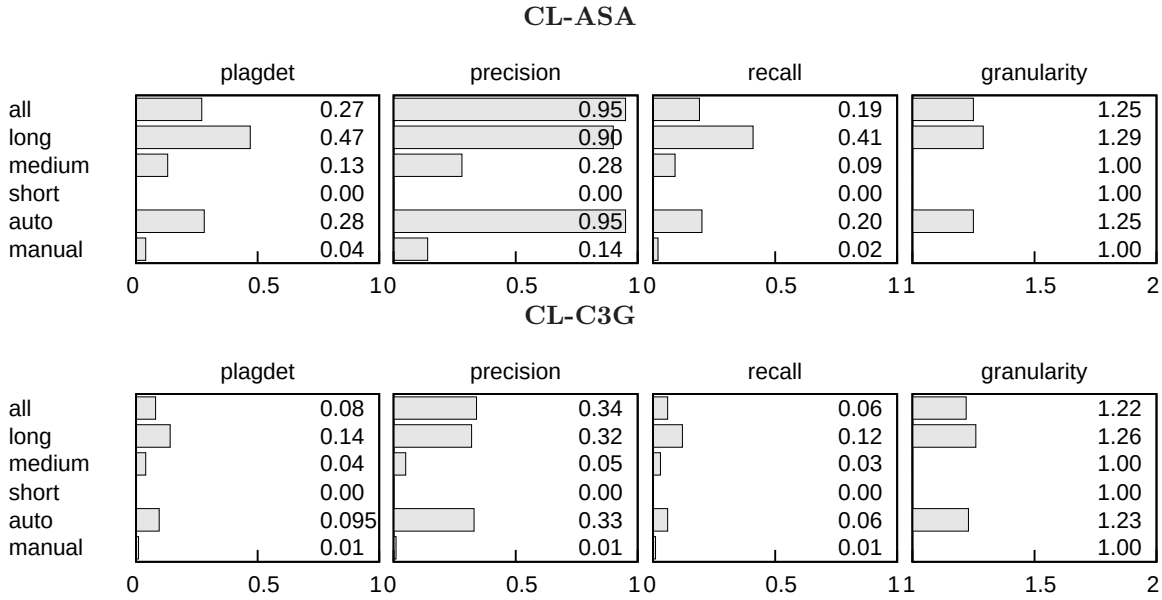


Figure 7.28: Entire process of cross-language plagiarism detection considering the entire corpus (Experiment 3). Plagdet, precision, recall and granularity computed for the entire collection C_3 and the different plagiarism types within it.

within the 50 retrieved documents, is roughly the same for both CL-ASA and CL-C3G: 31% and 29% respectively. This low accuracy is explained by the fact that the source and plagiarised documents are not on common topics. We believe that the results of this step would be better in more realistic scenarios, where s_q is extracted from a document on the same topic of d_q (this belief is supported by the results of Experiment 1).

The results obtained after the overall process (including detailed analysis and post-processing) are presented in Fig. 7.28. As in the previous experiments, CL-ASA outperforms CL-C3G, regardless of the length or nature of the plagiarism case. The overall recall for CL-ASA is 0.19 with a very high precision. As in experiment 2, CL-ASA performs at its best with long cases. Recall reaches a value of 0.41, still with a high precision. As expected from the previous experiments, exact translations are easier to detect than further obfuscated. However, in agreement with Experiment 1, neither CL-ASA nor CL-C3G are able to detect short cases. We consider that two reasons are behind the failure: (*i*) most of the further paraphrased translated cases in the corpus, the hardest to detect, are among the short cases, and (*ii*) the decision of defining a minimum length a candidate must surpass to consider it relevant (this decision is in commitment with precision).

The obtained results with CL-ASA and its reduced dictionary somehow contradict the findings of Ceska *et al.* (2008): that the incompleteness of the language resource (in their case a thesaurus, in ours a dictionary) causes difficulties to plagiarism detection. We tried with a "complete" and a limited dictionary. The second one performed best, regardless its bias and incompleteness. CL-ASA showed remarkable performance when detecting plagiarism of entire documents, including further paraphrased translations. When aiming at detecting specific borrowed fragments and their source, both short and further paraphrased cases certainly cause difficulties to the detection. In our strategy,

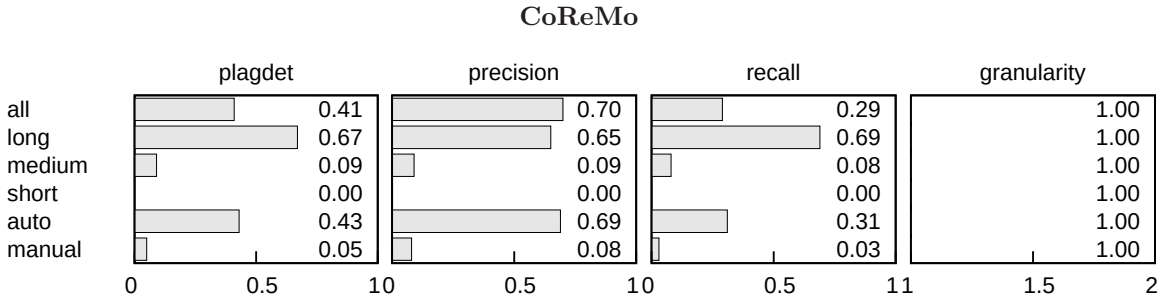


Figure 7.29: Evaluation of cross-language plagiarism detection with CoReMo (Rodríguez Torrejón and Martín Ramos, 2011), for comparison of Experiment 3.

we opted for betting for a high precision (for some types of plagiarism higher than 0.90). As a result, if the detector identifies a case of potential plagiarism, it is certainly worth analysing it.

As a point of comparison, we consider the results obtained by Rodríguez Torrejón and Martín Ramos (2011) and their cross-language section of the CoReMo system. We select only this approach because it operates under the same conditions than ours: without relying on the Google translation service. Their results are in Fig. 7.29. The first fact is that neither CoReMo is able to detect short cases of translated plagiarism (which include further paraphrased cases). We consider that two reasons are behind the imprecisions of CoReMo and our detector: (*i*) most of the further paraphrased translated cases in the corpus, the hardest to detect, are among the short cases, and (*ii*) as in our case, CoReMo defines a minimum length a candidate must surpass to consider it as relevant (this decision is in commitment with precision). The second fact is that our CL-ASA-based detector outperforms CoReMo in medium cases. In all the cases, the precision we obtain is much higher.

As in the monolingual setting, further obfuscated cases seem to remain an open issue in cross-language plagiarism detection.

7.6 Chapter Summary

With the advent of the International Competition on Plagiarism Detection in 2009, more than thirty plagiarism detection methods have been compared by means of the same data collection and evaluation measures. Researchers that had to use private collections of documents, which cannot be freely provided to others for ethical reasons, have found a common test-bed for improvement. As a result, more and better models are being generated as a countermeasure to plagiarism. Four of the developed systems are used in academic and commercial environments in Chile, Czech Republic, Slovakia, and Ukraine, some of them working on alphabets other than Latin (e.g. Cyrillic).

In the first part of the chapter, we analysed the different intrinsic and external models developed in the framework of the competition. From the results, it seems reasonable to consider that the external detection of verbatim copies is no more an open issue. Nevertheless, when plagiarism includes paraphrasing, state of the art detection models decrease their performance. We observed as well that cross-language plagiarism detection based on a preliminary translation of documents into a common language and their further comparison with monolingual models is *in vogue*. However, we believe that this does not represent the best approach, as translating every document on the Web is not doable.

Different lessons have been learned. From PAN 2009 we observed that the techniques based on dot-plot represent one of the best options to consider when trying to perform external analysis. From PAN 2010, we learned that more than "classic" word n-grams can be used to characterise a document, and that a simple ordering of an n-gram's words can considerably improve the results as it allows to detect some paraphrased cases, mainly those based on re-ordering. From PAN 2011, we realised that translating all the documents into a common language may not always be the best idea, mainly when the translation applied during plagiarism implies further paraphrasing. Regarding intrinsic detection, during the three editions we observed that just a few approaches exist up to date and their performance remains low. The best performing models are simple: based on character n-grams or words counting. Plenty of space for improvement exists. As a consequence of the competition, proper attention is now paid to the setting of offsets that divide plagiarised from original texts.

In the second part of the chapter, we investigated how some of the plagiarism detection models we had proposed in previous chapters worked. A simple model based on a word n-grams Boolean comparison, showed to be competitive, regardless its simplicity, when dealing with monolingual plagiarism.

We paid special attention to the cross-language plagiarism detection problem. The plagiarism detector relied on our proposed cross-language similarity model called CL-ASA. CL-ASA exploits a combination of bitext alignment, statistical machine translation, and cross-language information retrieval techniques. CL-ASA was tested on detecting the cases of cross-language re-use in the PAN-PC-11 and compared to CL-C3G. In every step of the process CL-ASA clearly outperformed CL-C3G. As both CL-C3G and CL-ASA do not depend on online MT systems they cannot be directly compared to the best performing cross-language models at PAN (as they reduce the problem to

a monolingual verbatim copy detection, exploiting the same translator used to generate the corpus cases, something unrealistic). When compared to other systems that do not rely on Google translation, CL-ASA showed to be competitive.

Related publications:

- Stein, Potthast, Rosso, Barrón-Cedeño, Stamatatos, and Koppel (2011a)
- Potthast, Eiselt, Barrón-Cedeño, Stein, and Rosso (2011b)
- Barrón-Cedeño, Basile, Degli Esposti, and Rosso (2010d)
- Potthast, Barrón-Cedeño, Eiselt, Stein, and Rosso (2010d)
- Barrón-Cedeño and Rosso (2010)
- Potthast, Stein, Eiselt, Barrón-Cedeño, and Rosso (2009)

Chapter 8

Plagiarism meets Paraphrasing

> The bees pillage the flowers here and there but they make honey of them which is all their own; it is no longer thyme or marjoram: so the pieces borrowed from others are transformed and mixed up into a work all their own.
>
> Michael de Montaigne

Paraphrasing is the linguistic mechanism many plagiarism cases rely on. Indeed, as discussed in Section 2.5.2, there is a confusion on whether paraphrasing implies plagiarism, as it certainly does. Nevertheless, little attention has been paid to the relationship between paraphrasing and plagiarism. A proof of such lack of resources is the report of Maurer *et al.* (2006, p. 1074), where he realises that neither Turnitin, Mydropbox or Docol©c are able to uncover paraphrase plagiarism. Indeed, by 2006, no system (or model) was available that approached paraphrase plagiarism detection Maurer *et al.* (2006, p. 1079). The results obtained by the detectors evaluated in the 2010 and 2011 competitions on plagiarism detection confirm that this remains an open issue (cf. Sections 7.2 and 7.3).

One of the reasons is that the linguistic phenomena underlying plagiarism have been hardly analysed in the design of the plagiarism detection models, which we consider to be a key issue for their improvement. Paraphrasing, generally understood as sameness of meaning between different wordings, is the linguistic mechanism underlying many plagiarism acts and the linguistic process in which plagiarism is based. In paraphrase plagiarism, different operations are performed over the text, such as substitution for semantic equivalents and grammar changes. Even translation could be considered as a type of paraphrasing (cf. Chapter 6).

As paraphrases are in the core of every kind of plagiarism, more efforts are to be done to detect them. Plagiarism detection experts are starting to turn their attention to paraphrases, such as Burrows *et al.* (2012) who stress that in the 2010 edition of the PAN plagiarism detection, no plagiarism detector achieved a recall value higher than 0.27 for the paraphrase plagiarism cases.[1] In order to create feasible mechanisms for crowd-

[1]In contrast, the best performing model obtained values around 0.95 for both precision and recall on

sourcing paraphrases acquisition, Burrows *et al.* aimed at automatically discriminating, given two text fragments, whether they composed a paraphrase or not. A particularity of this research is that duplicates and near-duplicates were not considered as positive samples, a decision that they themselves accept as questionable. Their classifier considers ten paraphrase similarity measures as features, including the Levenshtein distance and word n-grams overlapping. The best classifier obtained a $prec = 0.98$, with $rec = 0.52$ (note that in this task precision is the most important factor as the amount of noisy entries depends on it). The nature of our research is completely different: in order to determine what paraphrasing types make plagiarism harder to be uncovered, we are interested in analysing the different types of paraphrasing strategies applied during the text re-use process.

In this chapter we analyse the relationship between paraphrasing and plagiarism, paying special attention to which paraphrase phenomena underlay plagiarism acts and which of them are (un)covered by plagiarism detection systems. We focus on monolingual paraphrase plagiarism. The relationship between plagiarism and paraphrasing is analysed, and the potentials of such a relationship in automatic plagiarism detection are set out. We aim at not only investigating how difficult detecting paraphrase cases for state-of-the-art plagiarism detectors is, in particular those applied in the 2010 edition of the plagiarism detection competition (cf. Section 7.2), but especially, in understanding what types of paraphrases are the most difficult to be detected.

In order to do that, we created the P4P corpus, annotating a portion of the PAN-PC-10 corpus (cf. Section 4.2.3.4) on the basis of a new paraphrase typology, which consists in an updated version of the one designed recently by Vila, Martí, and Rodríguez (2011). We mapped the annotated cases of plagiarism to those in the PAN-PC-10 corpus. Such mapping allows for analysing the results of the participants to the Second International Competition on Plagiarism Detection when aiming at detecting cases of paraphrase plagiarism (in this chapter we refer to this edition of the competition as "PAN-10 competition").

The rest of the chapter is structured as follows. Section 8.1 describes the paraphrase typology we used. Section 8.2 illustrates the construction of the P4P corpus. Section 8.3 discusses our experiments and results derived from mapping the P4P corpus and the results in the PAN-10 competition.

Key contributions One of the main outcomes of this chapter is the P4P corpus (Section 8.2). The corpus was annotated by linguists from the Universitat de Barcelona. At this step, the contribution of the author of this dissertation was proposing the adaptation of some of the labels and the annotation criteria. An important issue was how to organise the resulting annotated cases for analysing the performance of the different plagiarism detectors (Section 8.3). The organisation strategy, based on unsupervised learning (clustering), was designed and performed by the author of this dissertation (Section 8.3.1). The analysis of the results was a joint work with the researchers from the Universitat de Barcelona.

verbatim copies (cf. Figure 7.6).

8.1 Paraphrase Typology

Typologies are a precise and efficient way to draw the boundaries of a certain phenomenon, identify its different manifestations, and, in short, go into its characterisation in depth. Moreover, typologies are in the basis of any corpus annotation process, which has its own effects on the typology itself: the annotation process tests the adequacy of the typology for the analysis of the data, and allows for the identification of new types or the revision of the existing ones. In this section, after setting out a brief state of the art on paraphrase typologies and the drawbacks they present, the typology used for the annotation and subsequent analysis of the P4P corpus are described.

A number of paraphrase typologies have been built from the perspective of NLP. Some of these typologies are simple lists of paraphrase types useful for a specific system or application, or the most common types found in a corpus. They are specific-work oriented and far from being comprehensive: Barzilay, McKeown, and Elhadad (1999), Dorr, Green, Levin, Rambow, Farwell, Habash, Helmreich, Hovy, Miller, Mitamura, Reeder, and Siddharthan (2004) and Dutrey, Bernhard, Bouamor, and Max (2011), among others. Other typologies come from paraphrase related fields like editing (Faigley and Witte, 1981). Yet others classify paraphrases in a very generic way, setting out two or three types only (Barzilay, 2003; Shimohata, 2004). These classifications should not reach the category of typologies *strictu sensu*.

Finally, there are more comprehensive typologies, such as the ones by Culicover (1968), Dras (1999), Fujita (2005), and Bhagat (2009). They usually take the shape of extensive and very fine-grained lists of paraphrasing types grouped into bigger classes following different criteria. They generally focus on the specific paraphrase mechanisms, leaving the general phenomena at a second level. However, only the latter can account for paraphrasing in a comprehensive way. A list of specific mechanisms will always be endless.

The paraphrase typology relies on the paraphrase concept defined by Recasens and Vila (2010) and Vila *et al.* (2011). It consists of an upgraded version of the one presented in the latter: it classifies paraphrases according to the linguistic nature of their difference in wording. It attempts to capture the general linguistic phenomena of paraphrasing, rather than presenting a long, fine-grained and inevitably incomplete list of concrete mechanisms. It consists of a two-level typology of 20 paraphrase types grouped into six classes as represented in Fig. 8.1. Paraphrase types reflect a paraphrase phenomenon. Four of these classes (1 to 4) follow the classical organisation in formal linguistic levels from morphology to discourse. Class 5 is a miscellaneous comprising types not directly related to one single level. Finally, paraphrases in class 6 are not based on the form but on the semantic content.

This typology is proposed from the point of view of paraphrases and, therefore, is general enough to cover different kinds of rewriting. Still it has a direct application to the task of text re-use and plagiarism detection, where some prototypical rewriting operations have been identified. In particular, Clough and Gaizauskas (2009) gathered three frequently applied operations in journalistic re-use which are comparable to some entries of our typology: deletion (of redundant context and resulting from syntactic

1. **Morphology-based changes**
   - Inflectional changes
   - Modal verb changes
   - Derivational changes
2. **Lexicon-based changes**
   - Spelling and format changes
   - Same polarity substitutions
   - Synthetic/analytic substitutions
   - Opposite polarity substitutions
   - Inverse substitutions
3. **Syntax-based changes**
   - Diathesis alternations
   - Negation switching
   - Ellipsis
   - Coordination changes
   - Subordination and nesting changes
4. **Discourse-based changes**
   - Punctuation and format changes
   - Direct/indirect style alternations
   - Sentence modality changes
5. **Miscellaneous changes**
   - Syntax/discourse structure changes
   - Change of order
   - Addition/deletion
6. **Semantics-based changes**

Figure 8.1: Overview of the paraphrases typology.

changes), lexical substitution (synonymous and phrases), changes in syntax (word order, tense passive and active voice switching) and summarisation.

Belonging to one of the level of language classes does not necessarily mean that the level of language affected by the paraphrase phenomenon is only that which gives name to the class, but that the paraphrase appears at that level and other levels can be affected: a morphology based change (derivational) like the one in (1),[2] where the verb *reigned* is changed for its nominal form *reign*, has obvious syntactic implications; however, the morphology basis is the change considered for classification.[3] Moreover, although types in our typology are presented in isolation, they can overlap or be embedded one into another: in (2), both changes of order of the subject and the adverb, and two same polarity substitutions (*answered/said* and *carefully/cautiously*) can be observed. Finally, a difference between the cases in (1) and (2) should be set out: in the former, the morphology and syntactic changes are mutually dependent, so only one single paraphrase phenomenon is considered; in the latter, the same polarity substitutions and changes of order are independent, they can take place in isolation, so three paraphrase phenomena are considered.

[2]The examples are extracted from the P4P corpus. For clarity reasons, in some of them, only the fragment we are referring to appears; in others, its context is also displayed (with the fragment in focus in italics). Note that some of them (e.g. examples (18) and (19) in page 214) are grammatically incorrect. As this is how they were generated and included in the corpus, we did not correct them.

[3]In case of doubt of which change should be considered to be the basis, the one which is in a lower level in the typology is chosen. Morphology based changes are in the lowest level; discourse based changes, in the highest. Here we only refer to those classes based on levels of language.

(1) a. The eleventh year that the king *reigned*
b. The eleventh year of that king's *reign*

(2) a. "Yes," I carefully answered
b. "Yes," said I cautiously

The six classes in which the typology is divided comprise morphology (Section 8.1.1), lexicon (Section 8.1.2), syntax (Section 8.1.3), discourse based changes (Section 8.1.4), miscellaneous changes (Section 8.1.5), and semantics (Section 8.1.6).

8.1.1 Morphology-based Changes

These are those that arise at the morphology level of language.

Inflectional changes consist in adding inflectional affixes to words. In (3), a singular/plural alternation (*street/streets*) can be observed.

Modal verb changes consist in changes of modality using modal verbs, like *could* and *might* in (4).

Derivational changes consist in changes of category by adding derivational affixes to words. These changes comprise a syntactic change in the sentence where they occur. In (5), the adverb *hopefully* is changed to its verbal original form *hope*, with the consequent structural reorganisation.

(3) a. You couldn't even follow the path of the *street*
b. it was with difficulty that the course of *streets* could be followed

(4) a. I was pondering who they *could be*
b. I [...] was still lost in conjectures who they *might be*

(5) a. I have heard many *different* things about him
b. I have heard many accounts of him [...] all *differing* from each other

8.1.2 Lexicon-based Changes

They consist in those paraphrases that arise at the lexical level. This type gathers phenomena that, all having a lexical basis, are different in nature.

Spelling and format changes comprise changes in the spelling and format of lexical units, like case changes, abbreviations or digit/letter alternations. In (6), case changes occur (*PEACE/Peace*).

Same polarity substitutions change one lexical unit for another one with approximately the same meaning. Among the linguistic mechanisms under this type, we find synonymy, general/specific substitutions or exact/approximate alternations. In (7), *very*

little is more general than *a teaspoon of*.[4]

Synthetic/analytic substitutions consist in changes of synthetic structures to analytic structures, or vice versa. This type comprises mechanisms such as compounding/decomposition, light element or semantically emptied specifier additions, or genitive/prepositional phrase alternations. In (8-b), a (semantically emptied) specifier (*sequence*) has been added: it does not add new semantic content to the lexical unit, but it only emphasizes its plural nature.

Opposite polarity substitutions. Two phenomena are considered within this type. First, there is the case of double change of polarity, where a lexical unit is changed for its antonym or complementary and, in order to maintain the same meaning, another change of polarity has to occur within the same sentence. In (9), *failed* is substituted by its antonym *succeed*, and a negation is added to the latter. Second, there is the case of change of polarity and argument inversion, where an adjective is changed for its antonym in comparative structures. Here an inversion of the compared elements has to occur. In (10), the adjectival phrases *less serious* and *less common* change to the opposite polarity ones *far deeper* and *more general*. To maintain the same meaning, the order of the compared elements (i.e., what the Church considers and what is perceived by the population) has to be inverted.[5]

Inverse substitutions take place when a lexical unit is changed for its inverse. In order to maintain the same meaning, an argument inversion has to occur. In (11), *received [...] from* is changed to *awarded to*, and the arguments *he* and *the Geological Society in London* are inverted.

(6) a. Yet still they shout *PEACE*! *PEACE*!
b. And yet they are calling for *Peace*!–*Peace*!!

(7) a. *very little* vanilla
b. *a teaspoonful of* vanilla

(8) a. ideas
b. A sequence of ideas

(9) a. he *did not succeed* in either case
b. Leicester [...] *failed* in both enterprises

(10) a. the Church considers that this scandal is *less serious* and *less common* than it really is
b. the sense of scandal given by this is *far deeper* and *more general* than the Church thinks

(11) a. [resulted in] him *receiving* the Wollaston medal *from* the Geological Society in London in 1855
b. the Geological Society of London in 1855 *awarded to* him the Wollaston medal

[4]It should be noted that many of the examples displayed are context dependent.

[5]Note that, as expected, in both (9) and (10) the opposite polarity substitutions trigger further changes to maintain the meaning of the phrase. During the generation of artificial paraphrase plagiarism cases in PAN such substitutions are considered at random, but no further rephrasing is carried out (cf. Section 4.2.3.2). As a result, such cases in the PAN-PC corpora do not express the same idea, and could not be considered actual paraphrases.

8.1.3 Syntax-based Changes

Syntax-based changes are those that arise at the syntactic level of language.

Diathesis alternation type gathers those diathesis alternations in which verbs can participate, such as the active/passive alternation (12).

Negation switching consists in changing the position of the negation within a sentence. In (13-a), the verb *need* is negated; in (13-b), it is the noun *reference* which is negated.

Ellipsis includes linguistic ellipsis, i.e, those cases in which the elided fragments can be recovered through linguistic mechanisms. In (14-a), the subject *he* appears in both clauses; in (14-b), it is only displayed in the first one.

Coordination changes consist in changes in which one of the members of the pair contains coordinated linguistic units, and this coordination is not present, or changes the position or form in the other member of the pair. The coordinated clauses with de conjunction *and* in (15-a) are juxtaposed with a full stop in (15-b).

Subordination/nesting changes consist in changes in which one of the members of the pair contains a subordination or nesting, which is not present or changes the position or form in the other member of the pair. In (16-a), *they barred his admission* is a consecutive clause; in (16-b), it is the main clause (with a slightly different wording). Moreover, in (16-a), *the Russian law had limits for Jewish students* is the main clause and, in (16-b), it is a relative clause. These changes are dependent one to the other and constitute a single paraphrase phenomenon.

(12) a. our attention *was drawn by* our guide to a little dungeon
b. the guide *drew* our attention to a gloomy little dungeon

(13) a. One *does not need* to recognize a tangible object to be moved by its artistic representation
b. In order to move us, it needs *no reference* to any recognised original

(14) a. *He* equaled Salvini, in the scenes with Iago, but *he* did not in any point surpass him or imitate him
b. In the scenes with Iago *he* equaled Salvini, yet did not in any one point surpass him

(15) a. Altogether these works cost him almost £10,000 *and* he wrote a lot of small papers as well
b. It is estimated that he spent nearly £10,000 on these works. In addition he published a large number of separate papers

(16) a. the Russian law had limits for Jewish students so they barred his admission
b. the Russian law, which limits the percentage of Jewish pupils in any school, barred his admission

8.1.4 Discourse-based Changes

Discourse-based changes are those affecting the discursive structure of the sentence. This group covers a broad range of discourse reorganisations.

Punctuation and format type consists in any change in the punctuation or format of the sentence (not of a lexical unit, cf. lexicon-based changes). In (17-b), the list appears numbered and, in (17-a), it does not.

Direct/indirect style alternations consist in changing direct style for indirect style, or vice versa. The direct style can be seen in (18-a), and the indirect in (18-b).

Sentence modality changes are those cases in which there is a change of modality (not provoked by modal verbs, cf. modal verb changes), but the ilocutive value is maintained. In (19-a), an affirmative sentence can be observed; this is changed to an interrogative sentence in (19-b).

(17) a. You will purchase a return ticket to Streatham Common and a platform ticket at Victoria station
b. At Victoria Station you will purchase (1) a return ticket to Streatham Common, (2) a platform ticket

(18) a. The Great Spirit said that she is her
b. "She is mine," said the Great Spirit

(19) a. He do it just for earning money or to please Theophilus P. Polk or vex Hariman Q. Kunz
b. The real question is, will it pay? will it please Theophilus P. Polk or vex Harriman Q. Kunz?

8.1.5 Miscellaneous Changes

This section groups together those changes that, for different reasons, are related to more than one of the previous classes.

Changes in the syntax/discourse structure gather a wide variety of syntax/discourse reorganisations not covered by the types in the syntax and discourse classes above. An example can be seen in (20).

Change of order includes any type of change of order from the word level to the sentence level. In (21), *first* changes its position in the sentence.

Addition/deletion consists of all deletions of lexical and functional units. In (22-a), *one day* is deleted.

(20) a. Peace is much desirable than war
b. Dear as war may be, a dishonorable peace will prove much dearer

(21) a. We got to some rather biggish palm trees *first*
b. *First* we came to the tall palm trees

(22) a. As a proof of bed treatment, she took a hot flat-iron and put it on my back after removing my clothes.
b. *One day* she took a hot flat-iron, removed my clothes, and held it on my naked back until I howled with pain.

8.1.6 Semantics-based Changes

Semantics based changes are those that imply a different lexicalisation of the same content units.[6] These changes affect more than one lexical unit and a clear cut of these units in the mapping between the two members of the paraphrase pair is not possible. In the example (23), the content units TROPICAL-LIKE ASPECT (*tropical appearance/scenery [...] tropical*) and INCREASE OF THIS ASPECT (*added/more*) are present in both fragments, but there is not a clear cut mapping between the two.

(23) a. Which added to the tropical appearance
b. The scenery was altogether more tropical

8.2 Building the P4P Corpus

In order to compose the P4P corpus we considered a sub-sample of the PAN-PC-10 (cf. Section 4.2.3). This corpus was selected because it represents the standard *de-facto* corpus for the development of (simulated) plagiarism detection systems.[7] As already discussed, the most of the obfuscated cases in this corpus were generated automatically, i.e., rewriting operations were simulated by a computational process. The rest (6%) were created by humans who aimed at simulating paraphrase cases of plagiarism, known onwards as "simulated plagiarism" (cf. Section 8.2). We already noted the difficulty to detect simulated cases of plagiarism in the PAN-PC corpora (cf. Sections 7.2 and 7.3), and this issue was stressed by Stein *et al.* (2011a) as well.

This section describes how the P4P (Paraphrase for Plagiarism) corpus was built.[8] P4P consists of a sample of the PAN-PC-10 annotated with the paraphrase typology. We limited ourselves to the cases of simulated plagiarism in the PAN-PC-10 (plg_{sim}). They consist of pairs of source and plagiarism fragments, where the latter fragment was manually created reformulating the former. From this set, we selected those containing 50 words or less ($|plg_{sim}| \leq 50$). Note that beside the difficulty that paraphrasing implies for plagiarism detection, the shorter a plagiarised case is, the harder it is to detect it. Finally, 847 are the paraphrase pairs that complied with these conditions and have been selected as our working subset. The decision was taken for the sake of simplicity and efficiency, and it is backed by state of the art paraphrases corpora. As a way of illustration, the corpus of Barzilay and Lee (2003) includes examples of about 20 words only, and the Microsoft Research Paraphrase Corpus (MSRPC) (Dolan and Brockett, 2005) contains 28 words per case on average.

Tagset and scope. After tokenisation of the working corpus, the annotation was performed by, on the one hand, tagging the paraphrase phenomena present in each source/plagiarism pair with the typology and, on the other hand, indicating the scope of

[6] This type departs form the ideas in Talmy (1985).

[7] Having defined a proper typology and annotation mechanisms, further corpora of text re-use can be annotated, for instance those identified by Clough and Gaizauskas (2009). This would be an interesting future work to carry out.

[8] The corpus is freely available at `http://clic.ub.edu/en/`

each of these tags, i.e., the range of the fragment affected by the paraphrase phenomenon. The tagset consists of the 20 paraphrase types plus IDENTICAL and NON-PARAPHRASE. The former refers to those text fragments in the source/plagiarism pairs that are exact copies; the latter refers to fragments in the source/target pairs that are not semantically related. The reason for adding these two tags is our interest in having them identified in order to see the performance of the plagiarism detection systems regarding them in comparison to the actual paraphrase cases (cf. Section 8.3).

Regarding the scope of the fragments to be annotated, we do not annotate strings but linguistic units. In (24), although a change takes place between the fragments *other brothers with* and *brotherhood among*, the paraphrase mapping has to be established between *the other brothers* and *the brotherhood*, and *with* and *among*, two different pairs of linguistic units, fulfilled by a nominal phrase and a preposition, respectively. They consist of two same polarity substitutions.

It is important to note that paraphrase tags can overlap or be embedded one into another. In example (25), a same polarity substitution overlaps with a change of order in *wisely/sagely*. Tags can also be discontinuous. In (26-b), an example of a discontinuous same polarity substitution can be observed: *distinct [...] from*, with respect to *unconnected to* in (26-a).

(24) a. the other brothers with whom they lived
b. the brotherhood among whom they had dwelt

(25) a. shaking his head *wisely*
b. *sagely* shaking his head

(26) a. Still, in my opinion, the use of "Gothic" might well have origins *unconnected to* the emergence of the pointed arch.
b. But yet I imagine that the application of the term "Gothic" may be found to be quite *distinct*, in its origin, *from* the first rise of the Pointed Arch.

The scope affects the annotation task differently regarding the classes and tags. In concrete, we distinguish three scope annotation methods:

Morphology, lexicon and semantics classes, and change of order and addition/deletion types: Only the substituted, transposed or added/deleted linguistic unit(s) is (are) tagged. As some of these changes entail other changes (mainly inflectional or structural), two different attributes are provided: LOCAL, which stands for those cases in which the change does not entail any other change in the sentence; and GLOBAL, which stands for those cases in which the change does entail other changes in the sentence.

In (27), an isolated same polarity substitution takes place (*aging/older*), so the scopes *aging/older*[9] are taken and the attribute LOCAL is used; in (28), the same polarity substitution (*however/but*) entails changes in the punctuation. In that case, only *however/but* are taken as scope using the attribute GLOBAL. For the entailed changes pointed by the GLOBAL attribute, neither the type of change nor the fragment suffering the change are

[9]In the examples, neither the fragment set out nor italics necessarily refer to the annotated scope, but to the phenomenon we want to present or emphasize, although sometimes they coincide.

specified in the annotation.

(27) a. The *aging* trees
b. The *older* trees

(28) a. [...] wouldn ' t have been. *However*, she ' s not too resentful
b. [...] would not have had to endure; *but* she does not seem embittered

Syntax and discourse classes, and syntax/discourse structure change type: The whole linguistic unit (phrase, clause or sentence) suffering the syntactic or discourse reorganisation is tagged. Moreover, most syntax and discourse based changes have a key element that gives rise to the change and/or distinguishes it from others. This key element was also tagged.

In (29), the coordination change affects two clauses in (29-a) and two sentences in (29-b), so all of them constitute the scope of the phenomenon. The conjunction *and* stands for the key element.

(29) a. They are the sons of the same Father and are born *and* brought up with the same plan.
b. They were born of the same universal fact. They are of the same Father!

In the case of identical and non-paraphrases, no attributes LOCAL/GLOBAL nor key elements are used, and only the affected fragment is tagged.

The annotation process. The annotation process was carried out at the University of Barcelona by three postgraduate linguists experienced in annotation. CoCo (España-Bonet, Vila, Martí, and Rodríguez, 2009)[10] was the interface used for the annotation. The annotation was performed in three phases: annotators training, inter-annotator agreement (20% of the corpus) and final annotation (remaining 80%). The inter-annotator agreement was calculated taking into account the scope and types involved, obtaining a value of 0.62. We consider this to be an acceptable result, considering that a much simpler task, the binary decision of whether two text fragments are paraphrases or not on the MSRPC corpus, obtained an agreement of 0.84 Dolan and Brockett (2005). These results show the suitability of the paraphrase typology for the annotation of plagiarism examples.

Annotation Results. Paraphrase type frequencies, and total and average lengths are collected in Tables 8.1 and 8.2. Same polarity substitutions represent the most frequent paraphrase type ($freq_{rel} = 0.46$). At a considerable distance, the second most frequent type is addition/deletion ($freq_{rel} = 0.13$). We hypothesise that the way paraphrases were collected has a high impact on these results. They were hand-made asking people to re-write a collection of text fragments, i.e., they were originated in a reformulation framework, where a conscious reformulative intention by a speaker exists. Our hypothesis is that the most frequent paraphrase types in the P4P corpus stand for the paraphrase mechanisms most accessible to humans when asked to reformulate. Same polarity sub-

[10] http://www.lsi.upc.edu/~textmess/

Table 8.1: Paraphrase type absolute and relative frequencies. Note that the values in bold are in fact sums of the corresponding subtypes.

	$freq_a$	$freq_r$
Morphology-based changes	**631**	**0.057**
Inflectional changes	254	0.023
Modal verb changes	116	0.010
Derivational changes	261	0.024
Lexicon-based changes	**6,264**	**0.564**
Spelling and format changes	436	0.039
Same polarity substitutions	5,056	0.456
Synthetic/analytic subst.	658	0.059
Opposite polarity subst.	65	0.006
Inverse substitutions	33	0.003
Syntax-based changes	**1,045**	**0.094**
Diathesis alternations	128	0.012
Negation switching	33	0.003
Ellipsis	83	0.007
Coordination changes	188	0.017
Subord. and nesting changes	484	0.044

	$freq_a$	$freq_r$
Discourse-based changes	**501**	**0.045**
Punctuation and format changes	430	0.039
Direct/indirect style alternations	36	0.003
Sentence modality changes	35	0.003
Miscellaneous changes	**2,331**	**0.210**
Syntax/discourse structure changes	304	0.027
Change of order	556	0.050
Addition/deletion	1471	0.132
Semantics-based changes	**335**	**0.030**
Others	**136**	**0.012**
Identical	101	0.009
Non paraphrases	35	0.003

stitutions and addition/deletion are mechanisms which are relatively simple to apply to a text by humans: changing one lexical unit for it synonym (understanding synonymy in a general sense) and deleting a text fragment, respectively. The above mechanisms could be equally applied to other kinds of text re-use.

It is interesting to note that, in general terms, $len_{src} > len_{plg}$, which means that, while reformulating, people tend to summarise, use shorter expressions for same meaning, or, as already said, just delete some fragments. Clearly, the different types of paraphrase mechanisms tend to be used to summarise the re-used contents. Finally, the paraphrase types with a largest average length are syntax and discourse based changes. The reason has to be found in the above distinction between the two ways to annotate the scope: in structural reorganisations, we annotate the whole linguistic unit suffering the change.

8.3 Analysis of Paraphrase Plagiarism Detection

As already said, paraphrase plagiarism has been identified as an open issue in plagiarism detection (Potthast *et al.*, 2010d; Stein *et al.*, 2011a). In order to figure out the limitations of actual plagiarism detectors when dealing with paraphrase plagiarism, we analyse their performance over the P4P corpus. Our aim is to understand what types of paraphrases make plagiarism more difficult to be detected.

Firstly, we analyse the detectors' performance when considering the entire PAN-PC-10 (cf. Section 7.2.2). The aim is giving a general perspective of how difficult detecting cases with a high paraphrase density is. Secondly, we analyse the detectors' performance when considering different partitions of the P4P corpus. We do so in order to identify those (combinations of) paraphrase operations that better cause a plagiarised fragment to go unnoticed. This analysis opens the perspective to research directions in automatic

Table 8.2: Paraphrase type total and average lengths (lengths $\pm\sigma$). On top the lengths corresponding to the entire source and plagiarised fragments.

	tot_{src}	tot_{plg}	avg_{src}	avg_{plg}
Entire fragments	210,311	193,715	248.30±14.41	228.71±37.50
Morphology-based changes				
Inflectional changes	1,739	1,655	6.85±3.54	6.52±2.82
Modal verb changes	1,272	1,212	10.97±6.37	10.45±5.80
Derivational changes	2,017	2,012	7.73±2.65	7.71±2.66
Lexicon-based changes				
Spelling and format changes	3,360	3,146	7.71±5.69	7.22±5.68
Same polarity substitutions	42,984	41,497	8.50±6.01	8.21±5.24
Synthetic/analytic substitutions	12,389	11,019	18.83±12.78	16.75±12.10
Opposite polarity substitutions	888	845	13.66±8.67	13.00±6.86
Inverse substitutions	417	314	12.64±8.82	9.52±5.93
Syntax-based changes				
Diathesis alternations	8,959	8,247	69.99±45.28	64.43±37.62
Negation switching	2,022	1,864	61.27±39.84	56.48±38.98
Ellipsis	4,866	4,485	58.63±45.68	54.04±42.34
Coordination changes	25,363	23,272	134.91±76.51	123.79±71.95
Subordination and nesting changes	48,764	45,219	100.75±69.53	93.43±60.35
Discourse-based changes				
Punctuation and format changes	51,961	46,894	120.84±79.04	109.06±68.61
Direct/indirect style alternations	3,429	3,217	95.25±54.86	89.36±50.86
Sentence modality changes	3,220	2,880	92.0±67.14	82.29±57.99
Miscellaneous changes				
Syntax/discourse structure changes	27,536	25,504	90.58±64.67	83.89±56.57
Change of order	15,725	14,406	28.28±30.89	25.91±24.65
Addition/deletion	16,132	6,919	10.97±17.10	4.70±10.79
Semantics-based changes	16,811	13,467	50.18±41.85	40.20±29.36
Others				
Identical	6,297	6,313	62.35±63.54	62.50±63.60
Non paraphrases	1,440	1,406	41.14±26.49	40.17±24.11

plagiarism detection that aims at detecting these kinds of borrowing. For our second level analysis, we first divide the cases of plagiarism in the P4P corpus according to the occurring paraphrase phenomena (Section 8.3.1). In Section 8.3.2, the results and discussion for both levels of analysis are set out.

8.3.1 Clustering Similar Cases of Plagiarism in the P4P Corpus

Paraphrase annotation and plagiarism detection are performed at different levels of granularity: paraphrase phenomenon's scopes go from word to (multiple-)sentence level and plagiarism detectors aim at detecting entire, in general long, plagiarised fragments and their source. Thus, data should be organised in a way that makes them comparable. We decided not to compare paraphrase annotation and detectors' results directly, but grouping together those cases of plagiarism in the P4P corpus with a similar distribution

of paraphrase operations or where a paraphrase type stands out from the rest. These groupings make granularity differences irrelevant. In order to perform this process, we used k-means (MacQueen, 1967), a popular clustering method. In brief, k-means performs as follows: (*i*) k, the number of clusters, is set up at the beginning, (*ii*) k points are selected as initial centroids of the corresponding clusters, for instance, by randomly selecting k samples, and (*iii*) the position of the centres and the members of each cluster are iteratively redefined to maximise the similarity among the members of a cluster (intra-cluster) and minimise the similarity among elements of different clusters (extra-cluster).

We composed a vector of 22 features that characterises each plagiarism case. Each feature corresponds to one paraphrase tag in our annotation, and its weight is the relative frequency of the type in the plagiarism case.[11] As same polarity substitutions occur so often in many different plagiarism cases (this type represents more than 45% of the paraphrase operations in the P4P corpus and 96% of the plagiarism cases include at least one same polarity substitution), they do not represent a good discriminant factor for grouping together plagiarism cases with a similar distribution of paraphrase operations or where a paraphrase type stands out from the rest. This was experimentally confirmed by a preliminary experiment we carried out considering different values of k. Therefore, k-means was applied by considering 21 features only. For every value of k we carried out 100 clustering proceedings with different random initialisations and considering $k = [2, 3, \ldots 20]$. Our aim was twofold: (*i*) obtaining the best possible clusters for every value of k (i.e., with highest intra-cluster similarity and lowest extra-cluster similarity) and (*ii*) determining the number of clusters to better organise the paraphrase plagiarism cases. In order to determine a convenient value of k, we applied the *elbow method* (cf. Ketchen and Shook (1996)), which calculates the clusters' distortion evolution (also known as cost function) for different values of k. The inflection point, i.e., "the elbow", was in $k = 6$.

On the basis of our findings, we identified a set of paraphrase types to characterise the clusters. We describe the obtained results in the clusters that show the most interesting insights from the perspective of the paraphrase cases of plagiarism. A summary is included in Fig. 8.2. Although same polarity substitutions are not taken into account in the final clustering, they remain in the plagiarism cases, and their numbers are displayed. They are similarly distributed among all the obtained clusters, and remain the most frequent in all of them. In terms of linguistic complexity, identical and semantics-based changes can be considered as the extremes of the paraphrase continuum: absolute identity in the form and complete change in the form, respectively. In c_5 and c_2, identical and semantic, respectively, are the most frequent type (after same polarity substitutions) and more frequent than in the other clusters.[12] The most common type in c_3 is spelling and format. We observed that 39.36% of the cases imply only case changes which can be easily mapped to identical by a case folding process. In the other clusters no relevant features are observed.

[11] Note that the lengths of the different paraphrase chunks could have been considered as well. Nevertheless, paraphrases are annotated at different levels of granularity (e.g. lexicon-based changes versus discourse-based changes), preventing from being proper features.

[12] Identical and semantic fragments are also longer in the respective clusters than in the others.

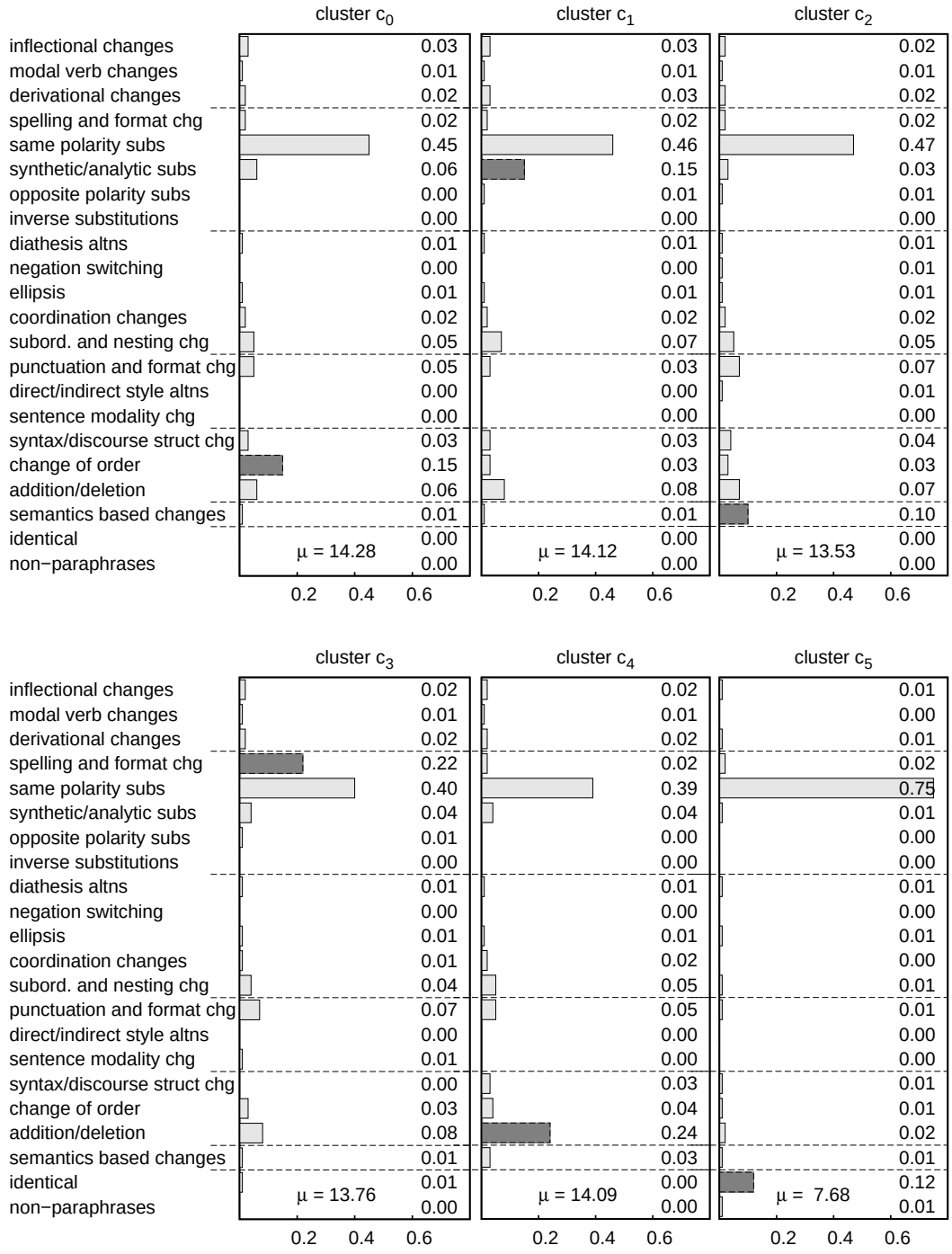


Figure 8.2: Average relative frequency of the different paraphrase phenomena in the plagiarism cases of each cluster. The feature that stands out (if we do not consider the non discriminative same polarity substitutions), also respect to the rest of clusters, is represented darker. The value of μ refers to the average absolute number of phenomena per case in each cluster. (chg→changes, altn→alternations, subs→substitutions, struct→structure.)

In terms of quantitative complexity, we regard at the amount of paraphrase phenomena occurring in the plagiarism cases. It follows that c_0 contains the cases with least phenomena on average. The remaining clusters have a similar number of phenomena.

8.3.2 Results and Discussion

Our in-depth analysis considers the evaluation measures of PAN (cf. Section 4.3.3). Due to our interest in investigating the amount of paraphrase cases that state of the art systems for plagiarism detection succeed to detect, we pay special attention to recall.

In order to understand the results obtained over the P4P corpus, it is worth recalling those obtained by the participants with the entire corpus (cf. Figure 7.5 in page 177). As we may observe there, the best recall values are above 0.60, with very good values of precision, some of them above 0.90. The results when considering manually paraphrased plagiarism were presented in Fig. 7.6 (page 178). It can be observed that, in the most of the cases, the quality of the detections decreases dramatically with respect to the entire corpus, which contains also translated, verbatim and automatically modified plagiarism. As we already observed, manually created cases seem to be much harder to detect than the other, artificially generated, cases.[13] The difficulty to detect simulated cases of plagiarism in the PAN-PC-10 corpus was stressed already by Stein *et al.* (2011a). This result does not necessarily imply that automatically generated cases were very easy to detect. When the simulated cases of the PAN-PC-10 corpus were generated, volunteers had specific instructions of creating rewritings with a high obfuscation level. This fact may have caused these cases to be even more difficult to detect than expected when analysing real documents for paraphrase plagiarism. It is worth noting that, in a few cases the plagdet level does not have an important variation between the overall and simulated cases. For instance, (Micol *et al.*, 2010) results are downgraded from 0.22 to 0.18 only. However, both precision and recall do experience an important decrease, but granularity is much improved, hence causing a lower punishment effect. The case of (Nawab *et al.*, 2010) is interesting as well. Their overall plagdet increases from 0.21 to 0.26. In this case, both precision and recall experience an improvement, as well as granularity. The fact is that both (Micol *et al.*, 2010) and (Nawab *et al.*, 2010) apply a very flexible retrieval strategy, based on single words. The detailed analysis is based on the search of the largest common sub-strings between d_q and d. None of them apply any special pre-processing, other than case folding or discarding tokens or non-alphanumeric strings.

Figure 8.3 shows the evaluation results when considering the P4P corpus only. All the detectors obtain a nearly perfect granularity. In most cases, this value does not mean that the detectors are better to define the limits of the plagiarism cases in this partition. The reason is that many of them are unable to detect (most of) them.[14] In some cases,

[13] This can be appreciated when looking at the difference in performance of the system applied at both competitions by Grozea *et al.* (2009) and Grozea and Popescu (2010a), practically the same implementation. At the first competition, which included artificial cases only, its recall was of 0.66 while in the second one, with simulated (i.e., paraphrase) cases, it decreased to 0.48.

[14] As a result, in the rest of figures of this section we do not display granularity and plagdet becomes practically equals to F-measure.

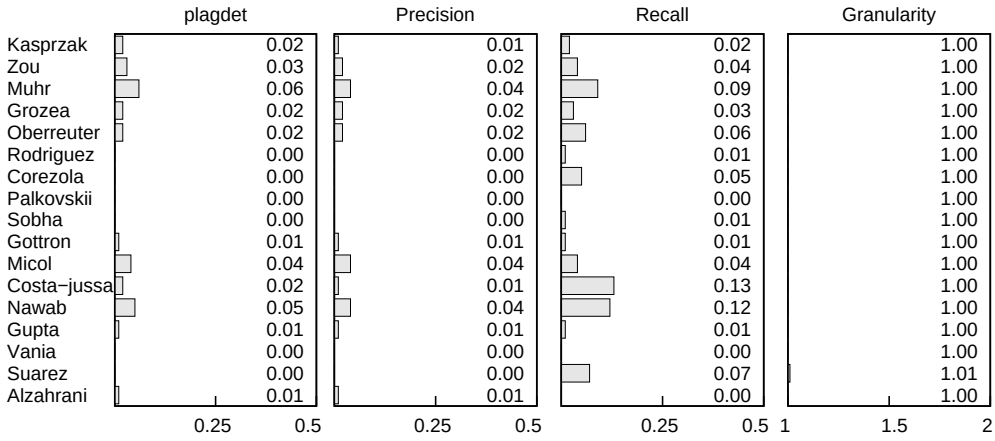


Figure 8.3: Evaluation of PAN 2010 participants' plagiarism detectors over the P4P corpus.

such as that of Rodríguez Torrejón and Martín Ramos (2010), this can be justified by the heuristics applied during post-processing, which include discarding cases which are shorter than a given threshold (which is very high). Note that beside the difficulty that paraphrasing implies to plagiarism detection, the shorter a plagiarised case is, the harder it is to detect, and the P4P corpus is composed precisely of the shortest cases of paraphrase plagiarism in the PAN-PC-10.

Plagdet, precision, and recall are measured for the partitions of the P4P corpus identified by the 6 clusters obtained in Section 8.3.1. Firstly we compare the results obtained over the extreme cases: c_5 versus c_2. Cluster c_5, which comprises the lowest linguistic and quantitative complexity, is the one containing the best detected plagiarism cases. Cluster c_2, which comprises the highest linguistic complexity, is the one containing the worst detected plagiarism cases. Cluster c_3 is somewhere in between c_5 and c_2 as the high presence of spelling and format changes (most of which are similar to identical cases), causes a plagiarism detector to have relatively more success on detecting them. These results are clearly observed through the values of recall obtained by the different detectors. Moreover, a correlation between recall and precision exists (in general terms, high values of recall come with higher values of precision). As can be seen, there exists a correlation between linguistic and quantitative complexity and performance of the plagiarism detection systems: more complexity implies worse performance of the systems.

Interestingly, the best performing plagiarism detection systems on simulated plagiarism are not the ones that perform the best at the PAN-10 competition. By still considering recall only, the best approaches on the P4P corpus, those of R. Costa-jussà *et al.* (2010) and Nawab *et al.* (2010) are far from the top detectors (Figure 8.3). On the one hand, Nawab *et al.* (2010) applies greedy string tiling, which aims at detecting as long as possible identical fragments. As a result, this approach clearly outperforms the rest of detectors when dealing with cases with a high density of identical fragments (c_5 in Figure 8.5). On the other hand, R. Costa-jussà *et al.* (2010) outperforms the rest of detectors when dealing with the cases in the rest of clusters (Figures 8.4 and 8.5). The reasons are twofold: (*i*) their pre-processing strategy (that includes case folding, stop-

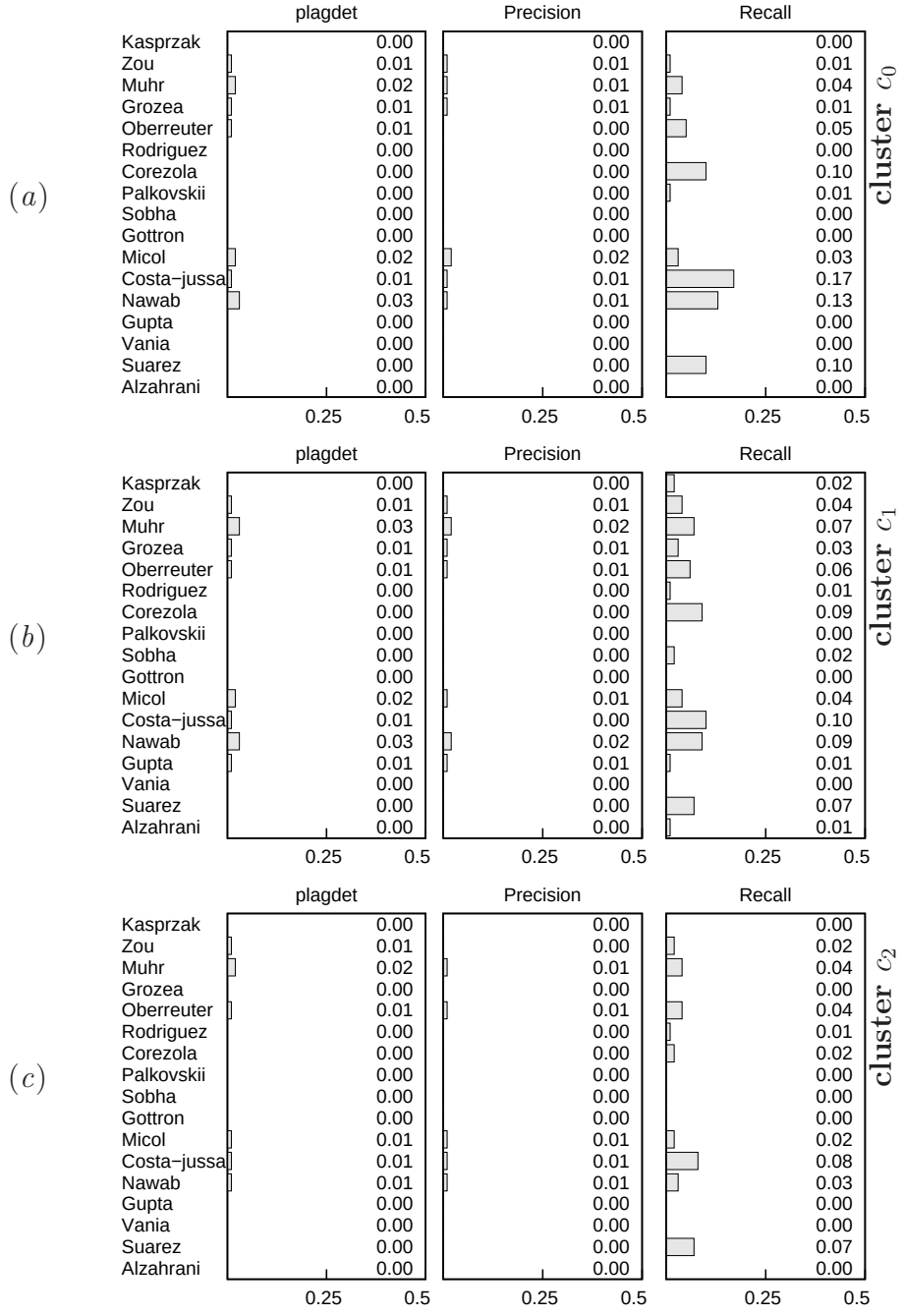


Figure 8.4: Evaluation PAN 2010 participants' plagiarism detectors for clusters (*a*) c_0; (*b*) c_1; (*c*) c_2.

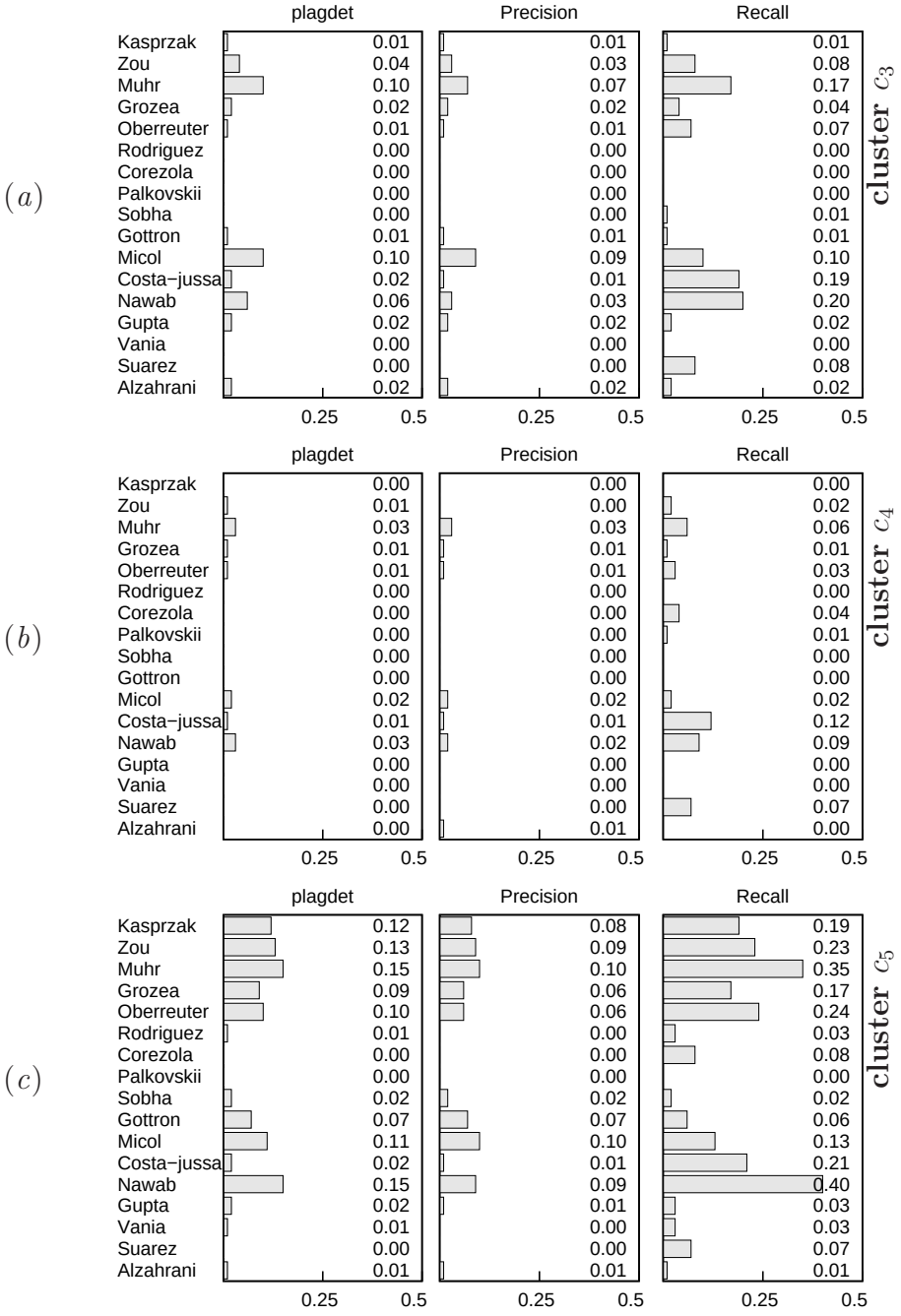


Figure 8.5: Evaluation PAN 2010 participants' plagiarism detectors for clusters (*a*) c_3; (*b*) c_4; (*c*) c_5.

word removal, stemming) looks at minimising the differences in the form caused by some paraphrase operations; (*ii*) their dot-plot-based technique (considering isolated words) is flexible enough to identify fragments that share some identical words only. Cluster c_3 is again somewhere in between c_5 and c_2. The results by Nawab *et al.* (2010) and R. Costa-jussà *et al.* (2010) are very similar in this case. The former shows a slightly better performance for the reasons already exposed: the high overlapping between the most frequent cases in clusters c_5 and c_3 (identical and spelling and format changes).

Now we compare the results obtained by R. Costa-jussà *et al.* (2010) to those of Grozea and Popescu (2010a). Regardless their comparison techniques are very similar (both based on dot-plot), the latter does not apply any pre-processing, hence ignoring any paraphrase operation. Moreover, their comparison unit is less flexible (character 16-grams versus word 1-grams).

8.4 Chapter Summary

Paraphrasing is the linguistic mechanism many text re-use and plagiarism cases rely on. Nevertheless, as discussed in the previous chapters, detectors do not address this issue. In this chapter, we aimed at not only investigating how difficult detecting paraphrase cases for state of the art plagiarism detectors is, but, especially, in understanding which types of paraphrases underlay text re-use acts and which are the most difficult to be detected.

In order to do that, we tagged a subset of the PAN-PC-10 corpus with the types of a new paraphrase typology. The obtained P4P corpus represents the only collection of plagiarism cases manually annotated at paraphrase level. We consider that this resource will allow the research community to investigate further the relationship between plagiarism and paraphrasing.

The annotated cases were mapped to those within the PAN-PC-10 corpus, allowing for an in-depth analysis of the approaches applied during the Second International Competition on Plagiarism Detection when dealing with these cases of paraphrase plagiarism. We observed that there exists a correlation between linguistic (i.e., kind of paraphrasing) and quantitative (i.e., amount of paraphrases) complexity and performance of the plagiarism detectors: more complexity implies worse performance of the systems.

We identified that lexical substitutions are the paraphrase mechanisms used the most when plagiarising. Moreover, all the paraphrase mechanisms tend to be used to produce a summarised version of the re-used text. These insights should be taken into account when developing the future generation of plagiarism detection systems in order to allow them to detect cases of plagiarism with a high level of paraphrasing.

Related publications:

- Barrón-Cedeño, Vila, and Rosso (2010b)

Chapter 9

Detection of Text Re-Use in Wikipedia

> Wikipedia is free content that anyone can edit, use, modify, and distribute.
>
> Wikipedia's third pillar

Wikipedia, the Free Encyclopedia, is identified as a Web-based collaborative authoring environment (Brandes and Lerner, 2007) in which articles, on a wide range of topics, are written worldwide in many different languages. Some people even consider its edition environment as *participatory journalism* (Lih, 2004). Editions of Wikipedia have been created in over 250 different languages. Some of them are more evolved than others, and by July 2011 it was estimated that it contained circa 19 million articles, each of which has been written, and re-written, by volunteers all around the world (Wikipedia, 2011p). Wikipedia has deserved the attention of researchers from a wide variety of environments.

A brief overview of the most related works to text re-use analysis over Wikipedia is given in Section 9.1. Taking advantage of its nature, in this chapter we focus on re-use inside and from Wikipedia. The study is carried out from both monolingual and cross-language perspectives. In the case of monolingual co-derivation, the dynamic authoring schema that this wiki represents is analysed. A small sample of the languages represented in Wikipedia is considered and revisions of a few topics are analysed. Whereas four different languages are considered, the analyses are performed independently, in a monolingual approach. Our findings on monolingual co-derivation in Wikipedia are described in Section 9.2. The cross-language analysis is divided in two parts. Firstly, we measure the cross-language relationships between editions of Wikipedia in ten different languages. Our aim is determining the extent of comparability between the different languages editions and comparing the expressiveness of a set of cross-language similarity models over Wikipedia articles. Our findings are discussed in Section 9.3. Secondly, we aim at analysing the cross-language co-derivation phenomenon, which seems to occur very often within Wikipedia: when the contents of an article a are re-used when editing an article a' in another language (in many cases a and a' are comparable, i.e., they cover the same topic in the two different languages). Our research on this phenomenon is described in Section 9.4.

The chapter closes with the discussion of an international challenge on cross-language text re-use detection from Wikipedia articles we recently organised. We simulated a framework where Wikipedia articles (in English) were used as the source for the generation of texts (in Hindi). The considered languages are nearly explored from the text re-use perspective. This is a common problem in different environments, such as academic plagiarism. The discussion is presented in Section 9.5.

Key contributions One of the main outcomes of this chapter is the estimation of similarity levels across editions of Wikipedia in different languages, which is useful for the selection of article pairs that could be exploited for enriching the resources necessary for CL-ASA (Section 9.3). The other outcome is the generation of an evaluation framework for the analysis of cross-language text re-use from Wikipedia articles in English to (simulated) instances in Hindi (Section 9.5).

9.1 Related Work over Wikipedia

Attention over Wikipedia has increased in the last years due to its nature: a perpetual evolution schema, multilingualism, and freedom. From a technological point of view, Wikipedia has shown to be a valuable resource for different monolingual and cross-language NLP and IR tasks, such as mono- and multilingual named entity recognition (NER) (Richman, 2008; Toral and Muñoz, 2006), query translation for CLIR (Gaillard, Boualem, and Collin, 2010; Nguyen, Overwijk, Hauff, Trieschnigg, Hiemstra, and de Jong, 2009), word sense disambiguation (WSD) (Mihalcea, 2007), near-duplicates detection (Potthast *et al.*, 2008a), and quality flaw prediction in Wikipedia.[1] However, little work has analysed text re-use within Wikipedia.[2]

9.1.1 Monolingual Analysis

Wikipedia has received a lot of attention considering it as a monolingual (in most cases English) corpus. For instance, Brandes and Lerner (2007) tried to cluster similar Wikipedia articles. Rather than considering the articles' contents, topical similarity was estimated on the basis of authorship and edition time intervals. The former feature measures the intersection among the authors of different articles' text. The latter has to do with eventual similarity, which can be measured on the basis of frequency of editions in similar time.

A quantitative analysis of Wikipedia was performed by Ortega Soto (2009). One of the outcomes of this research is the WikiXRay tool[3]. WikiXRay is a combination of Python and *R* scripts for automatically processing Wikipedia dumps. By performing

[1] At PAN 2010 and 2011, a task on Wikipedia vandalism detection was organised (Potthast and Holfeld, 2011; Potthast *et al.*, 2010c). In 2012 the task is on quality flaw prediction; e.g. poor writing style, unreferenced statements, or missing neutrality (cf. `pan.webis.de`).

[2] Refer to the Wikipedia article *Academic studies of Wikipedia* (`http://bit.ly/wikipedia_studies`) as well as the Zotero group *Wikipedia research* (`http://bit.ly/zotero_wikipedia`) for more information.

[3] `http://felipeortega.net/WikiXray`

different empirical (monolingual) analyses over ten of the most developed editions of Wikipedia, Ortega Soto tried to provide answers to different questions such as how the community of editors evolves and how the amount of information increases. He even tried to determine the reputation of an author on the basis of the articles she had participated in and are considered as featured.

Another problem that has received attention is that of automatically assessing the trustworthiness of an article. Zeng, Alhossaini, Fikes, and McGuinness (2006) performed a document-level analysis on the basis of an article's citation frequency. However, they noted that considering the revision history (including the trustworthiness of the previous revision, the author of the last revision, and the amount of text involved in the last revision) was a better option. While their initial approach considered entire articles, their new proposal goes in a deeper granularity: sentences (they claim that the semantics of a revision may be interpreted at a sentence level for the most part). The sentence comparison is computed considering the longest common subsequence algorithm and the analysis is carried out on the basis of a Bayesian network. Zeng *et al.* (2006) found a clear difference in the trustworthiness of featured, clean-up, and normal articles,[4]which reflects the usability of the model. Adler, Chatterjee, and de Alfaro (2008) went further and estimated the level of trust of every word w in an article. The decision is based on three main factors: (*i*) the extent of time w has remained in the article; (*ii*) the reputation of the author that inserted w; and (*iii*) the reputation of the following authors (which preserved w after review). By considering these and other features, they developed Wikitrust[5], a software that assess the trustworthiness of Wikipedia articles. This software has been successfully used in the PAN International Competition on Wikipedia Vandalism Detection (Adler, de Alfaro, Mola-Velasco, Rosso, and West, 2010; Potthast *et al.*, 2010c).

The work of Nelken and Yamangil (2008) is slightly closer to ours (cf. Section 9.2). Nelken and Yamangil used the differences between adjacent revisions of an article in order to support three tasks. (*i*) Automatic discovery of *eggcorns* (unusual, often wrong, spellings such as <funder, founder>). These tokens are retrieved on the basis of the edit distance between substituted words in neighbour revisions, which could be simple corrections of each other. (*ii*) Sentences compression, i.e., reducing sentences to the maximum level but maintaining the idea. They compare the editions at the sentence level to see how the different sentences evolve; (*iii*) Extractive text summarisation on the basis of articles' history. The intuitive idea is that the longer a sentence has survived across revisions, the more relevant it is and it is worth considering it for the summary.

9.1.2 Analysis across Languages

Ziggurat is a system that performs *information arbitrage* (Adar, Skinner, and Weld, 2009). However, it does not consider free text; instead, articles' *infoboxes* in different

[4]According to Wikipedia, featured articles are those that accomplish with the encyclopedia's criteria of accuracy, neutrality, completeness, and style; clean-up articles are those that have been identified as requiring major revisions because *flaws* are identified regarding spelling, grammar, typographical errors, sourcing, etc. A competition on Quality Flaw Prediction is organised at PAN 2012.

[5]`http://www.wikitrust.net`

languages are compared in order to detect inefficiencies (missing, inaccurate, or wrong information). The considered languages include English, French, German, and Spanish. As infoboxes include very concise information, a classifier is used considering the following features: (*i*) word level similarity (which could capture some common named entities), (*ii*) character 3-grams similarity (CL-C3G), (*iii*) correlation ("translation") of numerical values (km→hect, population, etc.), (*iv*) translations similarity (every word is represented by every possible translation in the other language in order to find potential matches,[6] and (*v*) structural level similarity.

The last feature is particularly interesting: the outlinks in the different fields are considered. As a Wikipedia article should point to articles in the same language only, a mapping between comparable articles has to be obtained beforehand. That is, if an article in L links to an article a_L and an article in L' links to b'_L and a and b' happen to be comparable (i.e., the corresponding articles covering the same topic in the respective language), the corresponding texts are considered to be more similar. As links could point to hypernyms, the first paragraph of the linked article is used to determine similarity. Comparisons based on Wikipedia outlinks have shown encouraging results on the extraction of similar sentences across languages as well (Adafre and de Rijke, 2006). Both Adafre and de Rijke (2006) and Adar *et al.* (2009) consider not only the target links but also the content of the target articles to deal with ambiguities and semantic relationships. In our work (cf. Section 9.3) we simply create a vector representation composed of outlinks in the documents that are mapped to another language using the structure of Wikipedia.

Several research works focus on detecting similar sentences across comparable documents to build parallel corpora for statistical MT. For instance, Munteanu *et al.* (2004) consider newspapers; Mohammadi and GhasemAghaee (2010) and Yasuda and Sumita (2008) consider Wikipedia. Patry and Langlais (2011) focus on a model for detecting parallel documents within a cross-language collection. They propose a classifier that considers hapax-legomena, numerical entities, and a set of punctuation marks. As a result, their strategy can be applied to many language pairs.

9.2 Monolingual Co-Derivation in Wikipedia

Due to Wikipedia's collaborative environment, an article a is the result of multiple editions through time T.[7] When a volunteer aims at editing an article a she is in fact modifying the product of a long co-derivation process. In an extreme case, the text in a_T might be the result of the contribution of T different authors. In this framework, a_1 represents the rise of the article and a new revision a_t could imply the deletion, addition or paraphrasing of its contents.

As discussed in Chapter 3, when analysing a set of documents for text re-use and, in particular, co-derivation, considering a good similarity model is a key factor. Therefore,

[6] The dictionary used is that of Etzioni, Reiter, Soderland, and Sammer (2007).

[7] For this analysis we ignore those cases of vandal modifications of Wikipedia contents, which use to be reverted. Cf. Potthast and Holfeld (2011); Potthast *et al.* (2010c) for an overview on automatic detection of this kind of "contribution".

we carried out an exhaustive comparison of similarity estimation models in order to study which one performed better on different scales of text and languages. We considered both entire articles and sections. Four different languages were analysed: English, German, Spanish, and Hindi (en, de, es, hi). It is worth noting that, while considering multiple languages, in this section the analysis is carried out independently for each language, on a monolingual setting.

English represents the language with more representation on the Web. This is reflected by the fact that the English Wikipedia contains more than 3.5 million articles.[8] This amount is only comparable to the four languages with most Wikipedia articles after English: German, French, Italian, and Polish. Altogether, they just surpass the 4 million articles. English is a Germanic language with influences such as Latin and Norman. German is the second language with the largest representation in Wikipedia (over 1.2 million articles, comparable to French). It is a more difficult language to work with than English, mainly because of its inflecting nature. Spanish is a Romance language with middle-high representation in Wikipedia (over 0.8 million articles, comparable to Italian and Polish). Finally, Hindi represents a language with relatively little representation in the encyclopedia (over 90 thousand articles). By contrast, it is one of the most widely spoken languages in the world. While the other three languages have a Latin-based alphabet, Hindi's is *Devanagari*. This fact aimed at representing a bigger challenge for the text representations considered.

9.2.1 Experimental Settings

For this experiments we use the co-derivatives corpus, described in Section 4.2.2. For our experiments it is worth considering the curves shown in Figure 4.1 (page 86): the difference in similarity trends for different Wikipedia's articles. While the English Wikipedia is the best established one, the Spanish and German ones are looking for such a status. At the same time, the Hindi Wikipedia can be considered still in its childhood.[9] A common pre-processing for all of the languages was carried out: (*i*) space normalisation, (*ii*) sentence detection, (*iii*) tokenisation, and (*iv*) case folding. As before mentioned, stemming is another common pre-processing step that uses to improve the results in other tasks. Nevertheless, it does not seem to be the case for text re-use and co-derivatives detection. Previous experiments show that in these tasks stemming does not improve the results importantly (Barrón-Cedeño and Rosso, 2009b; Hoad and Zobel, 2003).

The approached problem is the detection of co-derivatives of a given query text. In order to detect such a co-derivation relationship, we propose an IR exercise. Let d_q, the last revision of an article, be the query document. Let D be the entire collection of articles in the same language (including d_q). Retrieve those documents $d \in D$ that are co-derivatives, in our specific setting, revisions of d_q. The best co-derivative candidates would be those texts $d \in D$ which maximise $sim(d_q, d)$. We define r_q to be a ranking of documents, sorted in descending order with respect to the computed similarity values.

[8]The figures here mentioned were obtained from `http://stats.wikimedia.org`, where statistics about different Wikimedia projects are available. They are as for July 2011.

[9]The analysis and comparison of maturity levels among the different Wikipedia editions represents an interesting topic for further research.

Table 9.1: Co-derivatives corpus statistics at section level. The column headers stand for: D'_q collection of query-sections, D^* sections of all documents in D, $|d^*_{avg}|_t$ average number of types per section, $|d^*_{avg}|$ average number of tokens per section, $|D^*|_t$ types in D^*.

Language	$\|D'_q\|$	$\|D^*\|$	$\|d^*_{avg}\|_t$	$\|d^*_{avg}\|$	$\|D^*\|_t$
	before stopwords elimination				
de	7726	133,171	124	198	261,370
en	8043	114,216	187	378	183,414
es	4696	86,092	126	241	133,595
hi	345	27,127	76	125	78,673
	after stopwords elimination				
de	7726	133,171	98	132	261,146
en	8043	114,196	159	266	183,288
es	4696	86,076	103	142	133,339
hi	345	27,125	64	92	78,577

As co-derivation is not necessarily a phenomenon over entire documents, we performed two independently evaluated stages: (*i*) document level analysis and (*ii*) section level analysis.[10]

For our comparison, a total of eight monolingual similarity measures from three families were considered: (*i*) a Boolean vector space model: the Jaccard coefficient; (*ii*) real-valued vector space models: the cosine similarity measure, Okapi BM25, and word chunking overlap; (*iii*) probabilistic models: the Kullback-Leibler distance and a monolingual machine translation model (a kind of monolingual ASA, cf. Section 6.3). These similarity measures are all described in Section 3.3. As terms, we considered the simple BoW model (cf. Section 3.1.2). Additionally, two well known plagiarism detection models were explored: Winnowing and SPEX (cf. Section 5.1.2.1).

Experiment 1: Document level analysis. The aim is retrieving all the articles $d \in D$ co-derived from d_q, i.e., ten documents. For each document d_q, **in total 500**, the documents in D are ranked with respect to their similarity $sim(d, d_q)$. The top-k documents in the ranked list are retrieved, composing the collection r^*_q ($k = 50$). It is expected that the documents on top of r^*_q will be those co-derived from d_q; r^*_q composes the input collection for the section level analysis.

Experiment 2: Section level analysis. Documents in r^*_q are split into sections, composing the collection D^* of co-derivative candidate sections. Likewise, d_q is composed of the corresponding sections in d_q. For each section in $s_q \in d_q$ the sections in $s \in D^*$ are ranked with respect to their similarity $sim(s, s_q)$, generating r^*_q. Again, it is expected that those sections on top of r^*_q are actual co-derivatives of s_q. Statistics of the articles' sections used in these experiments are included in Table 9.1. The difference in the amount of sections among the languages is worth noting.

In order to perform an objective evaluation framework at this stage, two factors are assumed: (*a*) D^*_q is composed only of those sections of d_q which have been equally named in the corresponding 10 revisions; and (*b*) in case a co-derivative of d_q is missing from r_q*, it is included. Evidently, these two factors are not realistic, but try to allow for an evaluation of the similarity measures over sections in optimal conditions. The former factor allows controlling the text fragments we are interested in retrieving. The latter

[10]In order to detect borders between sections, we took advantage of the Wikipedia articles' inherent structure, where the different sections are explicitly tagged.

factor makes possible retrieving every co-derived section, as far as the similarity model achieves it.

9.2.2 Results and Discussion

In order to evaluate how well the co-derivative detection process performs, we opted for considering recall and precision (cf. Section 4.3.1) and two measures specially designed for evaluating documents versions retrieval: HFM and *sep* (cf. Section 4.3.2). $|r_q|$ is considered to be $k = \{10, 20, 50\}$. For the case of recall, the three values of k are considered, i.e., *rec*@10, *rec*@20, and *rec*@50. For the case of precision, only $k = 10$ is considered, i.e., *prec*@10. This decision is backed in the fact that only ten relevant documents exist for a given query. Note that, as the number of relevant and retrieved documents from r_q are the same, *prec*@10 = *rec*@10. HFM and *sep* are calculated by considering $k = 50$. As previously noted (cf. page 102), these measures require all relevant documents for d_q to be included in r_q^*. Hence, they are computed in those cases where $R@50 = 1.0$ only. The results of both experiments are displayed in Fig. 9.1 for the four languages.

Experiment 1: Document level analysis. In most cases, the values of *rec*@10 are quite similar to those of *rec*@20 and *rec*@50. This means that the relevant documents, the revisions of d_q, are concentrated on the top-10 positions. For the experiments at document level the different models obtained very similar results for English, German, and Spanish. The only exception appears when using Okapi BM25. The reason behind this behaviour is that this method is actually designed for keyword based retrieval. Even by trying to tune the implied parameters, the results were not comparable to those obtained by the other methods.

By comparing the results of the four languages it might be erroneously considered that the retrieval of documents in Spanish and Hindi is more complicated than for English and German. However, this is not necessarily true. The reason for the worse results is in fact justified by Fig. 4.1 (page 86). A more drastic decrease in the actual similarity between revisions (a faster evolution trend), causes the retrieval exercise to be more complicated. In English the difference between $sim(d_q, d_1)$ and $sim(d_q, d_{10})$ —i.e., the first and last revision of an article—, is in average 0.23 only. For Hindi, at the other extreme, the difference is 0.72. We advocate that further investigation should be done on the process of discriminating co-derivatives from documents on the same topic.

Different similarity measures obtain comparable results in terms of recall. In order to analyse how similar the outcomes are, we performed a *Pearson's Chi-square test* (χ^2).[11] Table 9.2 (left hand side) summarises the obtained results for the four languages considering the eight similarity measures altogether and rec@10. Statistically significant difference exists considering all the measures, hence we performed a *post hoc* analysis. The results are displayed in Table 9.3. In general no statistically significant difference exists in the results obtained either with Jaccard, cosine, KL, Winnowing, or SPEX. The recall values for these five measures are in general high and very similar. As expected, when comparing these five measures to the other three significant difference exists, which

[11] cf. (Vaughan, 2001, pp. 75–88).

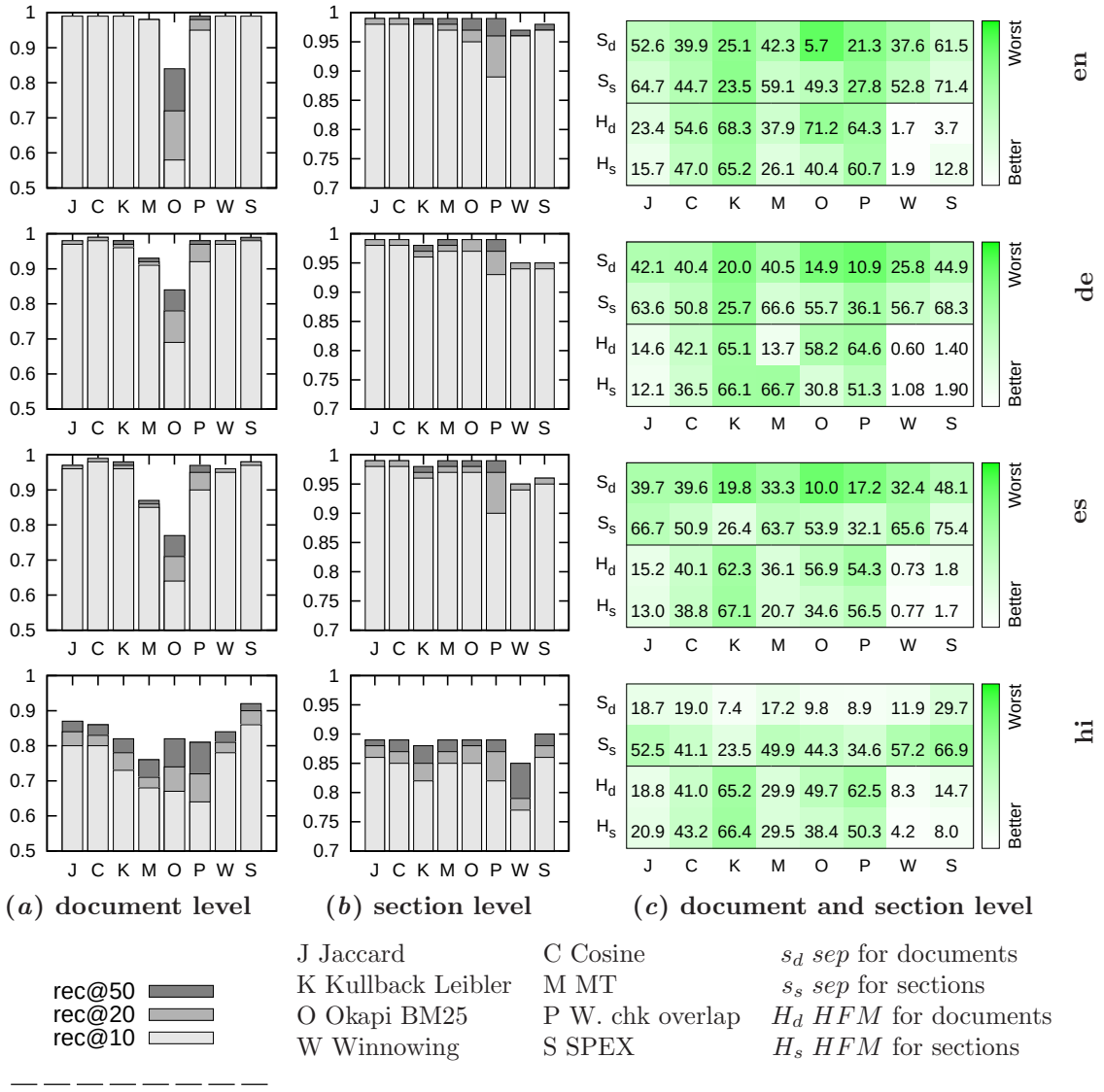


Figure 9.1: Co-derivatives results in terms of recall, HFM and sep. (a) and (b) show values of $rec@k$. (c) sep and HFM (for each square column the first and third row show sep/HFM at document-level $—_d—$ whereas the second and fourth rows show sep/HFM at section-level $—_s$).

is explained by the notably lower results.

In order to determine what measures are better, HFM and sep can be considered. At document level one thing results clear: the best approaches are those considering long representations, rather than BoW. As explored already during this document, this was an expected result. For the specific case of Winnowing, HFM is, in average, only 2.8%. For the case of SPEX the values are quite similar. The sep values obtained with these models are also the highest. This means that there is a clear border between relevant and irrelevant documents, letting for more accurate decisions. Again, this high precision

Table 9.2: Pearson's χ^2 test for the four languages and the eight similarity models considering rec@10. Results presented at document and section level; 7 represents the degrees of freedom($= |methods| - 1$), N is the number of retrieval experiments, and $p < 0.01$ stands for statistically significant difference.

	document level		section level	
de	$\chi^2(7, N = 500) = 466.82,$	$p < 0.01$	$\chi^2(7, N = 7726) = 33.98,$	$p < 0.01$
en	$\chi^2(7, N = 500) = 1138.715,$	$p < 0.01$	$\chi^2(7, N = 8043) = 83.33,$	$p < 0.01$
es	$\chi^2(7, N = 500) = 511.73,$	$p < 0.01$	$\chi^2(7, N = 4696) = 59.61,$	$p < 0.01$
hi	$\chi^2(7, N = 500) = 109.49,$	$p < 0.01$	$\chi^2(7, N = 345) = 23.95,$	$p < 0.01$

Table 9.3: Pairwise χ^2 test with Holm correction considering rec@10 at document and section level. The same nomenclature as in Fig. 9.1 is used. A black square indicates statistically significant difference between the pair ($p < 0.01$). A white square indicates non-statistically significant difference between the pair ($p > 0.01$).

	document level								section level								lan
	J	C	K	M	O	P	W		J	C	K	M	O	P	W		
C	□	-	-	-	-	-	-	C	□	-	-	-	-	-	-		
K	□	□	-	-	-	-	-	K	□	□	-	-	-	-	-		
M	■	■	□	-	-	-	-	M	□	□	□	-	-	-	-		de
O	■	■	■	■	-	-	-	O	□	□	□	□	-	-	-		
P	□	■	□	□	■	-	-	P	■	■	□	□	□	-	-		
W	□	□	□	■	■	□	-	W	□	□	□	□	□	□	-		
S	□	□	□	■	■	■	□	S	□	□	□	□	□	□	□		
	J	C	K	M	O	P	W		J	C	K	M	O	P	W		
C	□	-	-	-	-	-	-	C	□	-	-	-	-	-	-		
K	□	□	-	-	-	-	-	K	□	□	-	-	-	-	-		
M	□	□	□	-	-	-	-	M	□	□	□	-	-	-	-		en
O	■	■	■	■	-	-	-	O	□	□	□	□	-	-	-		
P	■	■	■	□	■	-	-	P	■	■	■	■	□	-	-		
W	□	□	□	□	■	■	-	W	□	□	□	□	□	■	-		
S	□	□	□	□	■	■	□	S	□	□	□	□	□	■	□		
	J	C	K	M	O	P	W		J	C	K	M	O	P	W		
C	□	-	-	-	-	-	-	C	□	-	-	-	-	-	-		
K	□	□	-	-	-	-	-	K	□	□	-	-	-	-	-		
M	■	■	■	-	-	-	-	M	□	□	□	-	-	-	-		es
O	■	■	■	■	-	-	-	O	□	□	□	□	-	-	-		
P	□	■	□	□	■	-	-	P	■	■	■	■	■	-	-		
W	□	□	□	■	■	□	-	W	□	□	□	□	□	□	-		
S	□	□	□	■	■	■	□	S	□	□	□	□	□	□	□		
	J	C	K	M	O	P	W		J	C	K	M	O	P	W		
C	□	-	-	-	-	-	-	C	□	-	-	-	-	-	-		
K	□	□	-	-	-	-	-	K	□	□	-	-	-	-	-		
M	■	■	□	-	-	-	-	M	□	□	□	-	-	-	-		hi
O	■	■	□	□	-	-	-	O	□	□	□	□	-	-	-		
P	■	■	□	□	□	-	-	P	□	□	□	□	□	-	-		
W	□	□	□	■	■	■	-	W	■	□	□	□	□	□	-		
S	□	□	■	■	■	■	□	S	□	□	□	□	□	□	■		

Table 9.4: Impact of stopwording in the co-derivatives experiments. *sw removal* reflects whether the best result was obtained when stopwords removal was applied. *Percentage* represents the percentage of experiments for which it was possible to estimate HFM and *sep* (i.e., $rec@50 = 1.0$) when processing documents / sections.

model	sw removal	percentage			
		en	de	es	hi
Jaccard coefficient	■	96 / 98	91 / 97	88 / 98	60 / 76
Cosine simiarity	■	97 / 98	96 / 97	95 / 98	63 / 76
Kullback Leibler		98 / 98	93 / 94	93 / 95	56 / 73
Machine Translation	■	94 / 97	77 / 95	65 / 96	37 / 74
Okapi BM25	■	79 / 98	79 / 97	69 / 98	62 / 77
Word chunking overlap	■	94 / 97	93 / 95	89 / 96	56 / 75
Winnowing		97 / 92	90 / 87	87 / 88	56 / 62
SPEX		96 / 95	82 / 92	80 / 92	48 / 61

comes at the cost of composing rigid comparison strategies, that may reduce the amount of relevant documents retrieved if further modified.

Experiment 2: Section level analysis. At section level, the supremacy of Winnowing and SPEX is not maintained any longer. The reason is simple: shorter documents are represented by fewer text chunks. As a result, very slight changes could prevent two similar —co-derived— sections to be retrieved. Another issue is that the input to this stage is a set of documents that are already highly related to the query (as not only the co-derived section, but all the rest in article a are there).

Whereas at document level Okapi BM25 performed worst, at section level it obtained comparable results to other vector and probabilistic models in terms of recall. Once again we performed a *Pearson's* χ^2 *test*. As observed in Table 9.2 (right hand side), statistically significant difference exists again considering all the measures. The results of the *post hoc* analysis are displayed in Table 9.3. In general, no difference exists between the different measures, except for the cases related to word chunking overlap, which offers the worst results. The recall values for the rest of measures are very similar. It is necessary to look at the HFM and *sep* values in order to figure out which performed best. Jaccard, Cosine and MT have practically the same quality in terms of *rec*, HFM and *sep*. Due to its simplicity, the Jaccard coefficient seems to be the best option (this is a result that we have seen over and over again!).

The experiments were carried out before and after stopwords removal. Table 9.4 specifies when the best result was obtained before or after stopwords removal (Table 9.1 only includes the best obtained results). The difference between whether removing stopwords or not is minimal in terms of recall. However, the percentage of comparisons in which it was possible to calculate HFM and *sep* specifies the amount of cases for which every relevant document was included among the top-50 documents in r_q. For entire documents cosine similarity and Kullback-Leibler distance performed better. At section level Okapi BM25, Jaccard and cosine represent the best options.

The results obtained in this set of experiments confirm the facts we had observed before: (i) in text re-use detection the frequency of a term is not so relevant as its

Table 9.5: Text pre-processing required for the three syntactic models. *tr*=translation, *cf*=case folding, *tk*=tokenisation, *wd*=stopwords deletion, *pd*=punctuation marks deletion, *bd*=blank space deletion, *sd*=symbols deletion, *dd*=diacritics deletion, *lm*=lemmatisation.

	tr	*cf*	*tk*	*wd*	*pd*	*bd*	*sd*	*dd*	*lm*
CL-C3G		■			■	■	■	■	
CL-COG		■	■				■	■	
T+MA	■	■	■	■					■

simple apparition is, (*ii*) considering long terms (for instance word n-grams with $n > 5$ causes to get a better separation between good and bad candidates, (*iii*) the certainty of retrieving only true positives comes at the cost of getting a high amount of false negatives, and vice versa (*iv*) considering flexible representations (BoW) gets very few false negatives at the cost of many true negatives.

9.3 Similarity of Wikipedia Articles across Languages

A subset of the topics in Wikipedia are available in multiple languages in a manner resembling a comparable corpus. However, despite the obvious and potential benefits of using Wikipedia as a multilingual data source, few studies have been undertaken to analyse the degree of similarity between articles written in multiple languages. The similarity and textual relationship between articles written in multiple languages on the same topic can vary widely (e.g. between being re-used to written independently, even covering different aspects of a topic). In particular, we consider the multilingual nature of Wikipedia for a subset of languages, including those which one might call "well-resourced" and those which are "under-resourced" (i.e., for which there exist few language resources). Different from previous studies of languages and Wikipedia, we compare the number of articles written in the selected languages and the links to Wikipedia articles on the same topic, but in different languages. We also compare methods for computing the cross-language similarity between the articles.

For our experiments, we consider four different cross-language similarity models: (*i*) cross-language character n-grams using 3-grams (CL-C3G), (*ii*) cognateness (CL-COG), (*iii*) translation plus monolingual analysis (T+MA), and (*iv*) outlinks.[12] The first three models have been discussed already in Section 6.2.2. In the case of T+MA, the translation is carried out with Apertium, an open-source shallow-transfer MT system (Armentano-Oller et al., 2005). The monolingual comparison considers texts codified as BoW. We avoid the use of word n-grams, that show good results in monolingual settings (cf. Chapter 5), because the translation of d into L' can result in semantically equivalent but syntactically different versions. The pre-processing necessary for CL-C3G, CL-COG and T+MA is summarised in Table 9.5.

[12]Outlinks are hyperlinks that connect the contents of a given article to the corresponding concepts inside of Wikipedia, i.e., other articles.

9.3.1 Experimental Settings

In these experiments we selected 10 languages to compare cross-language similarity measures to contrast well-resourced languages (English, German, Spanish) and those with fewer available linguistic resources (Romanian, Lithuanian, Slovenian, Croatian, Estonian, Latvian and Greek).[13] The languages also exhibit different characteristics, such as writing systems used (English vs. Greek) and syntactic structure (English vs. Spanish).[14] The Wikipedia for each language was downloaded independently[15]. To extract the textual content of the articles and outlinks we used JWPL[16] (Zesch, Müller, and Gurevych, 2008). For comparing similarity measures across languages, we used topics for which a Wikipedia topic was available in all the ten languages. This resulted in around 3,600 topics for which we computed (symmetric) cross-language similarity between all language pairs.

In order to analyse the relationships between the different languages, we designed a total of four experiments.

Experiment 1: Measuring links between editions of Wikipedia. We aim at measuring the amount of Wikipedia articles (topics) in the ten considered languages. As we are interested in the multilingual characteristics, we measure the extent of connectivity to (corresponding) articles in other languages.

Experiment 2: Categories in multiple languages. For the English Wikipedia articles, we ranked the topics by the number of language links it has. We then manually categorised the top 1,000 topics into the following: time (year, date, month); location (country, states, city, sea); person; subject (e.g. articles about "History", "Agriculture", "Physics", "Geography", etc.); language (e.g. page about the English language, Italian language, etc.); and other.

Experiment 3: Translation + monolingual analysis. As a first approach, we focus the cross-language similarity analysis by assessing Spanish-English document pairs. Translation of Spanish articles into English allows the comparison between language-independent measures and the approach based on translation and monolingual analysis.

Experiment 4: Cross-language similarity. For each of the topics and language pairs we computed three measures of similarity (not the translation measure as this requires language resources which are not readily available for the under-resourced languages), and compare their results. Note that in this case we are not talking about text reuse. The idea is simply determining how the different measures behave when trying to estimate similarity.

[13]Except for Spanish (the native language of the author), the languages are the focus of the Accurat project (`http://www.accurat-project.eu/`), whose aim is the development of "research methods and techniques to overcome one of the central problems of machine translation: the lack of linguistic resources for under-resourced areas of machine translation". In 2011, during an internship of four months, we interacted with members of this project in the University of Sheffield.

[14]As two of our models are based on syntax, Greek texts were transliterated using ICU4J (`http://site.icu-project.org`).

[15]Data downloaded from `http://dumps.wikimedia.org/` in March 2010.

[16]`http://code.google.com/p/jwpl`

Source Language	Total Articles	Articles in >1 Languages (%)	Avg. No. Links
English	3,110,586	1,425,938 (45.84%)	4.84
German	1,036,144	636,111 (61.39%)	7.76
Spanish	571,846	424,795 (74.28%)	9.99
Romanian	141,284	106,321 (75.25%)	17.24
Lithuanian	102,407	67,925 (66.33%)	22.28
Slovenian	85,709	58,489 (68.24%)	21.02
Croatian	81,366	60,770 (74.69%)	23.47
Estonian	72,231	49,440 (68.45%	25.47
Greek	49,275	37,337 (75.77%)	29.62
Latvian	26,227	22,095 (84.25%)	33.36

Table 9.6: Languages used for similarity comparison. Languages ranked in descending order of the total number of articles that exist in each Wikipedia language (second column). The third column shows the number (and percentage) of articles which have links to at least one other language version. The fourth column shows the average number of links to different language versions per article (computed only over articles which have language links).

Category	%	Category	%	Category	%
Time	59	Person	3	Language	2
Location	28	Subject	2	Other	6

Table 9.7: Categories of the 1,000 English articles linked to the most different language versions.

9.3.2 Results and Discussion

Experiment 1: Measuring links between editions of Wikipedia. Table 9.6 summarises the distribution of articles across the ten languages. It is interesting to observe that although the number of articles in each language will affect the proportion of language links, there is not a direct correlation between these variables. For example, on the one hand the English articles have the most documents but also have the lowest number of average links: this indicates that most documents are being created independently of other languages. On the other hand, Latvian documents have the fewest number of articles but the highest average of language links suggesting that there is a trend that many articles are linked to other editions and a reliance on other languages. Although these figures do not indicate the possible degree of overlap of content between languages, they are certainly indicative of the amount of re-use that could be expected (higher for under-resourced languages).

Experiment 2: Categories in multiple languages. Results are shown in Table 9.7. English articles about time and location are the most commonly found in different language versions of Wikipedia. The highest ranked article is *True Jesus Church* which, by March 2010, linked to 238 equivalent articles. It was followed by locations (e.g. *Uetersen*, *Europe*, *Germany*, etc.).

Experiment 3: Translation + monolingual analysis. Upon manual inspection the similarity scores between the T+MA and the rest of models differ widely because: (*i*) the quality of translation and (*ii*) the degree of comparability between articles (i.e., some differ altogether in their contents even though they contain the same concept). The average similarity between the English and translated Spanish articles (T+MA) is 0.54. Between the English and Spanish articles the average similarity is 0.37 for CL-COG; 0.44 for CL-C3G and 0.26 for outlinks. The correlation of scores from "language-independent" measures with the translation approach is ρ(T+MA,CL-COG) $= 0.507$ ($p = 0.01$); ρ(T+MA, CL-C3G) $= 0.653$ ($p = 0.01$) and ρ(T+MA, outlinks) $= -0.229$ ($p = 0.01$). For this language pair, CL-C3G works well as a measure of cross-language similarity.

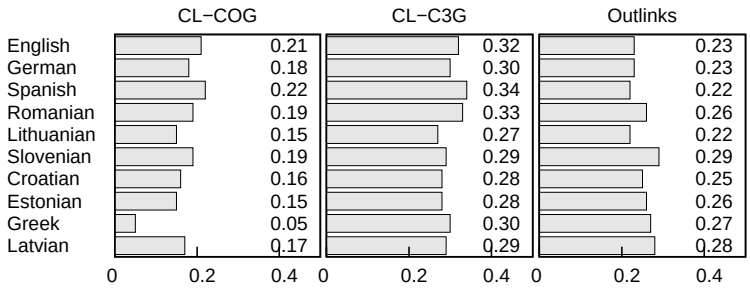


Figure 9.2: Average similarity scores across all topics and language pairs for each Wikipedia source language.

This shows that CL-C3G is a promising measure as it correlates well with translation, but it does not need any linguistic resource.

Experiment 4: Cross-language similarity. Figure 9.2 shows the average similarity scores computed across all language pairs (and topics) for each source language. Character n-grams show to be more stable than other models. In particular, they are less sensitive to transliteration, that causes cognateness similarity to decrease considerably. As in the Wikipedia cross-language experiments performed in Section 6.4, CL-C3G represents a robust measure of similarity among articles on the same topic in different languages. In fact, n-grams show to be barely sensitive to transliteration. No resources are needed to find similar documents using these measures. For the particular case of Wikipedia, considering its outlinks shows to be robust as well. Recently, Paramita, Clough, Aker, and Gaizauskas (2012) reported such human similarity judgements on a small set of Wikipedia articles in the ten languages, allowing for determining whether the similarity scores here obtained correlate with them in the future.

By analysing well-resourced and under-resourced languages, we found that articles of under-resourced languages have considerably higher number of language links on average, while documents from well-resourced languages tend to exist independently. However, the similarity scores of these languages are not too different from each other regardless of the language. Despite the obvious and potential benefits of using Wikipedia as a multilingual data source (at some stages, up to 282 editions of Wikipedia have existed, obviously representing the same amount of languages (Wikipedia, 2011p)), little work has explored the language distribution of Wikipedia and in particular the degree of overlap between articles written in multiple languages. Measuring the degree of similarity between articles written in multiple languages was mandatory.

9.4 Extracting Parallel Fragments from Wikipedia

The task of looking for re-used text fragments between different editions of Wikipedia is very similar to that of extracting parallel fragments from a comparable corpus. Here we discuss a model based on the IBM word alignment models for phrases alignments. As already mentioned, this problem has two definitions, depending on the discipline from which it is looked at: (*i*) analysing the phenomenon of re-use across Wikipedia editions in different languages and (*ii*) extracting parallel corpora from Wikipedia, useful

for statistical MT.[17] Within these experiments we analyse how feasible enriching our CL-ASA model is on the basis of comparable texts of Wikipedia. Once again, within Wikipedia, we take advantage of the explicit annotation: the langlinks that point from one article in L to the article on the same topic in L'. For our experiments we consider L =English and L' =Spanish.

9.4.1 Model Description

In order to extract re-used sentences, we consider pairs of Wikipedia articles identified beforehand: $X = (x_1, \ldots, x_j, \ldots, x_{|X|})$ and $Y = (y_1, \ldots, y_i, \ldots, y_{|Y|})$, where $X \in L$, $Y \in L'$ are articles on the same topic and x_j (y_i) is the $j(i)$-th sentence in X (Y). (x_j, y_i) is defined as an alignment between $x_j \in X$ and $y_i \in Y$ and A a finite set of alignments. Initially $A = (X \times Y)$, i.e., A contains every possible alignment between X and Y's sentences. The probability $p(x_j, y_i) \in A$ is computed by the IBM M4 (Brown *et al.*, 1993b), a word level alignment model broadly used in statistical MT.[18] On the left side, if the degree of co-occurrence between x_j and y_i vocabularies is high, $p(x_j, y_i)$ is high. On the right side, if few (or null) co-occurrences exist, $p(x_j, y_i)$ will be low. Once the alignments in A are computed $B \subseteq A$, a set of its most likely alignments, can be obtained by the following maximisation:

$$B \leftarrow (x_j, y_i) = \max_{y_i \in Y} p_{IBM}(x_j \mid y_i) \tag{9.1}$$

i.e., for every $x_j \in X$, the most likely alignment (x_j, y_i) according to the IBM M4 is kept in B. The final set is composed of those alignments in B that surpass a given threshold α. In our experiments we explore different values for α.

9.4.2 Parallel and Comparable Corpora

We used two corpora: one for training our translation model (parallel texts), and one for testing, composed of Wikipedia articles (comparable texts).

Learning corpora. In order to train the base translation model we used three corpora commonly used in statistical MT: (*i*) Europarl-v5 (Koehn, 2005), (*ii*) United Nations (Rafalovitch and Dale, 2009), and (*iii*) News-Commentary[19]. The overall statistics are included in Table 9.8. Our base translation model is trained with MGIZA (Gao and Vogel, 2008).[20]

[17] The alignment of bilingual corpora started more than twenty years ago (Brown *et al.*, 1991; Gale and Church, 1991). Nowadays, due to the increasing necessity of extracting parallel text for MT training, the extraction of parallel corpora from comparable ones is a hot topic (Eisele and Xu, 2010; Uszkoreit, Ponte, Popat, and Dubiner, 2010). Wikipedia has deserved special attention (Adafre and de Rijke, 2006; Mohammadi and GhasemAghaee, 2010).

[18] Fung and Cheung (2004) had used the M4 for extracting parallel sentences from "very non-parallel corpora" as well.

[19] `http://www.statmt.org/wmt11/translation-task.html`

[20] `http://geek.kyloo.net/software/doku.php/mgiza:overview`

Table 9.8: Overall statistics of the parallel corpora used for training. Number of sentences, tokens, and types included.

language	en	es
sentences	2.8M	
tokens	54M	58M
types	118k	164k

Table 9.9: Statistics of the Wikipedia articles test partition. Number of articles, sentences, potential alignments, tokens and types included.

language	en	es
articles	15	
sentences	661	341
pot. alignments	22,680	
tokens	24.5k	16.2k
types	3.4k	2.8k

The Wikipedia partition used for evaluation purposes is composed of a set of articles on similar domains to the MT training corpus. In particular, we selected a total of 15 article pairs on Economy and European Union administrative processes. The pre-processing applied to the articles plain text included sentence-splitting, tokenisation, and case folding. The statistics of the corpus after pre-processing are included in Table 9.9.

The articles were manually labelled applying a methodology inspired in Och and Ney (2003), but considering sentence instead of word's alignments. Two people manually annotated all the article pairs independently, identifying exact translations only. Two sets of alignments were defined and annotated: (a) P set of likely alignments and (b) S set of certain alignments. The former includes alignments between sentences that were not exact translations of each other. This includes sentences that are semantically equivalent or compose a $1 \times n$ or $n \times 1$ translation (i.e., one sentence is translated to more than one in the other language). The latter defines "quasi-exact" translations ($S \subseteq P$). Annotator 1 (2) generated the sets S_1 and P_1 (S_2 and P_2). These sets were combined into the global S and P as follows:

$$S = S_1 \cap S_2 \qquad\qquad P = P_1 \cup P_2$$

where $S \subseteq P$. P represents the pairs of sentences that our automatic process aims at retrieving. In particular, for the fifteen articles we obtained, $|S| = 10$ and $|P| = 115$ alignments.

9.4.3 Experimental Settings

We designed an exploratory experiment in order to assess how well our model performed on extracting re-used sentences from the Wikipedia articles. We exhaustively explored the influence of the parameter α in the retrieval quality.

For evaluation purposes, we used a set of five measures. Precision and recall were used considering the pairs of sentences in P as retrieval units. Moreover, as this is a problem close to the identification of parallel sentences, we opted for applying the *Sentence Alignment Error Rate* (SAER), inspired by Och and Ney (2003). Let X and Y

be a pair of documents (i.e., Wikipedia articles). Let S and P be the set of alignments between them, manually annotated. Let C be the set of alignments identified by our model. SAER is defined as:

$$SAER(S, P, C) = 1 - \frac{|C \cap S| + |C \cap P|}{|C| + |S|} \ . \tag{9.2}$$

As aforementioned, we aim at studying the influence of the α parameter. On the left side, the higher α, the more strict our model is. Intuitively, it would extract only parallel sentences, i.e., generated by exact translation. A higher than necessary α may cause retrieving no sentences at all. On the right side, a low α would admit further modified sentences as well, but at the cost of retrieving more noisy entries, i.e., false positives (fp). We aim at guaranteeing the retrieval of the highest ratio of true positives (TPR), minimising the ratio of false positives (FPR). Both proportions are computed as follows:

$$TPR = \frac{tp}{tp + fn} \ , \tag{9.3}$$

$$FPR = \frac{fp}{fp + tn} \ , \tag{9.4}$$

where tp represents the number of true positives, fn the number of false negatives, fp the number of false negatives and tn the number of true negatives.

9.4.4 Results and Discussion

First, we exhaustively explored the influence of α in the model. Figure 9.3 represents the relationship between the ratio of true positives (TPR, y axis) and ratio of false positives (FPR, x axis) respect to α. Firstly, due to the relative proportion of false positives, the curve would never reach a value of 1.0, as it is bounded by $fp/(fp + tn)$. Given that $fp \leq |X|$ (i.e., the maximum number of false positives is the number of sentences in the source document) and $TN \leq |X \times Y|$ (i.e., every possible alignment can be discarded), the value of the quotient is very small. Secondly, for the highest values of α the relation of false positives nearly reaches 0 for a ratio of 0.3 of true positives. For the smaller values of α, a value of 0.5 of TPR with a ratio of 0.02 of FPR can be obtained (the lower the value of α, the less restrictive the model becomes, hence accepting more possible alignments). In relative terms, the second point seems to be the best one, but by considering the absolute values, the differences are in the order of hundreds of false positives. Therefore, we select the first one, with $\alpha = 1.1 \cdot 10^{-3}$.

Table 9.10 shows the evaluation results together with some interesting statistics. We include the selected value of α and two extreme cases, aiming at stressing its influence in the obtained results. Additionally to the values of recall, precision and SAER, the number of retrieved alignments C, true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn) are included. Beside the simplicity of our model, when considering the best value of α, the alignment error is $SAER = 0.63$, with $prec = 0.59$

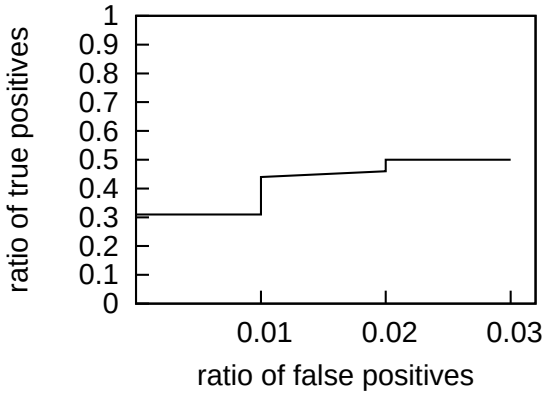


Figure 9.3: ROC representing the relationship between true and false positives respect to α. The value of α is inversely proportional to the evolution in both axes.

Table 9.10: Re-used sentences retrieval evaluation. $\alpha = 1.1 \cdot 10^{-3}$ is identified as best value. Two extreme values are included for comparison purposes.

α	$1 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$5 \cdot 10^{-2}$
Recall	1.00	0.90	0.10
Precision	0.09	0.59	0.50
F-measure	0.17	0.71	0.17
SAER	0.90	0.36	0.79
$\|C\|$	656	59	4
tp	58	35	2
tn	21,967	22,541	22,563
fp	598	24	2
fn	57	80	113

and $rec = 0.9$. Nevertheless, the recall value is very optimistic, as only 10 alignments are annotated as certain. Some interesting conclusions can be generated from the best and extreme values of α. With $\alpha = 1 \cdot 10^{-4}$ no single alignment is filtered, i.e., a candidate re-used sentence exists for every source sentence. With $\alpha = 5 \cdot 10^{-2}$ our model is too strict; hence, it would never be able to retrieve the 58 alignments that are in the reference. A countermeasure to this limitation would be considering the sentence alignments from both sides (i.e., from X to Y and from Y to X). Such an adaptation would require applying some heuristic to combine the union of alignments. As a result, more robust alignments would be obtained and alignments from 1 to m and vice versa could be considered (and even from m to m).

Table 9.11 shows some of the re-used sentences retrieved by our model. Undoubtedly the three sentences compose cases of cross-language re-use. We find that further analysing them to determine what is the source and what the target article (without relying on the Wikipedia intrinsic information) would be very interesting. In this preliminary experiment only cases of (nearly) exact translation can be identified. However, we consider that the obtained results are promising and open an avenue for the improvement of our cross-language similarity model CL-ASA.

Table 9.11: Instances of re-used sentence pairs properly retrieved.

English	Spanish
On 20 april 2005, the European Commission adopted the communication on Kosovo to the council "a european future for Kosovo" which reinforces the commission's commitment to Kosovo.	El 20 de abril de 2005, la Comisión Europea adoptó la comunicación sobre Kosovo en el consejo "un futuro europeo para Kosovo" que refuerza el compromiso de la comisión con Kosovo.
He added that the decisive factor would be the future and the size of the Eurozone, especially whether Denmark, Sweden and the UK would have adopted the euro or not.	Añadió que el factor decisivo será el futuro y el tamaño de la zona del euro, especialmente si Dinamarca, Suecia y el Reino Unido se unen al euro o no.
Montenegro officially applied to join the EU on 15 december 2008.	Oficialmente, Montenegro pidió el acceso a la UE el 15 de diciembre de 2008.

9.5 PAN@FIRE: Cross-Language !ndian Text Re-use

In the previous sections, we performed an analysis of text re-use within Wikipedia. However, the encyclopedia's contents can be used on other websites, in the formulation of laws (cf. Section 2.4.1.5), or as a favourite source for academic plagiarism (Head, 2010; Martínez, 2009). In this section we present the Cross-Language !ndian Text Re-Use detection task (CL!TR) (Barrón-Cedeño *et al.*, 2011), a challenge aimed at analysing cross-language text re-use between English and Hindi.[21] CL!TR is a branch of the PAN initiative (cf. Chapter 7). The corpora generated in the frame of PAN contain examples of automatically generated and simulated plagiarism including mono- and cross-language cases. The CL!TR@FIRE track is focussed on manually generated *cross-language text re-use*.[22]

9.5.1 Proposed Task

We targeted two languages: Hindi and English. The potentially re-used documents were all written in Hindi, whereas the potential source documents were in English. The corpus provided to participants is the CL!TR 2011 (cf. Section 4.2.5), and it is composed of a set of potentially re-used documents written in Hindi, D_{hi}, and a set of potential source documents written in English, D_{en}. The proposed task was to identify those documents in D_{hi} that were created by re-using fragments from a document $d \in D_{en}$. It can be described as follows:

[21] `http://www.dsic.upv.es/grupos/nle/fire-workshop-clitr.html`

[22] As already discussed, in the cross-language text re-use scenario the re-used text fragment and its source(s) are written in different languages, making the detection harder than when both texts are in the same language.

Let D_{en} be a collection of documents (Wikipedia articles). Let $d_q \in D_{hi}$ be a re-used document. Given d_q, retrieve those documents $d \in D_{en}$ that are likely source texts of d_q. Afterwards determine whether the pair $p(d_q, d)$ compose a case of re-use together with its source.

This is a document level task; no specific fragments inside of the documents were expected to be identified. Determining either a text has been re-used from its corresponding source is enough. Specifying the level of re-use (Exact, Heavy, or Light) was not necessary. For the training phase we provided an annotated corpus. The actual cases of re-use (re-used and source document) were labelled, as well as the specific kind of re-use they composed. During the test phase no annotation or hints about the cases were provided.

The success of a text re-use detection model was measured in terms of precision, recall, and F_1-measure on detecting the re-used documents together with their source in the test corpus. A detection is considered correct if the re-used document d_{hi} is identified together with its corresponding source document d_{en}. For the *prec*, *rec* and F_1 computation, we consider three sets:

- *total detected* is the set of suspicious-source pairs detected by the system,
- *correctly detected* is the subset of pairs detected by the system which actually compose cases of re-use, and
- *total re-used* is the gold standard, which includes all those pairs which compose actually re-used cases.

Precision and recall are defined as follows:

$$prec = \frac{\text{correctly detected}}{\text{total detected}} \qquad rec = \frac{\text{correctly detected}}{\text{total re-used}}$$

F_1-measure is used in order to compose the competition ranking.

9.5.2 Submissions Overview

Six teams from five different countries —India, Spain, Ireland, China (Hong Kong), and Ukraine— participated at the competition. They were allowed to submit up to three runs in order to encourage them to considering different approaches or parameter settings. A total of fifteen text re-use detection runs were submitted. Most of the participants opted for a "traditional" CLIR approach. They translated the suspicious documents in D_{hi} into English in order to perform a monolingual similarity estimation (Aggarwal, Asooja, and Buitelaar, 2011; Ghosh, Pal, and Bandyopadhyay, 2011a; Gupta and Singhal, 2011; Palkovskii and Belov, 2011; Rambhoopal and Varma, 2011). Most of these approaches exploited the Google or Bing translation services. The prototypical —IR— process that followed the language normalisation was as follows. D_{en} is indexed into a search engine (most of the participants use Nutch/Lucene[23]) and a document d_{hi} is queried to the

[23] `http://nutch.apache.org` and `http://lucene.apache.org`

search engine in order to retrieve the most similar documents $d \in D_{en}$. We now describe the information retrieval processes used by three approaches.

Aggarwal *et al.* (2011) do not apply any pre-processing to the documents in D_{en}, which are directly submitted into the index. Afterwards, the documents d_{hi} are queried against the index and the most relevant retrieved document is considered a candidate source document for d_{hi}. Ghosh *et al.* (2011a) splits the documents in D_{en} into paragraphs and expands their vocabulary on the basis of WordNet relationships (hyponyms, hypernyms and synsets). The enriched representation of each paragraph is fed into the index. The sentences in d_{hi} are queried against the index and the top 10 source paragraphs are retrieved. The best matches are considered in order to select pairs of re-used and source (entire) documents. Rambhoopal and Varma (2011) used IR process for their third run. After indexing D_{en}, key phrases were extracted from d_{hi} in order to independently query the index. The most frequently retrieved document $d_{en} \in D_{en}$ by the different key phrases in d_{hi} is selected as the source document.

Instead of translating the documents, Gupta and Singhal (2011) use a bilingual dictionary in order to map Hindi to English words. Words for which no possible translation exists in the dictionary are transliterated. Afterwards, a similarity estimation is carried out between the representations of d_{hi} and d_{en}. They submitted three runs that incrementally added processing stages: (*i*) only dictionary based mapping is applied to d_{hi} (run 1); (*ii*) mapping and transliteration are applied to d_{hi} (run 2); and (*iii*) in addition to the mapping and transliteration processes, a minimal similarity threshold has to be surpassed in order to consider that d_{hi} is re-used from d_{en} (run 3).

Palkovskii and Belov (2011) applies a fingerprinting model in order to detect exact string matches. After discarding non alpha-numeric characters, chunks of 5 words with a sliding window of 4 are hashed. All the matches from d_{en} to d_{hi} are merged and used to estimate whether a case of re-use is at hand. Their three runs considered different parameters for the fingerprinting process. The best settings are those just described.

In addition to the approach based on a search engine that was just described, Rambhoopal and Varma (2011) also submitted two more approaches based on machine learning. The model is based on a J48 decision tree classifier. For run 1 the features for the classifier were composed of the cosine similarity estimated over stemmed word 3-grams. For run 2 stopwords were removed and key phrases extracted. The relevance and length of the sequences compose the features for the classifier. The approach of Addanki and Wu (2011) is based on machine learning as well. They used an SVM classifier considering features of statistical MT and sentence alignment models. The features for the classification process are three: (*i*) and (*ii*) are the score of the most likely alignments at sentence and paragraph level between d_{hi} and d_{en}, respectively (these scores were computed with the length-based alignment algorithm proposed by Gale and Church (1993)) and (*iii*) is a lexical feature: a Hindi-English dictionary was used to gloss the Hindi documents and calculate an *idf*-based cosine similarity between suspicious and potential source documents.

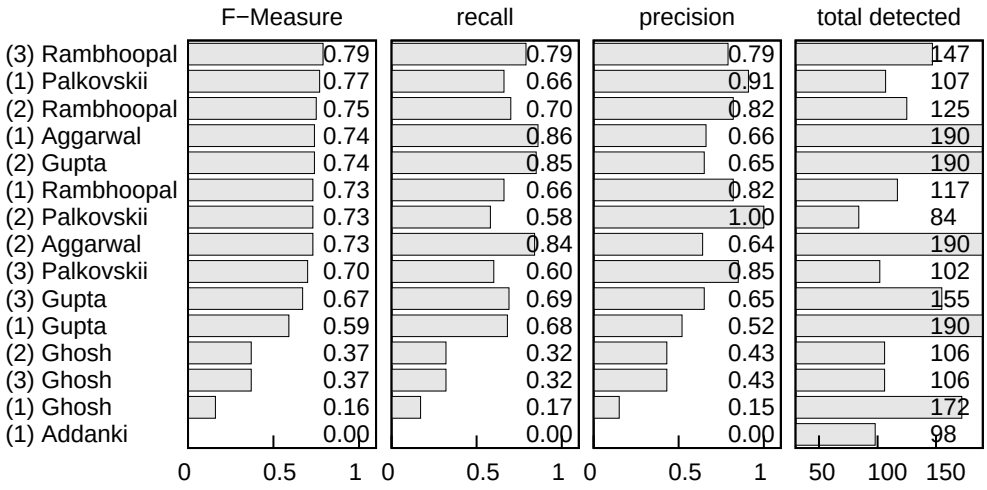


Figure 9.4: CL!TR *overall* evaluation results. Additionally to rank and evaluation of the runs, we show the number of suspicious documents identified as re-used.

9.5.3 Results and Discussion

The evaluation results are presented in Fig. 9.4. The most successful approaches for this task are based on standard CLIR techniques. After translating the suspicious documents into English and building a search engine, Rambhoopal and Varma (2011) composed the queries by selecting a set of key phrases from the suspicious document. This approach strikes a good balance between recall and precision, with an F-measure of 0.79. The second best approach considered word 5-grams as terms (Palkovskii and Belov, 2011). This kind of representation is very sensitive to changes in the original text and is better suited to identifying exact matches. As a result, their obtained precision is among the highest: 0.91, with still a reasonable level of recall: 0.66. Note that in terms of F-measure, the difference between the top three approaches is only of 0.04.

On the other hand, the highest recall value is obtained by Aggarwal *et al.* (2011): 0.86, at the cost of a slightly reduced precision: 0.66. They opt for a full representation of d_{hi} when generating the queries to the search engine. Moreover, Aggarwal *et al.* (2011) decided to assume that every document in D_{hi} was re-used and simply retrieved the most similar document in D_{en}. This assumption was made by Gupta and Singhal (2011) as well (note that in total four submissions reported 190 documents as re-used).

The bad results obtained by the approach of Addanki and Wu (2011) may be due to the nature of constructing re-use cases. As aforementioned, the documents in D_{hi} contain, in general, one single paragraph. For the re-used partition, this paragraph has been extracted from entire Wikipedia articles, causing the length factor to be less expressive (even if the length factors used are at sentence and paragraph level).

In order to perform a type-wise evaluation (i.e., regarding at exact, light and heavy revisions in isolation), the actual cases of re-use and the detections of a given detector were sub-sampled as follows. Let G be the set of actual cases of re-use (composed of re-used and original text). Let E, L, and H be the actual cases of exact, light and heavy

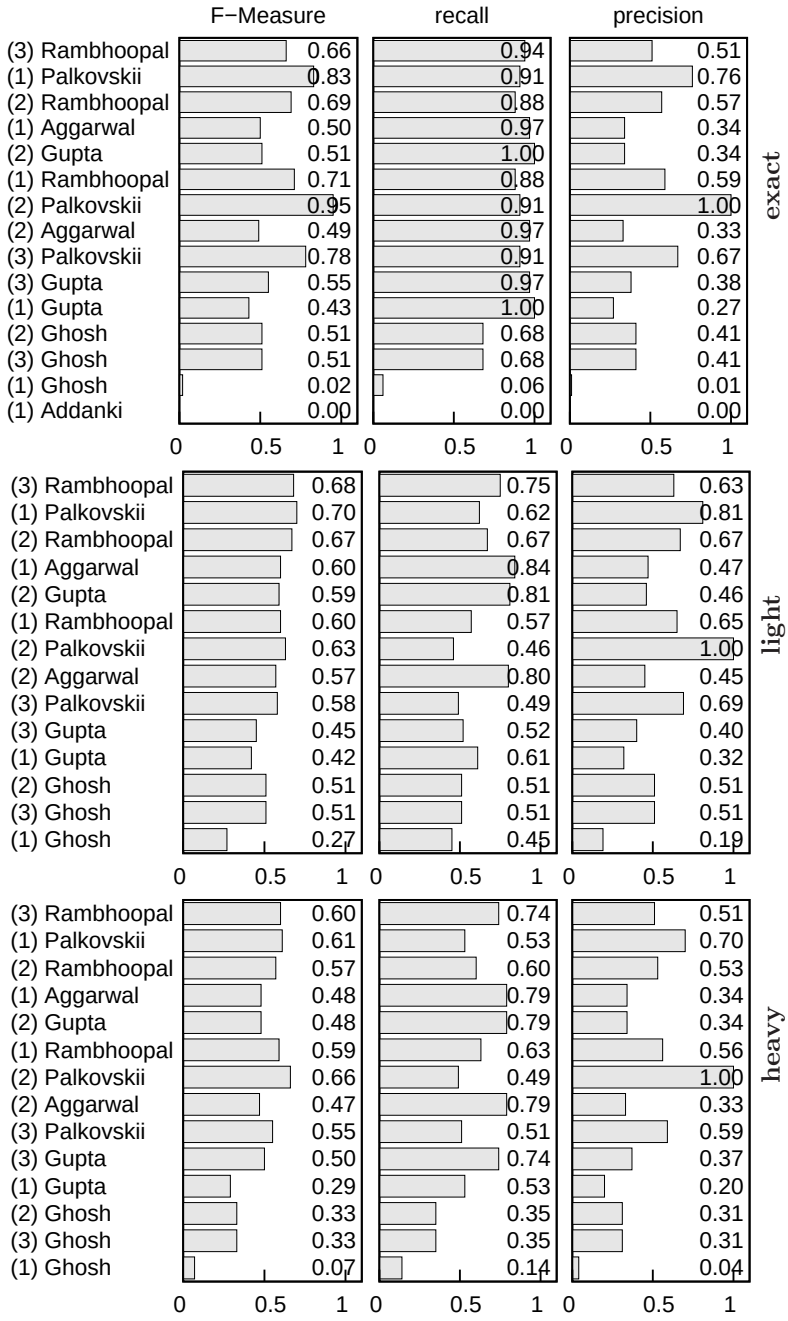


Figure 9.5: CL!TR evaluation results for *exact*, *light* and *heavy* cases.

revisions in G, i.e.,

$$G = \{E \cup L \cup H\} \ .$$

Let P_d be the set of cases identified as re-used by a given detector. The gold standard partition considered for the evaluation when analysing exact cases is simply $G_E = E$. The partition of detections considered is defined as:

$$P_{d,E} = P_d \setminus \{p_d \mid p_d \in P_d \cap (L \cup H)\} \ ,$$

i.e., those properly detected cases that correspond to light and heavy revisions are dis-

carded. The same procedure is followed when sub-sampling for evaluating cases of light and heavy revision. However, those cases in P_d which are not actual cases of re-use are considered in every resulting partition: $P_{d,E}$, $P_{d,L}$, and $P_{d,H}$. This does not effect recall, but reduces precision, and therefore F-measure. [24]

Figure 9.5 shows the results when considering the cases of *exact* cross-language re-use as well as *light* and *heavy* revisions. As aforementioned, these results have to be observed with caution. The precision bias caused by our sub-sampling strategy causes the approach of Palkovskii and Belov (2011) to outperform the other participants in the three cases. Once again this was expected as they pay special attention to precision.

The rest of our type-wise analysis of the results is centred in recall. As expected, the values of recall for the exact cases are the highest, as they are the easiest to detect. Indeed, Gupta and Singhal (2011) reached $rec = 1.0$ in two of their runs and many other participants obtain values above 0.9. Higher levels of paraphrasing cause these values to decrease. Whereas the average recall of the submissions (that managed to detect at least one case), on exact cases is of 0.84, for light and heavy revisions is of 0.61 and 0.57 only. As observed in Chapters 7 and 8, paraphrases, also across languages, cause a decrease of the performance of the plagiarism detectors. However, regardless of the level of paraphrasing, most of the top approaches still managed to properly retrieve more than half of the cases.

The surprisingly high results obtained by some of the approaches have to be read with caution. Most of them perform language normalisation based on on-line translators (such as that offered by Google). When generating the cases, the volunteers were allowed to use these and other automatic tools to translate the contents they had selected to answer a given question and further modify it.

9.6 Chapter Summary

In this chapter we analysed three re-use phenomena that have to do with Wikipedia: co-derivation among articles revisions, re-use across articles in different languages, and re-use from articles into external documents across different languages.

In the case of co-derivatives detection, different monolingual similarity were applied without further adaptation: Jaccard coefficient, cosine, word chunk overlap, Okapi BM25; and two models were adapted in order to measure similarity between texts: Kullback-Leibler distance and a model based on (monolingual) statistical machine translation. Additionally, two well-known models for plagiarism detection were included: Winnowing and `SPEX`. Our aim was determining which one performed the best within different conditions. Our evaluation, in terms of recall and precision as well as highest false match and separation, made possible to estimate not only whether every relevant document was retrieved, but the distance between the similarity values computed for relevant and irrelevant documents as well. By considering these three factors more comprehensive information was available to select the most suitable method. The obtained results show that, as it is expected, at document level Winnowing and `SPEX` have the

[24] This strategy for computing type-wise evaluations is similar to that used at PAN@CLEF.

best results. The advantage of Winnowing is that the generation of a fingerprint for a given document is independent from the others. However, it must be considered that if derivation or plagiarism implies further modifications, Winnowing does not seems to be the best option. This is reflected in the experiment carried out at section level. In this case the statistical and vector space models (Jaccard coefficient, cosine measure, etc.) outperform the rest of models. The results confirmed that in text-reuse detection the frequency of a term can be neglected in most cases.

The first part of our analysis across languages aimed at investigating the degree of similarity among Wikipedia articles considering a diverse subset of languages. We investigated a variety of similarity measures, ranging from some which do not need any linguistic resource to those which require machine translation. Our findings suggest that character n-grams in combination with outlinks represent a robust measure of similarity among articles on the same topic in different languages. In fact, n-grams show to be barely sensitive to transliteration. No resources are needed to find similar documents using these measures. By analysing well-resourced and under-resourced languages, we found that articles of under-resourced languages have considerably higher number of language links on average, while documents from well-resourced languages tend to exist independently. However, the similarity scores of these languages are not too different from each other regardless of the language.

The second part of our analysis across languages aimed at investigating how a heuristic model managed to extract re-used sentences from a set of comparable articles in Wikipedia. Our model is based on an adapted model for phrase alignment, borrowed from statistical machine translation. Our results showed promising as in a set of preliminary experiments on identifying re-used sentences (parallel and highly comparable in machine translation terms) we managed to obtain high success rates. We aim at investigating further this approach in the future.

Our aim with these analyses was determining the feasibility of better tune our cross-language similarity assessment model —CL-ASA— by considering comparable Wikipedia articles across languages. Our first study allows for determining what pairs of articles (and languages) are the most likely to include parallel (re-used) fragments. With this information, we can target in a set of languages and article pairs in order to extract good parallel sentences. Such sentences can be further used to train better statistical bilingual dictionaries for CL-ASA.

Finally, we presented an overview of the Cross-Language !ndian Text Re-Use Detection Competition, an event we recently organised in order to promote the development of better detectors of cross-language text re-use, particularly on distant languages. The challenge consisted of identifying, among a set of short documents written in Hindi, those texts that had been generated by re-use and the corresponding source document was written in English. The benchmark collection (the first text collection of this nature) allowed for the comparison of fifteen approaches. Most of them were based on standard cross-language information retrieval approaches and some others on statistical machine translation and machine learning techniques. The most successful approach was based on a preliminary translation of the documents, followed by key phrase-based retrieval process.

Related publications:

- Barrón-Cedeño, Eiselt, and Rosso (2009a)
- Barrón-Cedeño, Rosso, Lalitha Devi, Clough, and Stevenson (2011)

Chapter 10

Conclusions

> Plagiarism will die and be reborn with a positive connotation in the Information Age. What we now call plagiarism will become a basic skill. Instead of trying to prevent it, we will teach it... the student who can find, analyze, and display an elegant solution to a task possesses the skills necessary to prosper in the Information Age. Whether the solution is his/her own or someone else's is irrelevant.
>
> Rodney P. Riegle

The main focus of this research was on the development of models for automatic text re-use and plagiarism detection. Special attention was paid to those cases of text re-use carried out on the basis of translation and paraphrasing; also to text re-use in and from Wikipedia. Cross-language plagiarism represents a problem nearly approached that has received interest just recently, during the last years. The detection of strongly paraphrased plagiarism, which is closer to plagiarism of ideas is still in its infancy. Wikipedia represents an interesting co-derivation environment and is identified as a preferred source for plagiarism, hence requesting attention.

The research work was structured as follows:

1. Analysis of the phenomena behind text re-use and plagiarism.
2. Analysis of techniques for representation and comparison of texts and their further application to mono- and cross-language text re-use and plagiarism detection.
3. Contribution on the creation of better evaluation frameworks for mono- and cross-language plagiarism detection.
4. Study of the impact of paraphrasing on plagiarism and its relevance in the design of plagiarism detection models.
5. Development of a cross-language model for automatic text re-use and plagiarism detection based on statistical machine translation.

6. Comparison of cross-language text re-use detection models to asses how similar Wikipedia articles are across languages and their application to the detection of text re-use from Wikipedia.

The publications generated in the framework of this research work are included in Appendix B. The research work done on cross-language plagiarism detection was covered by media, and some references are included in Appendix C.

10.1 Contributions

The main contributions of this work have been:

- A survey on the amount of cross-language plagiarism committed in academia. Around 35% of the students we queried declared they have plagiarised, at least once, from sources written in a language different than their native one.
- Cooperation in the creation of the first standardised frameworks for automatic text re-use and plagiarism detection: PAN@CLEF and PAN@FIRE.
- A model based on statistical machine translation specially designed for cross-language plagiarism detection. The feasibility to feed in with parallel samples extracted from Wikipedia was also investigated.
- A seminal study on paraphrasing types to analyse which are the most difficult to detect. The insights will be helpful for the development of the future generation of plagiarism detectors. The P4P, a subset of the PAN-PC-10 corpus tagged with paraphrasing types, will allow for investigating further the core mechanisms of plagiarism: paraphrase.

10.2 Research Answers

In Section 1.2 we identified the three main difficulties of automatic text re-use and plagiarism detection: lack of collections with actual cases of plagiarism and the complexity that paraphrasing and translation cause to automatic plagiarism detection. Here we discuss our insights for the research questions raised from these difficulties.

1. How to build a standard collection of documents for the study and development of automatic plagiarism detection?

(a) What are the reasons behind the lack of freely available collections of documents with actual cases of plagiarism?

As plagiarism represents a fault, ethical and legal issues prevent from publishing collections with this kind of borrowing. Even ignoring these issues, a corpus with actual cases of plagiarism (and the corresponding sources) should be fully annotated to be useful. This allows for delimiting original and plagiarised text fragments. This task has been tackled on relatively small corpora, such as the METER corpus. This

is due to the difficulty that a corpus of bigger size would imply. Up to date, the only alternative has been the exploitation of corpora with cases of fair text re-use (see next question) and the generation of documents with synthetic cases of plagiarism, either generating them with software (e.g. the PAN-PC-09 corpus), or asking volunteers to act as plagiarists (e.g. the PAN-PC-{10,11} and CL!TR corpora).

(b) Under what circumstances currently available corpora of text re-use are useful for plagiarism detection?

Whereas their application has different aims (either deceiving or simply adhering to some writing style), the paraphrase phenomena behind plagiarism and, in general, text re-use, are common. As a result, from a natural language processing and information retrieval point of view, the difference between plagiarism and fair re-use seems not to be significant. Therefore, corpora including cases of journalistic text re-use and co-derivation, among others, are worth considering for the development of plagiarism detection models.

(c) How to build a corpus with artificial cases of plagiarism?

We helped in the generation of two "sister" corpora containing simulated cases of re-use and plagiarism. They share certain characteristics, but also have others that differentiate them. The PAN-PC series of corpora look at composing a realistic information retrieval challenge, hence they are composed of thousands of documents, including thousands of plagiarism cases ranging from verbatim copy to artificially and (a few) manually created cases, some of them across languages. The CL!TR corpus looks at composing a realistic cross-language challenge, hence it is composed of just a few thousand documents, including hundreds of re-use cases, all of them generated manually and across distant languages. It includes different levels of cross-language paraphrasing, resulting in exact as well as slightly and heavily modified translations. Though these two corpora are not the first of this nature, they (particularly the PAN-PC series) have become a reference in the development of models for plagiarism detection, filling an important gap in the area.

(d) How valid is a synthetic corpus for plagiarism detection?

The answer to this question depends on the area from which it is looked at. From the perspective of information retrieval, it seems that synthetic corpora are indeed a valid alternative. Synthetic obfuscation strategies have received particular criticism, but considering some of the most commonly used IR characterisation schemas, the difference between manual —well-formed— and randomly mixed up re-use seems irrelevant. From the forensic linguistics point of view, synthetic cases are completely useless, as the linguistic evidence they require to present during a legal process needs to accomplish with different conditions, including authenticity (i.e., being created by humans), under certain circumstances. From the perspective of linguistics, those cases of simulated (human-made) plagiarism in the synthetic corpora represent a valuable resource for paraphrases analysis. Finally, from a practical point of view, synthetic corpora have demonstrated to be an important resource in the development and tuning of better plagiarism detection systems. The success of different research teams that have applied their models on these synthetic corpora and run plagiarism detectors in their own institutions supports their validity.

2. What models perform the best to detect cases of re-use with high level of paraphrasing?

(a) Are simple —syntax-based— models enough for detecting this kind of borrowing?

The vast majority of models for text re-use detection consider syntax information only. Regardless of the simplicity of such models, they still manage to detect a few cases of paraphrase plagiarism. The reason is that they exploit those text fragments borrowed with null modification and assume that their context is re-used as well (for instance, when two pairs of strings are highly similar but surround a text fragment that was not identified as such). Moreover, we have identified that simple pre-processing strategies, such as case folding and stemming, have the potential of diminishing some of the effects of paraphrasing.

(b) How can paraphrases analysis techniques support text re-use and plagiarism detection?

The best way of supporting paraphrase text re-use and plagiarism detection is understanding which types of paraphrases underlay text re-use acts. For the first time, we analysed the paraphrase phenomena applied when a set of plagiarised text fragments had been manually generated. Our seminal study showed that lexical substitutions are the paraphrase mechanisms used the most when plagiarising. Moreover, all the paraphrase mechanisms tend to be used to generate a summarised version of the re-used text. Therefore, a model intended to succeed in detecting paraphrase re-use certainly requires to include modules that compose a robust enough text characterisation to consider such paraphrases: the expansion (or contraction) of related vocabulary, the normalisation of formatting and word forms, and the inclusion of mechanisms that model the expected length of a re-used fragment given its source.

(c) What are the paraphrasing operations that most mislead automatic detectors?

Regardless of any particular paraphrasing operation, we observed a strong correlation between linguistic (i.e., kind of paraphrasing) and quantitative (i.e., amount of paraphrases) complexity and the performance of plagiarism detectors. That is, the occurrence of more and more complex paraphrases implies worse detection performance. As a starting point, note that in general the success of a detector decreases to less than half when approaching paraphrase respect to verbatim plagiarism. Semantics-based paraphrase changes have shown to be among the most misleading: the amount of detected cases of plagiarism with a high concentration of these changes by various models was practically null. The reason is that they imply high lexical differences and unclear mapping between the source and plagiarised (paraphrased) fragments.

3. How can we detect cases of text re-use across languages?

(a) How can we build a collection of cross-language text re-use cases?

We have explored two approaches. On the one hand, we investigated the automatic generation of cross-language cases of re-use through automatic machine translation. Nevertheless, this naïve approach misled the efforts on cross-language plagiarism detection, which tended to defuscate these cases by a simple translation process.On

the other hand, we tried with the manual generation of cross-language cases, some of them still with assistance of machine translators, but implying further paraphrasing. These cases showed to be much harder to detect, stressing the necessity of more robust cross-language detection models.

(b) How well do (adapted) models for CLIR perform when aiming at detecting text re-use?

We explored a range of cross-language information retrieval techniques. In particular, we observed that a simple model based on the texts' characterisation by short character n-grams (CL-CNG) was worth considering when dealing with languages using common alphabets (and different alphabets, after transliteration), and particularly if they have some influence (e.g. recall levels of 0.8 for English-French versus 0.2 for English-Polish when aiming at retrieving documents translations). Still it is not better than other, more sophisticated, models (see next question). We also tried with a naïve model based on translating the documents into a common language. The results are relatively good on exact translations (e.g. recall values of 0.6 for sentence level detection for Spanish-Basque). However, this approach requires the previous translation of all the texts, which can be prohibitive for large collections (assuming a translator for the working languages is at hand) and potentially infeasible for "on the fly" applications or for detecting paraphrased plagiarism on the Web.

(c) How well do (adapted) models for machine translation perform when detecting re-use?

We studied the capabilities of our proposed model cross-language alignment-based similarity analysis. Such model is based on translation probabilities and length distributions between the analysed texts. Our empirical results showed that this model offers competitive results when looking for re-used texts, regardless of whether they have been generated by manual or automatic translation. It performs better than cross-language semantic analysis and cross-language character n-grams, identified as two of the most appealing models for cross-language similarity assessment when dealing with translations at document and fragment level. Moreover, it can be used across different languages,[1] regardless of their alphabet or influence.

10.3 Future Trends of Research

A number of issues remain open in the research on text re-use and plagiarism detection. Here we stress some of the most interesting ones from our point of view.

Identification of proper citations.

Probably one of the most interesting topics for future research on plagiarism detection is the automatic analysis of citations and references, a problem that, although mentioned

[1]Special attention should be paid to under-resourced languages, as their lack of resources increases the tendency to re-use from other languages.

already by Maurer *et al.* (2006), still remains unsolved. When a text fragment is properly re-used, i.e., including its corresponding citation or reference, it does not compose a case of plagiarism, and it would not be considered as suspicion. The development of techniques for discriminating between plagiarism and properly cited re-use requires, firstly, obtaining (or generating) diverse cases of quotations and, more general, proper and wrong citations. Afterwards, models that recognise proper citation patterns could be designed.

Text re-use and search engines.

As previously stressed, text re-use detection implies no looking for topical similarity, but for co-derivation of texts. Therefore, text re-use detection models cannot be considered when designing a "traditional" topic-based search engine. However, we consider that models for text re-use detection are still worth considering in modern search engines. For instance, in a monolingual setting, models for text re-use detection could be used to improve the search diversity: avoiding to present near-duplicate documents on top of the returned ranking.

In a cross-language setting, these models could be used to assist the user for cross-language searches. We imagine the following scenario: (*i*) a person queries a search engine in a language different to her native one (e.g. Spanish is her native language and English the language of the query); (*ii*) the search engine returns a relevant document written in English; on the basis of algorithms for cross-language text re-use detection, a translation in Spanish is found. (*iii*) instead of offering the possibility for translating the relevant document "on the fly", the search engine offers the available translation, with a high likelihood of representing a better translation.

Mono- and cross-language intrinsic plagiarism detection.

We consider that intrinsic plagiarism detection should be re-designed. Such a model has plenty of limitations and assumptions, for instance, the suspicious document is supposed to be written by one single author $\mathcal{A}$. As a result, intrinsic plagiarism detection becomes practically useless when handling co-operatively authored text. Another problem is, assuming the task can be solved (i.e., identifying fragments with unexpected style and complexity differences), determining which fragments have been originally written by $\mathcal{A}$ and which borrowed. This represents in fact an authorship attribution problem, for which preliminary texts produced by $\mathcal{A}$ are necessary (and by considering such documents, this is not an intrinsic approach any more!). Therefore, we propose to split the problem of intrinsic plagiarism detection into two: (*i*) "intra-document authorship delimitation" in which fragments written by different authors (no matter who such authors actually are) included in a document are identified; and (*ii*) determining, by means of authorship identification techniques, what fragments are likely to have been produced by $\mathcal{A}$.[2]

The problem of cross-language intrinsic plagiarism detection would be, given a suspicious document, determining whether (some of) its contents have been generated by means of translation. Some seminal efforts have been conducted recently (Baroni and Bernardini, 2006; Koppel and Ordan, 2011), and we believe it represents an interesting

[2]This idea is very similar to the way the task of intrinsic plagiarism detection is being proposed for PAN 2012.

path for future research.

Cross-language re-use in source code.

Whereas work on source code re-use detection started nearly forty years ago (Ottenstein, 1976), Internet has made available huge amounts of code. As for texts, "much plagiarism and adaptation is now of computer programs" (Wilks, 2004).[3] We look at finding cases of re-use across programming languages as an interesting problem.

When dealing with this problem, it is possible to take advantage of the certain similarity between the reserved words across programming languages. As a result, we have made some preliminary efforts on detecting re-use among codes in C++, Java, and Python, using the cross-language character n-grams model (Flores, Barrón-Cedeño, Rosso, and Moreno, 2011, 2012). We believe that our cross-language alignment-based analysis model may be valuable as well for two reasons: (*i*) building a statistical dictionary for reserved words is feasible and (*ii*) even the length model could be worth considering to determine, for instance, how short a code in Python should be respect to its source in Java.

Extracting parallel samples from comparable corpora.

We identify some potential improvements for our cross-language alignment based analysis. For instance, considering two-way dictionaries, i.e., computing double likelihoods from language L to L' and vice versa. Our future work is focussed on enriching the translation model's dictionary. We plan to do so by further exploiting Wikipedia as a multilingual resource: extracting parallel fragments from its comparable articles in different language pairs. The outcome will be a valuable resource for machine translation as well. We hope to investigate these issues in the framework of the EU-ERCIM post-doctorate at the *Universitat Politécnica de Catalunya BarcelonaTech*.

[3]Refer to Burrows (2010) for an overview of models for analysis of authorship in source code reuse.

References

The secret of creativity is knowing how to hide your sources

Albert Einstein

ABC (2011). Pérez-Reverte, tras ser condenado por plagio: "Es una emboscada" (Pérez-Reverte, after being condemned for plagiarism: "It is an ambush"). ABC Diario [`http://bit.ly/abc_perez`], Spain. Published: 6/May/2011; Accessed: 14/Oct/2011.

Academic Plagiarism (2011). Academic Plagiarism Checker. [`http://www.academicplagiarism.com/`]. Accessed 16/Aug/2011.

Adafre, S. and de Rijke, M. (2006). Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Adar, E., Skinner, M., and Weld, D. (2009). Information Arbitrage Across Multi-lingual Wikipedia. In Baeza-Yates, Boldi, Ribeiro-Neto, and Cambazoglu (2009).

Addanki, K. and Wu, D. (2011). An Evaluation of MT Alignment Baseline Approaches upon Cross-Lingual Plagiarism Detection. In FIRE (2011).

Adler, B., Chatterjee, I., and de Alfaro, L. (2008). Assigning Trust to Wikipedia Content. In *Proceedings of the 4th International Symposium on Wikis*, Porto, Portugal. ACM.

Adler, B., de Alfaro, L., Mola-Velasco, S., Rosso, P., and West, A. (2010). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. *Computational Linguistics and Intelligent Text Processing, 11th International Conference*, **LNCS (6609)**, 277–288. Springer-Verlag.

Aggarwal, N., Asooja, K., and Buitelaar, P. (2011). Cross Lingual Text Reuse Detection Using Machine Translation & Similarity Measures. In FIRE (2011).

Akiva, N. (2011). Using Clustering to Identify Outlier Chunks of Text. In Petras, Forner, and Clough (2011).

Alcázar, A. (2006). Towards Linguistically Searchable Text. In *Proceedings of the BIDE (Bilbao-Deusto) Student Conference in Linguistics 2005*, Bilbao, Basque Country.

Alegria, I., Forcada, M., and Sarasola, K., editors (2009). *Proceedings of the SEPLN 2009 Workshop on Information Retrieval and Information Extraction for Less Resourced Languages*, Donostia, Basque Country. University of the Basque Country.

Alsedo, Q. (2008). Bunbury y Pedro Casariego: ¿plagio, préstamo o imposible coincidencia? (Bunbury and Pedro Casariego: plagiarism, borrowing or impossible coincidence?). El Mundo [`http://mun.do/mundo_bunbury`], Spain. Published: 5/Sep/2008; Accessed: 14/Oct/2011.

Alzahrani, S. and Salim, N. (2010). Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection. In Braschler and Harman (2010).

Anderson, G. (1999). Cyberplagiarism. A Look at the Web Term Paper Sites. *College & Research Libraries News*, **60**(5), 371–373.

Appelt, D., editor (1991). *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, Berkeley, CA, USA. Association for Computational Linguistics.

Applied Linguistics LLC (2011). Grammarly. [http://www.grammarly.com/]. Accessed 17/Aug/2011.

Arce, M. (2010). Un diputado K presentó un proyecto de ley contra el plagio, plagiado (A K Deputy presented a law project against plagiarism, plagiarised). Clarín [`http://bit.ly/clarin_arce`], Argentina. Published: 14/May/2010;Accessed: 13/Oct/2011.

Argamon, S. and Juola, P. (2011). Overview of the International Authorship Identification Competition at PAN-2011. In Petras *et al.* (2011).

Argamon, S. and Levitan, S. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. In *Proceedings of the Association for Computers and the Humanities, and the Association for Literary and Linguistic Computing Conference*, Victoria, BC, Canada.

Armentano-Oller, C., Corbí-Bellot, A., Forcada, M., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., and Sánchez-Martínez, F. (2005). An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, pages 23–30, Phuket, Thailand.

Association of Teachers and Lecturers (2008). School Work Plagued by Plagiarism - ATL Survey. Technical report, Association of Teachers and Lecturers, London, UK. Press release.

Atserias, J., Casas, B., Comelles, E., Gonzáles, M., Padró, L., and Padró, M. (2006). FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library.

In Calzolari, Choukri, Gangemi, Maegaard, Mariani, Odijk, and Tapias (2006). http://www.lsi.upc.edu/ nlp/freeling.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press, New York, NY.

Baeza-Yates, R., Boldi, P., Ribeiro-Neto, B., and Cambazoglu, B., editors (2009). *Proceedings of the Second ACM International Conference on Web Search and Web Data Mining*, Barcelona, Spain. ACM.

Bani-Ahmad, S., Cakmak, A., Ozsoyoglu, G., and Hamdani, A. (2005). Evaluating Publication Similarity Measures. *IEEE Data Engineering Bulletin*, **28**(4), 21–28.

Baroni, M. and Bernardini, S. (2006). A New Approach to the Study of Translationese: Machine Learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, **21**(3), 259–274.

Barrio, C. (2011). Pérez-Reverte, condenado por plagio (Pérez-Reverte, condemned for plagiarism). Interviu [`http://bit.ly/intervieu_perez`], Spain. Published: 29/Apr/2011; Accessed: 14/Oct/2011.

Barrón-Cedeño, A. (2010). On the Mono- and Cross-Language Detection of Text Reuse and Plagiarism. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Geneva, Switzerland. ACM Press.

Barrón-Cedeño, A. and Rosso, P. (2008). Towards the Exploitation of Statistical Language Models for Plagiarism Detection with Reference. In Stein *et al.* (2008), pages 15–19. http://ceur-ws.org/Vol-377.

Barrón-Cedeño, A. and Rosso, P. (2009a). On Automatic Plagiarism Detection based on n-grams Comparison. *Advances in Information Retrieval. Proceedings of the 31st European Conference on IR Research*, **LNCS (5478)**, 696–700. Springer-Verlag.

Barrón-Cedeño, A. and Rosso, P. (2009b). On the Relevance of Search Space Reduction in Automatic Plagiarism Detection. *Procesamiento del Lenguaje Natural*, **43**, 141–149.

Barrón-Cedeño, A. and Rosso, P. (2010). Towards the 2nd International Competition on Plagiarism Detection and Beyond. In JISC Plagiarism Advisory Service (2010).

Barrón-Cedeño, A., Rosso, P., Pinto, D., and Juan, A. (2008). On Cross-Lingual Plagiarism Analysis Using a Statistical Model. In Stein *et al.* (2008), pages 9–13. http://ceur-ws.org/Vol-377.

Barrón-Cedeño, A., Eiselt, A., and Rosso, P. (2009a). Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In D. Sharma, V. Varma, and R. Sangal, editors, *ICON 2009*, pages 29–38, Hyderabad, India. Macmillan Publishers.

Barrón-Cedeño, A., Rosso, P., and Benedí, J. (2009b). Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. *Computational Linguistics and Intelligent Text Processing, 10th International Conference*, **LNCS (5449)**, 523–534. Springer-Verlag.

Barrón-Cedeño, A., Potthast, M., Rosso, P., Stein, B., and Eiselt, A. (2010a). Corpus and Evaluation Measures for Automatic Plagiarism Detection. In Calzolari, Choukri, Maegaard, Mariani, Odjik, Piperidis, Rosner, and Tapias (2010).

Barrón-Cedeño, A., Vila, M., and Rosso, P. (2010b). Detección automática de plagio: De la copia exacta a la paráfrasis. In Garayzábal Heinze *et al.* (2010), pages 76–96.

Barrón-Cedeño, A., Rosso, P., Agirre, E., and Labaka, G. (2010c). Plagiarism Detection across Distant Language Pairs. In Huang and Jurafsky (2010).

Barrón-Cedeño, A., Basile, C., Degli Esposti, M., and Rosso, P. (2010d). Word Length n-grams for Text Re-Use Detection. *Computational Linguistics and Intelligent Text Processing, 10th International Conference*, **LNCS (6008)**, 687–699. Springer-Verlag.

Barrón-Cedeño, A., Rosso, P., Lalitha Devi, S., Clough, P., and Stevenson, M. (2011). PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In FIRE (2011).

Barzilay, R. (2003). *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University.

Barzilay, R. and Lee, L. (2003). Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2003)*, Edmonton, Canada. ACL.

Barzilay, R. and McKeown, K. (2001). Extracting Paraphrases from a Parallel Corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barzilay, R., McKeown, K., and Elhadad, M. (1999). Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 550–557, Maryland, USA. ACL.

Basile, C., Benedetto, D., Caglioti, G., and Degli Esposti, M. (2009). A Plagiarism Detection Procedure in Three Steps: Selection, Matches and Squares. In Stein, Rosso, Stamatatos, Koppel, and Agirre (2009), pages 19–23. http://ceur-ws.org/Vol-502.

Baty, P. (2000). Copycats Roam in Era of the Net. Times Higher Education [`http://bit.ly/timeshigher_copycats`]. Published: 14/Jul/2000; Accessed: 03/Jan/2012.

Baum, L. (1972). An inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process. *Inequalities*, **3**, 1–8.

BBC (2011). Plagiarism: The Ctrl+C, Ctrl+V boom. BBC News Magazine [http://www.bbc.co.uk/news/magazine-12613617], UK. Published: 2/Mar/2011; Accessed: 25/May/2011.

Beléndez Vázquez, M., Comas Forgas, R., Marta., M., Muñoz González, A., and Topa Cantisano, G. (2011). Plagio y otras prácticas académicamente incorrectas entre el alumnado universitario de nuevo ingreso. In *IX Jornadas de redes de investigación en docencia universitaria*, Alicante, Spain.

Bendersky, M. and Croft, W. (2009). Finding Text Reuse on the Web. In Baeza-Yates *et al.* (2009), pages 262–271.

Bennett, C., Li, M., Vitányi, P., and Zurek, W. (1998). Information Distance. *IEEE Transactions on Information Theory*, **44**(4), 1407—1423.

Berger, A. and Lafferty, J. (1999). Information Retrieval as Statistical Translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, Berkeley, CA. ACM Press.

Bernstein, Y. and Zobel, J. (2004). A Scalable System for Identifying Co-Derivative Documents. *String Processing and Information Retrieval. Proceedings of the Symposium on String Processing and Information Retrieval*, **LNCS (3246)**, 1–11. Springer-Verlag.

Bhagat, R. (2009). *Learning Paraphrases from Text*. Ph.D. thesis, University of Southern California.

Bierce, A. (1911). *The Devil's Dictionary*. Doubleday, Page & Company.

Bigi, B. (2003). Using Kullback-Leibler Distance for Text Categorization. *Advances in Information Retrieval: Proceedings of the 25th European Conference on IR Research (ECIR 2003)*, **LNCS (2633)**, 305–319. Springer-Verlag.

Blackboard Inc (2011). Safe Assign. [`http://safeassign.com`]. Accessed 16/Aug/2011.

Blanch-Mur, C., Rey-Abella, F., and Folch-Soler, A. (2006). Nivel de conducta académica deshonesta entre los estudiantes de una escuela de ciencias de la salud. *Enfermería clínica*, **16**(2), 57–62.

Boisvert, R. and Irwin, M. (2006). Plagiarism on the Rise. *Communications of the ACM*, **49**(6), 23–24.

Brandes, U. and Lerner, J. (2007). Revision and Co-revision in Wikipedia. Detecting Clusters of Interest. In *Bridging the Gap between Semantic Web and Web 2.0 (SemNet 2007)*, pages 85–96.

Braschler, M. and Harman, D., editors (2010). *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy.

Brin, S., Davis, J., and Garcia-Molina, H. (1995). Copy Detection Mechanisms for Digital Documents. In M. Carey and D. Schneier, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 398–409, San Jose, California. ACM Press.

Broder, A. (1997). On the Resemblance and Containment of Documents. In *Compression and Complexity of Sequences (SEQUENCES'97)*, pages 21–29, Salerno, Italy. IEEE Computer Society.

Brown, P., Lai, J., and Mercer, R. (1991). Aligning Sentences in Parallel Corpora. In Appelt (1991), pages 169–176.

Brown, P., Della Pietra, S., Della Pietra, V., Goldsmith, M., Hajic, J., Mercer, R., and Mohanty, S. (1993a). But Dictionaries Are Data Too. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993b). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, **19**(2), 263–311.

Bull, J., Collins, C., Coughlin, E., and Sharp, D. (2001). Technical Review of Plagiarism Detection Software Report. Technical report, University of Lutton.

Burrows, S. (2010). *Source Code Authorship Attribution*. Ph.D. thesis, School of Computer Science and Information Technology, RMIT University, Melbourne, Australia.

Burrows, S., Potthast, M., and Stein, B. (2012). Paraphrase Acquisition via Crowdsourcing and Machine Learning (to appear). *ACM Transactions on Intelligent Systems and Technology*.

Burstein, J. (2009). Opportunities for Natural Language Processing Research in Education. *Computational Linguistics and Intelligent Text Processing , 10th International Conference*, **LNCS (5449)**, 6–27. Springer-Verlag.

Buttler, D. (2004). A Short Survey of Document Structure Similarity Algorithms. In *5th International Conference on Internet Computing*, pages 3–9, Las Vegas, NV.

Cabanes, I. (2009). Condenan a un catedrático de Murcia por plagiar la tesis de su alumna. Levante-EMV [`http://bit.ly/levante_murcia`], Spain. Published: 21/Apr/2009; Accessed: 29/Jan/2009.

Callison-Burch, C. (2007). *Paraphrasing and Translation*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors (2006). *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., editors (2010). *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Cavanillas, S. (2008). Cyberplagiarism in University Regulations. In Comas and Sureda (2008a), pages 7–12. [`http://bit.ly/cyberplagiarism_cs`].

Cerf, V. (2011). Interview with Susana Rivera Torres for Fr@ctal (in the framework of the 2011 Campus Party Mexico). Fr@ctal TV Show. ForoTV Channel [`http://bit.ly/fractal_cerf`], Mexico City, Mexico. Published: Jul/2011; Accessed: 12/Aug/2011.

Ceska, Z., Toman, M., and Jezek, K. (2008). Multilingual Plagiarism Detection. In *Proceedings of the 13th International Conference on Artificial Intelligence (ICAI 2008)*, pages 83–92, Varna, Bulgaria. Springer-Verlag.

Chapman, K. and Lupton, R. (2004). Academic Dishonesty in a Global Educational Market: A Comparison of Hong Kong and American University Business Students. *International Journal of Educational Management*, **18**(7), 425–435.

Chomsky, N. (1957). *Syntactic Structure*. Mouton & Co.

Chong, M., Specia, L., and Mitkov, R. (2010). Using Natural Language Processing for Automatic Detection of Plagiarism. In JISC Plagiarism Advisory Service (2010).

Church, K. (1993). Char_align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993)*, pages 1–8, Columbus, OH, USA. ACL.

Church, K. and Helfman, J. (1993). Dotplot: A Program for Exploring Self-Similarity in Millions of Lines for Text and Code. *Journal of American Statistical Association, Institute for Mathematical Statistics and Interface Foundations of North America*, **2**(2), 153–174.

Clough, P. (2000). Plagiarism in Natural and Programming Languages: an Overview of Current Tools and Technologies. Research Memoranda: CS-00-05, Department of Computer Science. University of Sheffield, UK.

Clough, P. (2001). A Perl Program for Sentence Splitting Using Rules. Technical report, Department of Computer Science, University of Sheffield.

Clough, P. (2003). Old and new challenges in automatic plagiarism detection. National UK Plagiarism Advisory Service.

Clough, P. (2010). *Copyright and Piracy: An Interdisciplinary Critique*, chapter Measuring Text Re-Use in the News Industry. Cambridge University Press, Cambridge, UK.

Clough, P. and Gaizauskas, R. (2009). Corpora and Text Re-Use. In A. Lüdeling, M. Kytö, and T. McEnery, editors, *Handbook of Corpus Linguistics*, Handbooks of Linguistics and Communication Science, pages 1249—1271. Mouton de Gruyter.

Clough, P. and Stevenson, M. (2011). Developing a Corpus of Plagiarised Short Examples. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, **45**(1), 5–24.

Clough, P., Gaizauskas, R., Piao, S., and Wilks, Y. (2001). METER: Measuring Text Reuse. Technical Report CS-01-03, Department of Computer Science, University of Sheffield.

Clough, P., Gaizauskas, R., and Piao, S. (2002). Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1678–1691, Las Palmas, Spain. European Language Resources Association (ELRA).

Collberg, C. and Kobourov, S. (2005). Self-Plagiarism in Computer Science. *Communications of the ACM*, **48**(4), 88–94.

Comas, R. and Sureda, J., editors (2008a). *Academic Cyberplagiarism [online dossier]*, volume 10 of *Digithum. Iss*. UOC. [`http://bit.ly/cyberplagiarism_cs`].

Comas, R. and Sureda, J. (2008b). Academic Cyberplagiarism: Tracing the Causes to Reach Solutions. In Comas and Sureda (2008a), pages 1–6. [`http://bit.ly/cyberplagiarism_cs`].

Comas, R., Sureda, J., Nava, C., and Serrano, L. (2010). Academic Cyberplagiarism: A Descriptive and Comparative Analysis of the Prevalence amongst the Undergraduate Students at Tecmilenio University (Mexico) and Balearic Islands University (Spain). In *Proceedings of the EDULEARN'10 Conference*, Barcelona, Spain.

Comas-Forgas, R., Sureda-Negre, J., and Salva-Mut, F. (2010). Academic Plagiarism Prevalence among Spanish Undergraduate Students: An Exploratory Analysis. *Biochemia Medica, Special Issue on Responsible Writing in Science*, **20**(3), 301–306.

Comas Forgas, R., Sureda Negre, J., and Oliver Trobat, M. (2011). Prácticas de citación y plagio académico en la elaboración textual del alumnado universitario. *Teoría de la educación: Educación y cultura en la sociedad de la información*, **12**(1), 359–385.

Cooke, N., Gillam, L., Wrobel, P., Cooke, H., and Al-Obaidli, F. (2011). A High-performance Plagiarism Detection System - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Copy Tracker (2011). Copy Tracker. The Free Plagiarism Detection Tool. [`http://copytracker.ec-lille.fr`]. Accessed 17/Aug/2011.

Corezola Pereira, R., Moreira, V., and Galante, R. (2010a). A New Approach for Cross-Language Plagiarism Analysis. *Multilingual and Multimodal Information Access Evaluation, International Conference of the Cross-Language Evaluation Forum*, **LNCS (6360)**, 15–26. Springer-Verlag.

Corezola Pereira, R., Moreira, V., and Galante, R. (2010b). UFRGS@PAN2010: Detecting External Plagiarism Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Coulthard, M. (2004). Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, **25**, 431–447.

Coulthard, M. (2010). The Linguist as Detective: Forensic Applications of Language Description. [http://bit.ly/madrid_lingforense], Madrid, Spain. Talk at: Jornadas (In)formativas de Lingüística Forense ((In)formative Conference on Forensic Linguistics).

Coulthard, M. and Alison, J. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge, Oxon, UK.

Crot Anti Plagiarism Solutions (2011). Crot. [Online; accessed 16-August-2011].

Culicover, P. (1968). Paraphrase Generation and Information Retrieval from Stored Text. *Mechanical Translation and Computational Linguistics*, **11**(1 and 2), 78–88.

Culwin, F. (2008). A Longitudinal Study of Nonoriginal Content in Final-Year Computing Undergraduate Projects. *IEEE Transactions on Education*, **51**(2), 189–194.

de Montaigne, M. (1802). *Essais. Tome Premier*. L'Imprimerie et la Fonderie Stéréotypes de Pierre Didot, Paris, France. Available at [books.google.com/books?id=1jc9AAAAYAAJ] (November 2011).

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dice, L. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, **26**, 297–302.

Dolan, W. and Brockett, C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)*, Jeju, Korea.

Dorr, B., Green, R., Levin, L., Rambow, O., Farwell, D., Habash, N., Helmreich, S., Hovy, E., Miller, K., Mitamura, T., Reeder, F., and Siddharthan, A. (2004). Semantic Annotation and Lexico-Syntactic Paraphrase. In *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Corpora*.

Dowd, M. (1987). Biden's Debate Finale: An Echo from Abroad. The New York Times [http://nyti.ms/widka3], USA. Published: 12/Sep/1987; Accessed: 13/Oct/2011.

Dras, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. thesis, Macquarie University, Sydney, Australia.

Dumais, S., Letsche, T., Littman, M., and Landauer, T. (1997). Automatic Cross-Language Retrieval Using Latent Semantic Indexing. In *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*, pages 24—26. Stanford University.

Duplichecker (2011). Duplichecker. http://www.duplichecker.com. Accessed 16/Aug/2011.

Dutrey, C., Bernhard, D., Bouamor, H., and Max, A. (2011). Local Modifications and Paraphrases in Wikipedia's Revision History. *Procesamiento del Lenguaje Natural*, **46**, 51–58.

Eisele, A. and Xu, J. (2010). Improving Machine Translation Performance using Comparable Corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora LREC 2010*, pages 35–41. ELRA.

El País (1990). Una juez condena a Vázquez Montalbán por plagio de una traducción (A judge condemns Vázquez Montalbán for plagiarising a translation). El País [http://bit.ly/pais_vazquez], Spain. Published: 5/Jul/1990; Accessed: 14/Oct/2011.

El País (2008). Bunbury: "No me defiendo de la acusación de plagio" (Bunbury: "I am not defending myself about the plagiarism accusation"). El País [http://bit.ly/pais_bunbury], Spain. Published: 10/Sep/2008; Accessed: 14/Oct/2011.

Ephorus (2011). Ephorus. http://www.ephorus.com/. Accessed 16/Aug/2011.

España-Bonet, C., Vila, M., Martí, M. A., and Rodríguez, H. (2009). CoCo, a Web Interface for Corpora Compilation. *Procesamiento del Lenguaje Natural*, **43**, 367–368.

eTBLAST (2011). eTBLAST 3.0. http://etest.vbi.vt.edu/etblast3/. Accessed 17/Aug/2011.

Etzioni, O., Reiter, K., Soderland, S., and Sammer, M. (2007). Lexical Translation with Application to Image Search on the Web. In *Machine Translation Summit XI*, Copenhagen, Denmark.

Faigley, L. and Witte, S. (1981). Analyzing revision. *College Composition and Communication*, **32**(4), 400–414.

FIRE, editor (2011). *FIRE 2011 Working Notes. Third Workshop of the Forum for Information Retrieval Evaluation*.

Fitzgerald, J. (2010). Authorial Attibution and Alleged Suicide Communications: Forensic Linguistic Analysis and Expert Testimony in Three 2007 Violent Crimes. [http://bit.ly/madrid_lingforense]. Talk at: Jornadas (In)formativas de Lingüística Forense ((In)formative Conference on Forensic Linguistics).

Flores, E., Barrón-Cedeño, A., Rosso, P., and Moreno, L. (2011). Towards the Detection of Cross-Language Source Code Reuse. *International Conference on Applications of Natural Language to Information Systems*, **LNCS (6716)**, 250–253. Springer-Verlag.

Flores, E., Barrón-Cedeño, A., Rosso, P., and Moreno, L. (2012). DeSoCoRe: Detecting Source Code Re-Use across Programming Languages. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montreal, Canada. ACL.

Franklyn-Stokes, A. and Newstead, S. (1995). Undergraduate Cheating: Who Does What and Why? *Studies in Higher Education*, **20**, 159–172.

Fuglede, B. and Topse, F. (2004). Jensen-Shannon Divergence and Hilbert Space Embedding. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT'04)*, page 31, Chicago, IL. IEEE.

Fujita, A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Ph.D. thesis, Nara Institute of Science and Technology, Nara, Japan.

Fullam, K. and Park, J. (2002). Improvements for Scalable and Accurate Plagiarism Detection in Digital Documents. In *Proceedings of the 8th International Conference on Parallel and Distributed Systems*, pages 8–23, Troy, NY.

Fung, P. and Cheung, P. (2004). Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In Lin and Wu (2004), pages 57—63.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Gaillard, B., Boualem, M., and Collin, O. (2010). Query Translation using Wikipedia-based Resources for Analysis and Disambiguation. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*, Saint-Raphaël, France. European Association for Machine Translation .

Gale, W. and Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. In Appelt (1991), pages 177–184.

Gale, W. and Church, K. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, **19**, 75–102.

Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In K. Cohen and B. Carpenter, editors, *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA. Association for Computational Linguistics.

Garayzábal Heinze, E., Jiménez Bernal, M., and Reigosa Riveiros, M., editors (2010). *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. (Jornadas (In)formativas de Lingüística Forense)*. Euphonia Ediciones SL., Madrid, Spain.

Gellerstam, M. (1985). Translationese in Swedish Novels Translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, pages 88–95, Lund, Sweden. CWK Gleerup.

Ghosh, A., Pal, S., and Bandyopadhyay, S. (2011a). Cross-Language Rext Re-Use Detection Using Information Retrieval. In FIRE (2011).

Ghosh, A., Bhaskar, P., Pal, S., and Bandyopadhyay, S. (2011b). Rule Based Plagiarism Detection using Information Retrieval. In Petras *et al.* (2011).

Gibbons, J. and Turell, M. (2008). *Dimensions of Forensic Linguistics*. John Benjamins Publishing Company, Amsterdam, The Netherlands. [`http://bit.ly/gbooks_gibbons`].

Gillam, L., Marinuzzi, J., and Ioannou, P. (2010). TurnItOff - Defeating Plagiarism Detection Systems. In *Proceedings of the 11th Higher Education Academy-ICS Annual Conference*. Higher Education Academy.

Gipp, B., Meuschke, N., and Beel, J. (2011). Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches Using GuttenPlag. In *11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Ottawa, Canada. IEEE.

Girón Castro, S. (2008). Creatividad: Plagio no detectado. Universidad Sergio Arboleda [`http://bit.ly/usarboleda_creatividad`]. Accessed 17/Aug/2011.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-Document Summarization By Sentence Extraction. In *NAACL-ANLP 2000 Workshop on Automatic Summarization*, pages 40–48, Seattle, WA. Association for Computational Linguistics.

Gottron, T. (2010). External Plagiarism Detection based on Standard IR. Technology and Fast Recognition of Common subsequences. In Braschler and Harman (2010).

Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenization.

Grman, J. and Ravas, R. (2011). Improved Implementation for Finding Text Similarities in Large Collections of Data - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Grozea, C. and Popescu, M. (2010a). Encoplot - Performance in the Second International Plagiarism Detection Challenge Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Grozea, C. and Popescu, M. (2010b). Who's the Thief? Automatic Detection of the Direction of Plagiarism. *Computational Linguistics and Intelligent Text Processing, 10th International Conference*, **LNCS (6008)**, 700–710. Springer-Verlag.

Grozea, C. and Popescu, M. (2011). The Encoplot Similarity Measure for Automatic Detection of Plagiarism - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Grozea, C., Gehl, C., and Popescu, M. (2009). ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In Stein *et al.* (2009), pages 10–18. http://ceur-ws.org/Vol-502.

Gunning, R. (1968). *The Technique of Clear Writing*. McGraw-Hill.

Gupta, P. and Singhal, K. (2011). Mapping Hindi-English Text Re-use Document Pairs. In FIRE (2011).

Gupta, P., Sameer, R., and Majumdar, P. (2010). External Plagiarism Detection: N-Gram Approach using Named Entity Recognizer. Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

HaCohen-Kerner, Y., Tayeb, A., and Ben-Dror, N. (2010). Detection of Simple Plagiarism in Computer Science Papers. In Huang and Jurafsky (2010), pages 421–429.

Haines, V., Diekhoff, G., LaBeff, G., and Clarck, R. (1986). College Cheating: Immaturity, Lack of Commitment, and the Neutralizing Attitude. *Research in Higher Education*, **25**(4), 342–354.

Hariharan, S. and Srinivasan, R. (2008). A Comparison of Similarity Measures for Text Documents. *Journal of Information & Knowledge Management*, **7**(1), 1–8.

Hartrumpf, S., vor der Brück, T., and Eichhorn, C. (2010). Semantic Duplicate Identification with Parsing and Machine Learning. *Text, Speech and Dialogue (TSD 2010)*, **LNAI (6231)**, 84–92. Springer-Verlag.

Head, A. (2010). How today's college students use Wikipedia for course-related research. *First Monday*, **15**(3). [`http://bit.ly/first_head`].

Heintze, N. (1996). Scalable Document Fingerprinting. In *USENIX Workshop on Electronic Commerce*.

Hoad, T. and Zobel, J. (2003). Methods for Identifying Versioned and Plagiarized Documents. *Journal of the American Society for Information Science and Technology*, **54**(3), 203–215.

Holmes, D. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society*, **155**(1), 91–120.

Honore, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin*, **7**(2), 172–177.

Huang, C.-R. and Jurafsky, D., editors (2010). *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China. Coling 2010 Organizing Committee.

IEEE (2008). A Plagiarism FAQ. [`http://bit.ly/ieee_plagiarism`]. Published: 2008; Accessed 3/Mar/2010.

Ingendaay, P. (2011). War die Guttenberg-Affäre denn zu gar nichts gut? (Was the Guttenberg good for nothing?). Frankfurter Allgemeine Feuilleton [`http://bit.ly/frankfurter_mejuto`], Frankfurt, Germany. Published: 20/Jun/2011; Accessed: 12/Oct/2011.

iParadigms (2010). Turnitin. [`http://www.turnitin.com`]. Accessed 3/Mar/2010.

Irribarne, R. and Retondo, H. (1981). *Plagio de obras literarias. Ilícitos civiles y penales en derecho de autor*. IIDA, Buenos Aires, Argentina.

Iyer, P. and Singh, A. (2005). Document Similarity Analysis for a Plagiarism Detection System. In *Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI 2005)*, pages 2534–2544, Pune, India.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, **37**, 547–579.

Jackson, P. and Moulinier, I. (2002). *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins Publishing Company, Amsterdam/Filadelfia.

Jackson, R. and Jackson, J. (2008). *Forensic Science. Second Edition*. Pearson Education Limited, Essex, England.

JISC Plagiarism Advisory Service, editor (2004). *Plagiarism: Prevention, Practice and Policies Conference*, Newcastle upon Tyne, UK. Plagiarism Advice.

JISC Plagiarism Advisory Service, editor (2010). Newcastle upon Tyne, UK. Plagiarism Advice.

Johnson, S. (1755). *A Dictionary of the English Language: in which the Words Are Deduced from Their Originals, Explained in Their Different Meanings, and Authorised by the NAMES of the WRITERS in whose Works They Are Found*. Richard Bentley, UK. [`http://bit.ly/gbooks_johnson`] Accessed Aug/2011.

Jones, K., Reid, J., and Bartlett, R. (2008). Cyber Cheating in an Information Technology Age. In Comas and Sureda (2008a), pages 19–27. [`http://bit.ly/cyberplagiarism_cs`].

Jurafsky, D. and Martin, J. (2000). *Speech and Language Processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall Inc.

Kang, N., Gelbukh, A., and Han, S. (2006). PPChecker: Plagiarism Pattern Checker in Document Copy Detection. *Text, Speech and Dialogue (TSD 2006)*, **LNAI (4188)**, 661–667. Springer-Verlag.

Karp, R. M. and Rabin, M. O. (1987). Efficient Randomized Pattern-Matching Algorithms. *IBM Journal of Research and Development*, **31**(2), 249–260.

Kasprzak, J. and Brandejs, M. (2010). Improving the Reliability of the Plagiarism Detection. System Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Kasprzak, J., Brandejs, M., and Kripač, M. (2009). Finding Plagiarism by Evaluating Document Similarities. In Stein *et al.* (2009), pages 24–28. http://ceur-ws.org/Vol-502.

Keck, C. (2006). The Use of Paraphrase in Summary Writing: A Comparison of L1 and L2 Writers. *Journal of Second Language Writing*, **15**, 261–278.

Kent, C. and Salim, N. (2009). Web Based Cross Language Plagiarism Detection. *Journal of Computing*, **1**(1).

Kent, C. and Salim, N. (2010). Web Based Cross Language Plagiarism Detection. In *Second International Conference on Computational Intelligence, Modelling and Simulation*, pages 199–204. IEEE.

Kestemont, M., Luyckx, K., and Daelemans, W. (2011). Intrinsic Plagiarism Detection Using Character Trigram Distance Scores: Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Ketchen, D. and Shook, C. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, **17**(6), 441–458.

Kincaid, J., Fishburne, R., Rogers, R., and Chissom, B. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Memphis TN Naval Air Station*, **Research B**.

Kisiel, R. (2009). JK Rowling Sued for £ 500m in Plagiarism Lawsuit by Family of Late WillY the Wizard Author. Daily Mail [`http://bit.ly/mail_rowling`]. Published: 16/Jun/2009; Accessed: 22/Aug/2011.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Machine Translation Summit X*, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.

Koppel, M. and Ordan, N. (2011). Translationes and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, Portland, OR.

Kornai, A. (2007). *Mathematical Linguistics*. Springer-Verlag.

Krovetz, R. (1993). Viewing Morphology as an Inference Process. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, Pittsburgh, PA. ACM Press.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, **22**(1), 79–86.

La Jornada (2007). Paula Mues niega haber plagiado texto de académica de la UNAM (Paula Mues denies having plagiarised UNAM scholar's text. La Jornada [http://bit.ly/lajornada_mues], Mexico. Published: 31/Oct/2007; Accessed: 12/Oct/2011.

La Vanguardia (2010). El juez dicta la apertura de juicio oral contra Planeta por un presunto plagio de Camilo José Cela. La Vanguardia [http://bit.ly/lavanguardia_cela], Spain. Published: 22/Aug/2010; Accessed: 22nd August, 2011.

Lalitha Devi, S., Rao, P., Sundar Ram, V., and Akilandeswari, A. (2010). External Plagiarism Detection. Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Lancaster, T. and Culwin, F. (2004). A Visual Argument for Plagiarism Detection Using Word Pairs. In JISC Plagiarism Advisory Service (2004).

Lancaster, T. and Culwin, F. (2005). Classifications of Plagiarism Detection Engines. *ITALICS*, **4**(2).

Lee, C., Wu, C., and Yang, H. (2008). A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection. In *Proceedings of the 3rd International Conference on Innovative Computing Information (ICICIC'08)*. IEEE Computer Society.

Lembersky, G., Ordan, N., and Wintner, S. (2011). Language Models for Machine Translation: Original vs. Translated Texts. In R. Barzilay and M. Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland. Association for Computational Linguistics.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, **10**, 707–710.

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag.

Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism*, Austin, TX.

Limited, C. S. (2011). Copycatch. http://cflsoftware.com. Accessed 17/Aug/2011.

Lin, D. and Wu, D., editors (2004). *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain.

Lindey, A. (1952). *Plagiarism and Originality*. Harper and Brothers, New York, NY.

Littman, M., Dumais, S., and Landauer, T. (1998). *Cross-Language Information Retrieval, chapter 5*, chapter Automatic Cross-language Information Retrieval Using Latent Semantic Indexing, pages 51–62. Kluwer Academic Publishers.

Liu, Y., Zhang, H., Chen, T., and Teng, W. (2007). Extending Web Search for Online Plagiarism Detection. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2007)*, pages 164–169, Las Vegas, Nevada. IEEE.

Luhn, H. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, pages 159–165.

Lukashenko, R., Graudina, V., and Grundspenkis, J. (2007). Computer-based Plagiarism Detection Methods and Tools. In *Proceedings of the 2007 International Conference on Computer Systems and Technologies (CompSysTech 2007)*. ACM.

Lynch, J. (2006). The Perfectly Acceptable Practice of Literary Theft: Plagiarism, Copyright, and the Eighteenth Century. *Colonial Williamsburg*.

Lyon, C., Malcolm, J., and Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In L. Lee and D. Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 118–125, Pennsylvania. Association for Computational Linguistics.

Lyon, C., Barret, R., and Malcolm, J. (2004). A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector. In JISC Plagiarism Advisory Service (2004).

MacQueen, J. (1967). Some Methods for Classification and Analysis of MultiVariate Observations. In L. Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Malcolm, J. and Lane, C. (2009). Tackling the PAN'09 External Plagiarism Detection Corpus with a Desktop Plagiarism Detector. In Stein *et al.* (2009), pages 29–33. http://ceur-ws.org/Vol-502.

Malcolm, J. and Lane, P. (2008). An Approach to Detecting Article Spinning. In *Plagiarism: Prevention, Practice and Policies Conference*, Newcastle upon Tyne, UK. Plagiarism Advice.

Mallon, T. (2001). *Stolen Words. The Classic Book on Plagiarism*. Harcourt Inc.

Manning, C. and Schütze, H. (2002). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Manning, C., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Markov, Z. and Larose, D. (2007). *Data Mining the Web*. John Wiley & Sons, Inc., New Jersey, USA. [`http://bit.ly/markov_dmining`].

Martin, B. (1994). Plagiairsm: a Misplaced Emphasis. *Journal of Information Ethics*, **3**(2), 36–47.

Martínez, I. (2009). Wikipedia Usage by Mexican Students. The Constant Usage of Copy and Paste. In *Wikimania 2009*, Buenos Aires, Argentina. [http://wikimania2009.wikimedia.org].

Maurer, H., Kappe, F., and Zaka, B. (2006). Plagiarism - A Survey. *Journal of Universal Computer Science*, **12**(8), 1050–1084.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.

Mcnamee, P. and Mayfield, J. (2004). Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval*, **7**(1-2), 73–97.

Merriam-Webster (2011). Merriam-Webster Online Dictionary. [http://www.merriam-webster.com/dictionary/plagiarize]. (11 May 2011).

Metzler, D., Bernstein, Y., Croft, W., Moffat, A., and Zobel, J. (2005). Similarity Measures for Tracking Information Flow. In Chowdhury, Fuhr, Ronthaler, Schek, and Teiken, editors, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 517–524, Bremen, Germany. ACM Press.

Meyer zu Eißen, S. and Stein, B. (2006). Intrinsic Plagiarism Detection. *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research (ECIR 2006)*, **LNCS (3936)**, 565–569. Springer-Verlag.

Meyer zu Eißen, S., Stein, B., and Kulig, M. (2007). Plagiarism Detection without Reference Collections. *Advances in Data Analysis*, pages 359–366.

Micol, Danieland Ferrández, O., Llopis, F., and Muñoz, R. (2010). A Textual-based Similarity Approach for Efficient and scalable External Plagiarism Analysis. Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Mihalcea, R. (2007). Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL 2007)*, Rochester. ACL.

Milićević, J. (2007). *La paraphrase*. Peter Lang, Bern.

Mohammadi, M. and GhasemAghaee, N. (2010). Building Bilingual Parallel Corpora based on Wikipedia. In *Second International Conference on Computer Engineering and Applications*, volume 2, pages 264–268.

Monostori, K., Finkel, R., Zaslavsky, A., and Hodász, G. (2002). Comparison of Overlap Detection Techniques. *Computational Science. Proceedings of the International Conference on Computational Science*, **LNCS (2329)**, 51–60. Springer-Verlag.

Mora, M. (2000). Ana Rosa Quintana culpa de los plagios en 'Sabor a hiel' a un ayudante de toda confianza (Ana Rosa Quintana blames a trustworthy assistant for the plagiarism in 'Sabor a hiel'). El País [http://bit.ly/pais_quintana], Spain. Published: 23/Oct/2000; Accessed: 14/Oct/2011.

Morgan, C. and Foster, W. (1992). Student Cheating: An Ethical Dilemma. In L. P. Grayson, editor, *Proceedings of the Frontiers in Education Conference*, pages 678–682, Nashville, TE. IEEE.

Muhr, M., Kern, R., Zechner, M., and Granitzer, M. (2010). External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System. In Braschler and Harman (2010).

Munteanu, D., Fraser, A., and Marcu, D. (2004). Improved Machine Translation Performace via Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*, Boston, MA.

Nawab, R., Stevenson, M., and Clough, P. (2010). University of Sheffield Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Nawab, R., Stevenson, M., and Clough, P. (2011). External Plagiarism Detection using Information Retrieval and Sequence Alignment - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Nelken, R. and Yamangil, E. (2008). Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. In *AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, Chicago, Il. AAAI.

Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D., Hiemstra, D., and de Jong, F. (2009). WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. *Proceedings of the Cross-Language Evaluation Forum*, **LNCS (5706)**. Springer-Verlag.

Oberreuter, G., L'Huillier, G., Ríos, S., and Velásquez, J. (2010). FASTDOCODE: Finding Approximated Segments of N-Grams for Document Copy Detection. Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Oberreuter, G., L'Huillier, G., Ríos, S., and Velásquez, J. (2011). Approaches for Intrinsic and External Plagiarism Detection: Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Och, F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1), 19–51. See also [`http://www.fjoch.com/GIZA++.html`].

Olsson, J. (2008). *Forensic Linguistics. An Introduction to Language, Crime and the Law*. Continuum International Publishing Group, New York, NY.

Ortega Soto, J. (2009). *Wikipedia: A Quantitative Analysis*. PhD thesis, Universidad Rey Juan Carlos, Madrid, Spain.

Ottenstein, K. (1976). An Algorithmic Approach to the Detection and Prevention of Plagiarism. *ACM SIGCSE Bulletin*, **8**(4), 30–41.

Palkovskii, Y. (2009). "Counter Plagiarism Detection Software" and "Counter Counter Plagiarism Detection" Methods. In Stein *et al.* (2009), pages 67–68. http://ceur-ws.org/Vol-502.

Palkovskii, Y. and Belov, A. (2011). Exploring Cross Lingual Plagiarism Detection in Hindi-English with n-gram Fingerprinting and VSM based Similarity Detection. In FIRE (2011).

Palkovskii, Y., Belov, A., and Muzika, I. (2010). Exploring Fingerprinting as External Plagiarism Detection Method Lab Report for PAN at CLEF 2010. In Braschler and Harman (2010).

Palkovskii, Y., Belov, A., and Muzika, I. (2011). Using WordNet-based Semantic Similarity Measurement in External Plagiarism Detection - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Paramita, M., Clough, P., Aker, A., and Gaizauskas, R. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. to appear.

Park, C. (2003). In Other (People's) Words: Plagiarism by University Students–Literature and Lessons. *Assessment & Evaluation in Higher Education*, **28**(5), 471–488. Carfax Publishing.

Pataki, M. (2006). Distributed Similarity and Plagiarism Search. In *Proceedings of the Automation and Applied Computer Science Workshop*, Budapest, Hungary.

Patry, A. and Langlais, P. (2011). Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In P. Zweigenbaum, R. Rapp, and S. Sharoff, editors, *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon. Association for Computational Linguistics.

Petras, V., Forner, P., and Clough, P., editors (2011). *Notebook Papers of CLEF 2011 LABs and Workshops*, Amsterdam, The Netherlands.

Piao, S. (2001). Detecting and Measuring Text Reuse via Aligning Texts. Research Memorandum CS-01-15, Department of Computer Science. University of Sheffield, UK.

Pinto, D., Jiménez-Salazar, H., and Rosso, P. (2006). Clustering Abstracts of Scientific Texts Using the Transition Point Technique. *Computational Linguistics and Intelligent Text Processing, 7th International Conference*, **LNCS (3878)**, 536–546. Springer-Verlag.

Pinto, D., Benedí, J., and Rosso, P. (2007). Clustering Narrow-Domain Short Texts by Using the Kullback-Leibler Distance. *Computational Linguistics and Intelligent Text Processing, 8th International Conference*, **LNCS (4394)**, 611–622. Springer-Verlag.

Pinto, D., Civera, J., Barrón-Cedeño, A., Juan, A., and Rosso, P. (2009). A Statistical Approach to Crosslingual Natural Language Tasks. *Journal of Algorithms*, **64**(1), 51–60.

Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2011). A Self-Enriching Methodology for Clustering Narrow Domain Short Texts. *The Computer Journal*, **54**(7), 1148–1165.

Piron, A. (1846). *La métromanie*. Libraire de Firmin Didot Frères. Available at `http://bit.ly/piron_metromanie`.

Plagiarism Detect (2011). PlagiarismDetect.com. [`http://www.plagiarismdetect.com`]. Accessed 16/Aug/2011.

Plagiarism.org (2011). What is plagiarism? [`http://bit.ly/plagiarism_org`]. Accessed 22/Aug/2011.

Plagscan (2011). Plagscan The online service for plagiarism detection . [`http://www.plagscan.com`]. Accessed 16/Aug/2011.

Potthast, M. and Holfeld, T. (2011). Overview of the 2nd International Competition on Wikipedia Vandalism Detection. In Petras *et al.* (2011).

Potthast, M. and Stein, B. (2008). New Issues in Near-Duplicate Detection. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, editors, *Data Analysis, Machine Learning and Applications*, pages 601—609, Berlin Heidelberg New York. Springer-Verlag.

Potthast, M., Stein, B., and Anderka, M. (2008a). A Wikipedia-Based Multilingual Retrieval Model. *Advances in Information Retrieval, 30th European Conference on IR Research*, **LNCS (4956)**, 522–530. Springer-Verlag.

Potthast, M., Stein, B., and Gerling, R. (2008b). Automatic Vandalism Detection in Wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, editors, *ECIR 2008*, volume LNCS (4956), pages 663–668, Glasgow, UK. Springer-Verlag.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P. (2009). Overview of the 1st International Competition on Plagiarism Detection. In Stein *et al.* (2009), pages 1–9. http://ceur-ws.org/Vol-502.

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010a). An Evaluation Framework for Plagiarism Detection. In Huang and Jurafsky (2010), pages 997–1005.

Potthast, M., Trenkmann, M., and Stein, B. (2010b). Netspeak: Assisting Writers in Choosing Words. *Advances in Information Retrieval. 32nd European Conference on Information Retrieval (ECIR 10)*, **LNCS (5993)**, 672. Springer-Verlag.

Potthast, M., Stein, B., and Holfeld, T. (2010c). Overview of the 1st International Competition on Wikipedia Vandalism Detection. In Braschler and Harman (2010).

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., and Rosso, P. (2010d). Overview of the 2nd International Competition on Plagiarism Detection. In Braschler and Harman (2010).

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011a). Cross-Language Plagiarism Detection. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, **45**(1), 1–18.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011b). Overview of the 3rd International Competition on Plagiarism Detection. In Petras *et al.* (2011).

Pouliquen, B., Steinberger, R., and Ignat, C. (2003). Automatic Identification of Document Translations in Large Multilingual Document Collections. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 401–408, Borovets, Bulgaria.

Project Gutenberg (2011). Project Gutenberg Website. [`http://www.gutenberg.org`].

Público (2008). Enrique Bunbury: "Dos frases no hacen un plagio" (Enrique Bunbury: "Two phrases are not plagiarism"). Público [`http://bit.ly/publico_bunbury`], Spain. Published: 10/Sep/2008; Accessed: 14/Oct/2011.

Pupovac, V., Bilić-Zulle, L., and Petrovečki, M. (2008). On Academic Plagiarism in Europe. An Analytical Approach based on Four Studies. In Comas and Sureda (2008a), pages 13–18. [`http://bit.ly/cyberplagiarism_cs`].

Quirck, C., Brockett, C., and Dolan, W. (2004). Monolingual Machine Translation for Paraphrase Generation. In Lin and Wu (2004), pages 142–149.

R. Costa-jussà, M., Banchs, R., Grivolla, J., and Codina, J. (2010). Plagiarism Detection Using Information Retrieval and Similarity Measures based on Image Processing Techniques. In Braschler and Harman (2010).

Rabin, M. (1981). Fingerprinting by Random Polynomials. Technical Report TR-CSE-03-01, Center for Research in Computing Technology, Harvard University, Cambridge, MA.

Rafalovitch, A. and Dale, R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In *Proceedings of the MT Summit XII*, pages 292–299. International Association of Machine Translation.

Rambhoopal, K. and Varma, V. (2011). Cross-Lingual Text Reuse Detection Based On Keyphrase Extraction and Similarity Measures. In FIRE (2011).

Rao, S., Gupta, P., Singhal, K., and Majumdar, P. (2010). External & Intrinsic Plagiarism Detection: VSM & Discourse Markers based Approach. Notebook for PAN at CLEF 2011. In Braschler and Harman (2010).

Recasens, M. and Vila, M. (2010). On Paraphrase and Coreference. *Computational Linguistics*, **36**(4), 639–647.

Richman, A. (2008). Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics with the Human Language Technology Conference of the North American Chapter of the ACL (ACL/HLT 2008)*, pages 1—9, Columbus, OH.

Ríos, P. (2009). La juez ve plagio en 'La Cruz de San Andrés' de Cela. El País [`http://bit.ly/pais_cela_09`], Spain. Published: 21/Apr/2009; Accessed: 22/Aug/2011.

Ríos, P. (2010). El presunto plagio de Cela irá a juicio. El País [`http://bit.ly/pais_cela_10`], Spain. Published: 17/Oct/2010; Accessed: 22/Aug/2011.

Rivera, A. (2011). Ciencia china 'duplicada' en Galicia (Chinese Science 'duplicated' in Galicia). El País [`http://bit.ly/pais_china_galicia`], Spain. Published: 20/May/2011; Accessed: 25/May/2011.

Rivest, R. (1992). The MD5 Message-Digest Algorithm. RFC 1321 (Informational) [`http://www.ietf.org/rfc/rfc1321.txt`]. Updated by RFC 6151.

Rodríguez, E. (2000a). Ana Rosa Quintana culpa del plagio a un estrecho colaborador (Ana Rosa Quintana blames a collaborator for the plagiarism). El Mundo [`http://mun.do/mundo_quintana2`], Spain. Published: 23/Oct/2000; Accessed: 14/Oct/2011.

Rodríguez, E. (2000b). Planeta crea un precedente al retirar el libro que "escribió" Ana Rosa Quintana (Planeta makes precedent at retiring the book that Ana Rosa Quintana "wrote"). El Mundo [`http://mun.do/mundo_quintana`], Spain. Published: 16/Oct/2000; Accessed: 14/Oct/2011.

Rodríguez Torrejón, D. and Martín Ramos, J. (2010). CoReMo System (Contextual Reference Monotony). In Braschler and Harman (2010).

Rodríguez Torrejón, D. and Martín Ramos, J. (2011). Crosslingual CoReMo System - Notebook for PAN at CLEF 2011. In Petras *et al.* (2011).

Ruml, B., editor (1952). *American Bar Association Journal*, volume 38 (7). American Bar Association, Chicago, Il. Available at [`http://bit.ly/aba_ruml`] Accessed: 24/Aug/2011.

Runeson, P., Alexandersson, M., and Nyholm, O. (2007). Detection of Duplicate Defect Reports Using Natural Language Processing. In *Proceedings of the 29th International Conference on Software Engineering*. IEEE Computer Society.

Ryu, C., Kim, H., Ji, S., Gyun, W., and Hwan-Gue, C. (2008). Detecting and Tracing Plagiarized Documents by Reconstruction Plagiarism-Evolution Tree. In *Proceedings of the 8th IEEE International Conference on Computer and Information Technology*, pages 119–124. IEEE Computer Society.

Samuelson, P. (1994). Self-Plagiarism or Fair Use. *Communications of the ACM*, **37**(8), 21–25.

Sánchez-Vega, F., Villaseñor-Pineda, L., Montes-y Gómez, M., and Rosso, P. (2010). Towards Document Plagiarism Detection based on the Relevance and Fragmentation of the Reused Text. *Proceedings of the 9th Mexican International Conference on Artificial Intelligence (MICAI 2010)*, **LNAI (6437)**, 24–31. Springer-Verlag.

Scherbinin, V. and Butakov, S. (2009). Using Microsoft SQL Server Platform for Plagiarism Detection. In Stein *et al.* (2009), pages 36–37. http://ceur-ws.org/Vol-502.

Schleimer, S., Wilkerson, D., and Aiken, A. (2003). Winnowing: Local Algorithms for Document Fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, San Diego, CA. ACM.

Seaward, L. and Matwin, S. (2009). Intrinsic Plagiarism Detection Using Complexity Analysis. In Stein *et al.* (2009), pages 56–61. http://ceur-ws.org/Vol-502.

Shimohata, M. (2004). *Acquiring Paraphrases from Corpora and Its Application to Machine Translation*. Ph.D. thesis, Nara Institute of Science and Technology, Nara, Japan.

Shivakumar, N. and García-Molina, H. (1995). SCAM: A Copy Detection Mechanism for Digital Documents. In *Proceedings of the 2nd Annual Conference on the Theory and Practice of Digital Libraries*.

Si, A., Leong, H., and Lau, R. (1997). CHECK: A Document Plagiarism Detection System. In *Proceedings of the 1997 ACM Symposium on Applied Computing*, pages 70–77, San Jose, CA.

Sichel, H. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, **70**(351), 542–547.

Sidorov, G., Barrón-Cedeño, A., and Rosso, P. (2010). English-Spanish Large Statistical Dictionary of Inflectional Forms. In Calzolari *et al.* (2010).

Silvestre-Cerdà, J., García-Martínez, M., Barrón-Cedeño, A., and Rosso, P. (2011). Extracción de corpus paralelos de la Wikipedia basada en la obtención de alineamientos bilingües a nivel de frase. In A. Barrón-Cedeño, J. Civera, P. Rosso, M. Vila, A. Barreiro, and I. Alegria, editors, *SEPLN-ICL: Workshop on Iberian Cross-Language NLP Tasks*, Huelva, Spain.

Simard, M., Foster, G. F., and Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Six Degrés, Compilatio.net, and Le Sphinx Développement (2008). Los usos de Internet en la educación superior: de la documentación al plagio. Technical report, Six Degrés.

SkyLine, Inc. (2011). Plagiarism Detector. [`http://www.plagiarism-detector.com`]. Accessed 17/Aug/2011.

Somers, H., Gaspari, F., and Niño, A. (2006). Detecting Inappropriate Use of Free Online Machine Translation by Language Students – A Special Case of Plagiarism Detection. In *Proceedings of the Eleventh Annual Conference of the European Association for Machine Translation*, pages 41–48, Oslo, Norway.

Sorokina, D., Gehrke, J., Warner, S., and Ginsparg, P. (2006). Plagiarism Detection in arXiv. In *Proceedings of the 6th International Conference on Data Mining*.

Spärck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, **28**(1), 11–21.

Stamatatos, E. (2009a). A Survey of Modern Authorship Attribution. *Journal of the American Society for Information Science and Technology*, **60**(3), 538–556.

Stamatatos, E. (2009b). Intrinsic Plagiarism Detection Using Character n-gram Profiles. In Stein *et al.* (2009), pages 38–46. http://ceur-ws.org/Vol-502.

Stamatatos, E. (2011). Plagiarism Detection Based on Structural Information. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2011)*, Glasgow, Scotland. ACM Press.

Stein, B. and Meyer zu Eissen, S. (2007). Intrinsic Plagiarism Analysis with Meta Learning. In B. Stein, E. Stamatatos, and M. Koppel, editors, *SIGIR 2007 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*, pages 45–50, Amsterdam, The Netherlands.

Stein, B. and Potthast, M. (2007). Applying Hash-based Indexing in Text-Based Information Retrieval. In M. Moens, T. Tuytelaars, and A. de Vries, editors, *Proceedings of the 7th Dutch-Belgian Information Retrieval Workshop (DIR 07)*, pages 29–35, Leuven, Belgium. Faculty of Engineering, Universiteit Leuven.

Stein, B., Meyer zu Eissen, S., and Potthast, M. (2007). Strategies for Retrieving Plagiarized Documents. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 825–826, Amsterdam, The Netherlands. ACM Press.

Stein, B., Stamatatos, E., and Koppel, M., editors (2008). *ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008)*, volume 377, Patras, Greece. CEUR-WS.org. http://ceur-ws.org/Vol-377.

Stein, B., Rosso, P., Stamatatos, E., Koppel, M., and Agirre, E., editors (2009). *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009)*, volume 502, San Sebastian, Spain. CEUR-WS.org. http://ceur-ws.org/Vol-502.

Stein, B., Potthast, M., Rosso, P., Barrón-Cedeño, A., Stamatatos, E., and Koppel, M. (2011a). Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. *ACM SIGIR Forum*, **45**, 45–48.

Stein, B., Lipka, N., and Prettenhofer, P. (2011b). Intrinsic Plagiarism Analysis. *Language Resources and Evaluation (LRE), Special Issue on Plagiarism and Authorship Analysis*, **45**, 63–82.

Steinberger, R., Pouliquen, B., and Hagman, J. (2002). Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Computational Linguistics and Intelligent Text Processing. Proceedings of the CICLing 2002*, **LNCS (2276)**, 415—424. Springer-Verlag.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In Calzolari *et al.* (2006).

Stolcke, A. (2002). SRILM - An Extensible Language Modeling toolkit. In *Intl. Conference on Spoken Language Processing*, Denver, Colorado.

Suárez, P., González, C., and Villena-Román, J. (2010). A plagiarism detector for intrinsic plagiarism. In Braschler and Harman (2010).

Sureda, J. and Comas, R. (2008). El plagio y otras formas de deshonestidad académica entre el alumnado universitario. Resultados generales de los datos de una encuesta realizada a los usuarios del portal Universia. Technical report, Grupo Educación y Ciudadanía. Palma: Universitat de les Illes Balears. Departamento de Pedagogía Aplicada y Psicología de la Educación, Spain. [`http://bit.ly/cyberplagio_deshonestidad`] Accessed Aug/2011.

Sureda, J., Comas, R., and Morey, M. (2008). Cyberplagiarism Webliography. References to Academic Cyberplagiarism on the Internet. In Comas and Sureda (2008a), pages 29–39. [`http://bit.ly/cyberplagiarism_cs`].

Suri, H. (2007). Evaluation of Two Turnitin Trials in the Faculty of Law. Technical report, Monash University, Australia.

Talmy, L. (1985). Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen, editor, *Language Typology and Semantic Description. Grammatical Categories and the Lexicon*, volume III, chapter II, pages 57–149. University of Cambridge.

Taylor, F. (1965). Cryptomnesia and Plagiarism. *The British Journal of Psychiatry*, **111**, 1111–1118.

The Martin Luther King, Jr. Research and Education Center (2011). The Martin Luther King, Jr. Research and Education Center. [`http://bit.ly/stanford_king`]. (12 Oct 2011).

The New York Times (1991). Boston U. Panel Finds Plagiarism by Dr. King. The New York Times [`http://nyti.ms/nytimes_king`], USA. Published: 11/Oct/1991; Accessed: 11/Oct/2011.

Tilstone, W., Savage, K., and Clarck, L. (2006). *Encyclopedia of Forensic Science. An Encyclopedia of History, Methods and Techniques*. ABC-CLIO Inc., Santa Barbara, CA.

Toral, A. and Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia. In *Proceedings of the EACL Workshop on New Text 2006*, Trento, Italy. Association for Computational Linguistics.

Turell, M. (2011). La tasca del lingüista detectiu en casos de detecció de plagi i determinació d'autoria de textos escrits (The Work of the Detective Linguist in Cases of Plagiarism Detection and Authorship Attribution in Written Texts). *Llengua, societat i comunicació*, **9**, 69–85.

Turell, M. and Coulthard, M. (2011). Forensic Plagiarism Detection and Authorship Attribution: On the Linguists' Achievements and the Challenges for Computerized Analysis. Keynote at Lab PAN: Uncovering Plagiarism, Authorship and Social Software Misuse [`http://www.webis.de/research/events/pan-11`].

Turnitin (2010). iParadigms Licenses Language Weaver's Automated Translation Technology to Extend Plagiarism Prevention Capabilities. [`http://bit.ly/turnitin_cl`]. Published: 5/Jul/2010.

Urkund (2011). Urkund. [`http://www.urkund.com`]. Accessed 16/Aug/2011.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In Huang and Jurafsky (2010), pages 1101–1109.

Vallés Balaguer, E. (2009). Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind Tool. In Stein *et al.* (2009), pages 34–35. http://ceur-ws.org/Vol-502.

Vania, C. and Adriani, M. (2010). Automatic External Plagiarism Detection using Passage Similarities. In Braschler and Harman (2010).

Vargas Aignasse, G. (2010). Proyecto de ley. Reforma codigo penal: plagio. Agravamiento de penas (Draft law. Criminal Code Reform: plagiarism. Aggravation of punishment). H. Cámara de Diputados de la Nación [`http://bit.ly/proyecto_vargas`], Argentina. Published: 6/May/2010; Accessed: 13/Oct/2011.

Vaughan, L. (2001). *Statistical Methods for the Information Professional*. American Society for Information Science and Technology, Canada.

Vila, M., Martí, M., and Rodríguez, H. (2011). Paraphrase Concept and Typology. A Linguistically Based and Computationally Oriented Approach. *Procesamiento del Lenguaje Natural*, **46**, 83–90.

Vinokourov, A., Shawe-Taylor, J., and Cristianini, N. (2003). Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems (NIPS-02)*, pages 1473–1480. MIT Press.

Weber, S. (2007). *Das Google-Copy-Paste-Syndrom. Wie Netzplagiate Ausbildung und Wissen gefahrden*. Telepolis.

Wikipedia (2008). Gunning fog index. [http://bit.ly/wikipedia_gunning]. Accessed 19/Oct/2008.

Wikipedia (2010a). Basque language. [http://bit.ly/wikipedia_basque]. Accessed 5/Feb/2010.

Wikipedia (2010b). Party of European Socialists | Partido Socialista Europeo | Europako Alderdi Sozialista. [http://bit.ly/wikipedia_socialists]. Accessed 10/Feb/2010.

Wikipedia (2011a). Ana Rosa Quintana. [http://bit.ly/wikipedia_quintana]. Accessed 14/Oct/2011.

Wikipedia (2011b). Flesch–Kincaid readability test. [http://bit.ly/wikipedia_flesch]. Accessed 4/Jul/2011.

Wikipedia (2011c). Forensic science. [http://bit.ly/wikipedia_forensic]. Accessed 3/Oct/2011.

Wikipedia (2011d). Gerónimo Vargas Aignasse. [http://bit.ly/wikipedia_vargas]. Accessed 13/Oct/2011.

Wikipedia (2011e). Gerónimo Vargas Aignasse. [http://bit.ly/wikipedia_es_vargas]. Accessed 13/Oct/2011.

Wikipedia (2011f). How Opal Mehta Got Kissed, Got Wild, and Got a Life. [http://bit.ly/wikipedia_howopal]. Accessed 13/Oct/2011.

Wikipedia (2011g). Joe Biden. [http://bit.ly/wikipedia_biden]. Accessed 13/Oct/2011.

Wikipedia (2011h). Karl-Theodor zu Guttenberg. [http://bit.ly/wikipedia_zuguttenberg]. Accessed 12/Oct/2011.

Wikipedia (2011i). Legal disputes over the Harry Potter series. [http://bit.ly/wikipedia_potter]. Accessed 22/Aug/2011.

Wikipedia (2011j). Martin Luther King, Jr. authorship issues]. [http://bit.ly/wikipedia_lking]. Accessed 12/Oct/2011.

Wikipedia (2011k). My Sweet Lord. [http://bit.ly/wikipedia_sweet]. Accessed 22/Aug/2011.

Wikipedia (2011l). Plagiarism. [http://bit.ly/wikipedia_plagiarism]. Accessed 28/Jun/2011.

Wikipedia (2011m). Plagio. [http://bit.ly/wikipedia_plagio]. Accessed 13/Oct/2011.

Wikipedia (2011n). Ruth Shalit. [http://bit.ly/wikipedia_shalit]. Accessed 12/Oct/2011.

Wikipedia (2011o). The Chiffons. [http://bit.ly/wikipedia_chiffons]. Accessed 22/Aug/2011.

Wikipedia (2011p). Wikipedia. [http://bit.ly/es_wikipedia]. Accessed 27/Jul/2011.

Wikipedia (2011q). Wikipedia: Five pillars. [http://bit.ly/wikipedia_pillars]. Accessed 6/Dec/2011.

Wilks, Y. (2004). On the ownership of text. *Computers and the Humanities*, **38**, 115–127.

Wise, M. (1993). Running Karp-Rabin Matching and Greedy String Tiling. Technical Report 463, The University of Sydney, Sidney, Australia.

Wood, G. (2004). Academic Original Sin: Plagiarism, the Internet and Librarians. *The Journal of Academic Librarianship*, **30**, 237–242.

Yasuda, K. and Sumita, E. (2008). Method for Building Sentence-Aligned Corpus from Wikipedia. In *Association for the Advancement of Artificial Intelligence*.

Young, E. (1759). *Conjectures on Original Composition. In a Letter to the Author of Sir Charles Grandison*. A. Millar and R. and J. Dodsley, Great Britain. Available at [http://bit.ly/young_conjectures] (August 2011).

Yule, G. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zaka, B. (2009). Empowering Plagiarism Detection with a Web Services Enabled Collaborative Network. *Journal of Information Science and Engineering*, **25**, 1391–1403.

Zechner, M., Muhr, M., Kern, R., and Granitzer, M. (2009). External and Intrinsic Plagiarism Detection Using Vector Space Models. In Stein *et al.* (2009), pages 47–55. http://ceur-ws.org/Vol-502.

Zeng, H., Alhossaini, M., Fikes, R., and McGuinness, D. (2006). Mining Revision History to Assess Trustworthiness of Article Fragments. In *Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 1–10.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wikictionary. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zhang, H. and Chow, T. W. (2011). A Coarse-to-Fine Framework to Efficiently Thwart Plagiarism. *Pattern Recognition*, **44**, 471–487.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining and Knowledge Discovery*, **10**, 141–168.

Zhou, D. (2006). Student's Novel Faces Plagiarism Controversy. The Harvard Crimson [http://www.thecrimson.com/article/2006/4/23/students-novel-faces-plagiarism-controversy-beditors/], USA. Published: 23/Apr/2006; Accessed: 13 Oct, 2011.

Ziv, J. and Lempel, A. (1977). A Universal Algorithm for Sequential Data Compression. *IEEE Transactions on Information Theory*, **IT-24**(5).

Zou, D., Long, W.-j., and Ling, Z. (2010). A Cluster-based Plagiarism Detection Method. In Braschler and Harman (2010).

Appendix A

Generation of Dictionaries for CL-ASA

CL-ASA, our cross-language similarity assessment model, relies on a statistical bilingual dictionary (cf. Section 6.3). In this appendix we describe the two strategies we followed to generate such a dictionary. Section A.1 describes how we empirically built a dictionary, obtained from parallel data. Section A.2 describes the process we followed to obtain a dictionary from a "traditional" (lexicographic) one.

A.1 Dictionary Built from Parallel Corpora

In this section we describe the statistical model and the Expectation-Maximisation (EM) method for the estimation of the probabilities of the bilingual dictionary. It is possible to derive an EM algorithm to perform the maximum likelihood estimation of the statistical dictionary with respect to a collection of training samples $(X, Y) = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. In Chapter 6, such collection was composed of documents from the JRC-Acquis corpus for our experiments between English-{Dutch, French, German, Polish, Spanish}. The same corpus was used for our cross-language experiments in Chapter 7, for the pair English-Spanish. For our experiments with Basque-{English, Spanish} we used two parallel corpora: *Software*, an English-Basque memory of software manuals and *Consumer*, a corpus extracted from a consumer oriented magazine that includes articles written in Spanish along with their Basque, Catalan, and Galician translations (cf. Section 6.5).

The *(incomplete)* log-likelihood function is:

$$L(\vec{\Theta}) = \sum_{n=1}^{N} \log \sum_{a_n} p(y_n, a_n | x_n) \ , \tag{A.1}$$

with

$$p(y_n, a_n | x_n) = \frac{1}{(|x_n|+1)^{|y_n|}} \prod_{i=1}^{|y_n|} \prod_{j=0}^{|x_n|} p(y_{ni} \mid x_{nj})^{a_{nij}} \ , \tag{A.2}$$

where, for convenience, the alignment variable, $a_{ni} \in \{0, 1, \ldots, |x_n|\}$, has been rewritten as an indicator vector, $a_{ni} = (a_{ni0}, \ldots, a_{ni|x_n|})$, with 1 in the suspicious fragment position to which it is connected, and zeros elsewhere.

The so-called *complete* version of the log-likelihood function (A.1) assumes that the hidden (missing) alignments $a_1, \ldots, a_N$ are also known:

$$\mathcal{L}(\vec{\Theta}) = \sum_{n=1}^{N} \log p(y_n, a_n | x_n) \ . \tag{A.3}$$

An initial estimate for $\vec{\Theta}$, $\vec{\Theta}^{(0)}$, is required for the EM algorithm to start. This can be done by assuming that the translation probabilities are uniformly distributed; i.e.,

$$p(w \,|\, v)^{(0)} = \frac{1}{|\mathcal{Y}|} \ \forall v \in \mathcal{X}, w \in \mathcal{Y} \ . \tag{A.4}$$

After this initialisation, the EM algorithm maximises (A.1) iteratively, through the application of two basic steps in each iteration: the E(xpectation) step and the M(aximisation) step. At iteration k, the E step computes the expected value of (A.3) given the observed (incomplete) data, (X, Y), and a current estimation of the parameters, $\vec{\Theta}^{(k)}$. This reduces to the computation of the expected value of a_{nij}:

$$a_{nij}^{(k)} = \frac{p(y_{ni} \,|\, x_{nj})^{(k)}}{\sum_{j'} p(y_{ni} \,|\, x_{nj'})^{(k)}} \ . \tag{A.5}$$

Then, the M step finds a new estimate of $\vec{\Theta}$, $\vec{\Theta}^{(k+1)}$, by maximising (A.3), using (A.5) instead of the missing a_{nji}. This results in:

$$P(w|v)^{(k+1)} = \frac{\sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w) \, \delta(x_{nj}, v)}{\sum_{w'} \sum_n \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} a_{nij}^{(k)} \, \delta(y_{ni}, w') \, \delta(x_{nj}, v)} \tag{A.6}$$

$$= \frac{\sum_n \frac{p(w \,|\, v)^{(k)}}{\sum_{j'} p(w \,|\, x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w) \, \delta(x_{nj}, v)}{\sum_{w'} \left[\sum_n \frac{p(w' \,|\, v)^{(k)}}{\sum_{j'} p(w' \,|\, x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w') \, \delta(x_{nj}, v) \right]} \ , \tag{A.7}$$

for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}$; where $\delta(a, b)$ is the Kronecker delta function; i.e., $\delta(a, b) = 1$ if $a = b$; 0 otherwise. In order to do this computation we used Giza++ (Och and Ney, 2003).[1]

An example of the dictionary's entries is included in Table A.1 for Spanish-English. Note that given the Spanish word *tomó* (took), the most logical entries are *took*, *taken* and *take*, but many other potential translations are considered, though with low probabilities. Indeed, our experiments of Sections 6.4, 6.5, and 7.5 we experimentally show that

[1] `http://code.google.com/p/giza-pp/`

considering reduced probability masses filters most of these noisy entries. For instance, if we considered a probability mass of 30%, only *took* and *taken* would be considered as potential translations for *tomó*, as $p(\text{took}|\text{tomó}) + p(\text{taken}|\text{tomó}) > 0.30$.

Table A.1: Example entries in the empirically built dictionary for Spanish-English. We consider the entries for the Spanish word *tomó* (took).

translation	$p(w'\,|\text{tomó})$	**translation**	$p(w'\,|\text{tomó})$
took	0.27797	proposed	0.02138
taken	0.14961	take	0.02138
was	0.14908	look	0.02138
noted	0.08473	failed	0.02138
has	0.06409	adopted	0.02138
it	0.03916	eu's	0.02135
adopt	0.02138	cooperation	0.02130
consequently	0.02138	council	0.00028
falls	0.02138	steering	0.00000
fifth	0.02138		

A.2 Dictionary Built from Lexicographic Data

In a bilingual dictionary, a word w in a language L is linked to all its potential translations w' in a language L'. In a traditional bilingual dictionary, w and w' are usually lemmas, i.e., a morphologically normalised word form. Its translation very often is also a lemma, or a set of possible lemmas. This is a typical situation, see below the discussion of more complex situations when the translation is a multi-word expression. In this section we describe (i) how we generated a bilingual dictionary that includes a complete variation of words inflections, i.e., all possible word forms for each lemma for languages L and L' (though any pair of languages can be considered, in this case we considered L = English and L' = Spanish) and (ii) how we estimated the translation probability for each words pair, on the basis of monolingual frequencies of grammar classes in large corpora (Sidorov *et al.*, 2010).

The typical situation in a bilingual dictionary is the presence of a head word (lemma) in L and one or several translation equivalents (lemmas) in L'. Sometimes, the situation is more complex when the translation equivalents are represented by a combination of words. A question arises for our task: how a word that is not a head word should be treated in the word combinations? That is, should they be considered also as possible translation equivalents? In some specialised dictionaries, like terminological dictionaries, even a head word can be represented as a multi-word expression, for example, *concept album - disco monográfico*. The simplest solution that we adapt in this case is the usage of some heuristics or partial syntactic analysis for determining the syntactic structure of the word combination and then processing only the top head word. Translations of the head word often are lemmas as well. Nevertheless, in this case it is much more frequent having translation equivalents represented as multi-word expressions. The same considerations as above are applied. For the moment, we use just the top head word

Algorithm. Input: $Dict_{en-es}$

Initialise the set $T_{en,es}$
for each pair $\{en, es\} \in Dict_{en-es}$
    $en_l = lemma(en)$;
    $es_l = lemma(es)$
    $F[en_l] \leftarrow word\_forms(en_l, English)$
    $F[es_l] \leftarrow word\_forms(es_l, Spanish)$
    Add $F[en_l] \times Fes_l$ to $T_{en,es}$
Return: $T_{en,es}$

Figure A.1: Morphological generation algorithm. $T_{en,es}$ = set of generated translation pairs; $Dict_{en-es}$ = input bilingual dictionary; $lemma(x)$ obtains the lemma for word x; and $word\_forms(x)$ generates all word forms for the lemma x.

(*nucleus*) of the multi-word expression.

Generally speaking, translation equivalents can be either a generalisation, or, more often, a specification of the translated word. This specification can be either (*i*) a set of adjectives that depend on the head word; (*ii*) a multi-word expression where the translation equivalent is a lemma and the depending words have morphological forms that correspond to its government pattern; or (*iii*) a subordinate clause. It is desirable to treat somehow the dependant words because they represent part of the meaning of the word in the other language. However, they cannot be treated in the same way as the head word because these words are not translation equivalents of the head word in the other language but only specifiers.

We developed a corresponding algorithm for the pair of languages {English, Spanish}. The algorithm is divided in two main steps: (*i*) morphological generation: creation of a complete list of word forms for a list of translation equivalents in each language; and (*ii*) calculation of translation probabilities: estimation of the probabilities $p(w' \mid w)$ for all $w' \in L'$, $w \in L$. As a word form can correspond to various lemmas it has several sets of possible inflectional correspondences in the other language.

The morphological generation step is based on a list of bilingual correspondences. Its source is a traditional bilingual dictionary containing about 30,000 entry words and including around 64,000 translations. In order to generate the English and Spanish word forms we used the morphological dictionaries available in the FreeLing package (Atserias, Casas, Comelles, Gonzáles, Padró, and Padró, 2006). The idea was to consider not only those pairs included in a traditional translation dictionary, but also all the possible inflectional forms of each pair of words "source word–translation word(s)". The generation process is summarised in Fig. A.1. An example of the list of inflectional forms obtained for a word form in Spanish is presented in Table A.4. It includes a word form of the English verb *to take*, in this case *took*, with its valid translations into Spanish word forms.

A problem arises how to assign the probability for each translation $p(w', w)$. We use the idea that the probability of a word form is proportional to the distribution of the corresponding grammar sets in a large corpus. We use the term *grammar set* as part of a complete grammar paradigm for a given lemma. We consider that a paradigm is a well-structured table where all word forms can be placed, and grammar set characterises each

Table A.2: Distribution of English grammar classes.

freq	grammar	freq	grammar	freq	grammar
163935	NN	26436	VBZ	3087	NNPS
121903	IN	24865	VBN	2887	WP
114053	NNP	21357	PRP	2625	WRB
101190	DT	18239	VBG	2396	JJS
75266	JJ	15377	VBP	2175	RBR
73964	NNS	11997	MD	555	RBS
38197	RB	10801	POS	441	PDT
37493	VBD	10241	PRP$	219	WP$
32565	VB	4042	JJR	117	UH
29462	CC	3275	RP		

cell of this table. In this case, for example, *take* as a noun has two possible grammar sets (*Singular* and *Plural*), and *take* as a verb has at least four grammar sets that correspond to *take, takes, took, taken*. The exact number of grammar sets depends on how many cells we postulate for a verb in its paradigm for English language. An important point here is that we count probabilities for *take* as a noun and *take* as a verb separately and independently, because they have different grammar paradigms.

We considered frequencies of grammar sets for English and Spanish. The frequency distribution of English grammar sets was estimated by considering a version of the WSJ corpus (cf. Table A.2).[2] The frequency distribution of Spanish grammar sets was computed using a corpus marked with grammar information (cf. Table A.3).[3] The English and Spanish corpora contain about 950,000 and 5.5 million word forms, respectively; a sufficient amount of words for our purposes. The frequencies included in Tables A.2 and A.3 give us the possibility to assign probabilities to word forms according to the proportion of their grammar sets (grammar information) in the corpora.

Though in theory a word form w can be translated by any word form w' with some probability, in most of the cases, these translations are highly improbable. In other words, *a priori* not every w can be likely translated into any w'. In order to estimate such probability we use a similarity measure between grammar classes in languages L and L'. For example, a noun in singular is more likely to be translated into a noun in singular than in plural. It is not expected that a verb in present tense would be translated into a verb in past tense. In order to calculate this similarity measure we developed an algorithm for our specific language pair, though the majority of its steps and conditions are rather universal. Indeed, the algorithm is applied to the language pair where Spanish has relatively rich morphology, while English has a relatively poor morphological system. Therefore, we consider that the algorithm is rather universal and can be applied to any pair of languages.

The algorithm returns a Boolean value indicating if the grammar class in language L is compatible with the grammar class in language L'. The algorithm includes verification

[2]Data obtained by José-Miguel Benedí, Universidad Politécnica de Valencia; `http://users.dsic.upv.es/~jbenedi/`.

[3]`http://www.lsi.upc.edu/~nlp/web/`

Table A.3: Distribution of Spanish grammar classes.

freq	grammar	freq	grammar	freq	grammar
779175	SPS00	78262	AQ0MS0	2	VAIS2P0
350406	NCFS000	73092	DI0MS0	2	P02CP000
343046	NCMS000	71255	VMP00SM	2	AQXFS0
219842	DA0MS0	67882	P0000000	2	AQXCP0
201115	CC	64774	AQ0FS0	1	VSSF2S0
197969	RG	59394	VMIS3S0	1	VSM02S0
187499	DA0FS0	57661	DI0FS0	1	VSM02P0
170729	NP00000	56185	RN	1	VMSF3S0
147818	NCMP000	52512	VMII1S0	1	VASF3P0
137967	CS	81613	DA0MP0	1	VAM01P0
	...			1	VAIC2P0
116310	NCFP000	3	VSSF3P0	1	PX2MP0P0
106492	VMIP3S0	3	VASF1S0	1	PX1FP0S0
93495	PR0CN000	3	VAM02P0	1	PT0FS000
88735	AQ0CS0	3	AQXMS0	1	AQXMP0
81613	DA0MP0	2	VASI2P0	1	AQACP0

of conditions like those mentioned above, e.g., `if (English word is <Noun, Sg> and Spanish word is <Noun, Sg>) then return true`.

Still, we would like to comment on one language-specific decision that we made: given an English verb, we consider that English past participle and gerund are compatible with practically any Spanish verb form in indicative. This decision is made because such verb forms are often part of compound tenses (perfect tenses and continuous tenses). For the same reason, Spanish participle and gerund are considered compatible with any English verb form.

In those cases where the grammar classes are incompatible, a very low probability is assigned to the translation into the implied word form. We use a threshold ϵ for the sum of all "incompatible" forms. Thus, all "compatible" word forms are equally distributed with the value of $1 - \epsilon$. For instance, consider that, for a set of potential translations $p(w', w)$, the set of word forms w' consist of two compatible and three incompatible forms. The probability associated to the compatible forms will be $p(w', w) = (1 - \epsilon)/2$, and for the incompatible forms, it will be $p(w', w) = \epsilon/3$.[4]

Once we obtain the similarity estimations for all possible translations of word forms from one language into another on the basis of compatibility of the corresponding grammar classes, we follow on with the estimation of probabilities based on grammar distribution. This distribution establishes how likely is the appearance of the word form w with the given grammar class GC, computed as:

$$g_d(w_{GC}) = \frac{freq(GC)}{\sum_{GC \in L} freq(GC)} \quad . \tag{A.8}$$

This estimation is based on the relative frequency of the grammar class GC in a sig-

[4]The value of ϵ must be estimated empirically. In this case we considered $\epsilon = 0.025$.

nificantly large corpus of language L. This process is carried on separately for each language. Finally, the translation probability for a pair (w, w') is estimated as follows:

$$p(w', w) = g_d w' \cdot g_d w \cdot \varrho(w' \mid w) \ . \tag{A.9}$$

Note that we are interested in the probability of translations of a word form. If several grammar tags correspond to only one word form (for instance, consider the form *toma* in Table A.4), the probability of the corresponding translation is the result of the sum of probabilities associated to each grammar tag, i.e.:

$$\varrho(w' \mid w) = \sum_{GC} p(w'_{GC} \mid w) \ . \tag{A.10}$$

Finally, in order to obtain actual probabilities, the obtained values are scaled such that:

$$\sum_{w'} p(w' \mid w) = 1 \ . \tag{A.11}$$

An example of the dictionary's entries is included in Table A.4.[5] Note that, for illustrative purposes, only inflections of the verb *tomar* are included. However, just as in the dictionary of Table A.1 many more (sometimes no so logical) entries exist.

On the basis of this dictionary, we generated a stemmed version, where the probabilities were accumulated and distributed over the entries' stems. Heading back to the example of Table A.4, in the stem dictionary only three entries relate the English *took* with a Spanish stem related to the verb *tomar*: (a) p(took,tomareis) = 0.00000609, (b) p(took,tom) = 0.0156869, and (c) p(took,tomar) = 0.00001265.

Both versions of the dictionary have been used in our experiments on cross-language plagiarism detection in Section 7.5.

[5]The dictionary is freely available at http://users.dsic.upv.es/grupos/nle/downloads.html

Table A.4: Example entries in the inflectional dictionary. We consider some Spanish entries for the English word *took* (grammar information included for illustration purposes only).

possible Spanish translation	$p(w' \mid took)$	**possible Spanish translation**	$p(w' \mid took)$
tomó_VMIS3S0	0.3016546	tomaría_VMIC3S0;VMIC1S0	0.0006075
tomaba_VMII3S0;VMII1S0	0.2752902	tomará_VMIF3S0	0.0005070
tomaban_VMII3P0	0.0800329	tomen_VMSP3P0;VMM03P0	0.0004208
tomaron_VMIS3P0	0.0670665	tomas_VMIP2S0	0.0004094
tomé_VMIS1S0	0.0528457	tomabais_VMII2P0	0.0002844
tomamos_VMIS1P0;VMIP1P0	0.0494479	tomasteis_VMIS2P0	0.0002235
tomase_VMSI3S0;VMSI1S0	0.0424848	tomarán_VMIF3P0	0.0001992
tomara_VMSI3S0;VMSI1S0	0.0424848	tomaseis_VMSI2P0	0.0001874
tomasen_VMSI3P0	0.0121436	tomarais_VMSI2P0	0.0001879
tomaran_VMSI3P0	0.0121436	tomarían_VMIC3P0	0.0001489
tomar_VMN0000	0.0113312	tomemos_VMSP1P0;VMM01P0	0.0001304
toma_VMM02S0;VMIP3S0	0.0091485	tomes_VMSP2S0	0.0001065
tomábamos_VMII1P0	0.0087611	tomaré_VMIF1S0	0.0000988
tomado_VMP00SM	0.0059050	tomaremos_VMIF1P0	0.0000946
tomaste_VMIS2S0	0.0044491	tomarás_VMIF2S0	0.0000477
toman_VMIP3P0	0.0033597	tomaríamos_VMIC1P0	0.0000433
tomabas_VMII2S0	0.0033013	tomarens_VMSF3P0	0.0000413
tomando_VMG0000	0.0023740	tomáremos_VMSF1P0	0.0000410
tomada_VMP00SF	0.0019706	tomareis_VMSF2P0	0.0000410
tomásemos_VMSI1P0	0.0017167	tomáis_VMIP2P0	0.0000320
tomáramos_VMSI1P0	0.0017167	tomad_VMM02P0	0.0000258
tomo_VMIP1S0	0.0014987	tomarías_VMIC2S0	0.0000136
tomados_VMP00PM	0.0014060	toméis_VMSP2P0	0.0000111
tome_VMSP3S0;VMSP1S0;VMM03S0	0.0011019	tomaréis_VMIF2P0	0.0000062
tomadas_VMP00PF	0.0008767	tomare_VMSF3S0;VMSF1S0	0.0000017
tomases_VMSI2S0	0.0007872	tomares_VMSF2S0	0.0000015
tomaras_VMSI2S0	0.0007872	tomaríais_VMIC2P0	0.0000008

Appendix B

Related Publications

This research has generated a total of 21 scientific publications. Section B.1 contains journal papers and Section B.2 contains international conferences. Finally, Section B.3 and B.4 show book chapters and workshops. For each publication we show its impact (in terms of citations)[1] and mention the related chapter(s) of this thesis. The contributions to each paper are described in each section, after the publications list.

B.1 Journals

1. M. Potthast, **A. Barrón-Cedeño**, B. Stein, and P. Rosso. Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis, 45(1):1-18, 2011.
2. D. Pinto, J. Civera, **A. Barrón-Cedeño**, A. Juan, and P. Rosso. A Statistical Approach to Crosslingual Natural Language Tasks. Journal of Algorithms, 64(1):51-60, 2009.
3. B. Stein, M. Potthast, P. Rosso, **A. Barrón-Cedeño**, E. Stamatatos, and M. Koppel. Fourth International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. ACM SIGIR Forum 45, no. 1 (May 2011): 45-48. DOI: 10.1145/1988852.1988860, 2011
4. **A. Barrón-Cedeño** and P. Rosso. On the Relevance of Search space Reduction in Automatic Plagiarism Detection. Procesamiento del Lenguaje Natural, 43:141-149, 2009.

Table B.1 shows the impact of these publications and their related chapters.

[1] Numbers obtained through `http://scholar.google.com/` on Feb. 12 2012.

Table B.1: Overview of publications in journals. The information includes related chapter in the thesis, impact factor, and number of citations (with self citations) including: (*i*) books, (*ii*) journals, (*iii*) conferences, and (*iv*) theses (in (*i*) to (*iv*) no self-citations are included).

publication	chapter(s)	impact factor	citations (self):	books	journals	conf.	theses
1	3, 6	0.615	21 (9)	1	2	7	1
2	3, 6, 7	0.667	14 (11)		2	1	
3	4, 7, 8						
4	5		3 (1)		1		1

B.2 Conferences

1. **A. Barrón-Cedeño**, P. Rosso, E. Agirre, and G. Labaka. Plagiarism Detection across Distant Language Pairs. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, August 2010.
2. **A. Barrón-Cedeño**. On the Mono- and Cross-Language Detection of Text Reuse and Plagiarism. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010.
3. M. Potthast, B. Stein, **A. Barrón-Cedeño**, and P. Rosso. An Evaluation Framework for Plagiarism Detection. In COLING 2010 Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 997-1005, Beijing, China, August 2010.
4. E. Flores, **A. Barrón-Cedeño**, P. Rosso, and L. Moreno. DeSoCoRe: Detecting Source Code Re-Use across Programming Languages. In Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal Canada, 2012. ACL.
5. **A. Barrón-Cedeño** and P. Rosso. On Automatic Plagiarism Detection based on n-grams Comparison. Advances in Information Retrieval. In Proceedings of the 31st European Conference on IR Research, LNCS (5478):696-700, 2009.
6. **A. Barrón-Cedeño**, P. Rosso, and J.-M. Benedí. Reducing the Plagiarism Detection Search Space on the Basis of the Kullback-Leibler Distance. In Computational Linguistics and Intelligent Text Processing, 10th International Conference, LNCS (5449):523-534, 2009.
7. **A. Barrón-Cedeño**, C. Basile, M. Degli Esposti, and R. Paolo. Word Length n-grams for Text Re-Use Detection. In Computational Linguistics and Intelligent Text Processing, 10th International Conference, LNCS (6008):687-699, 2010.
8. **A. Barrón-Cedeño**, P. Rosso, S. Lalitha Devi, P. Clough, and M. Stevenson. PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. In Proceedings of the Forum for Information Retrieval Evaluation (FIRE 2011), Bombay, India, 2011
9. **A. Barrón-Cedeño**, M. Potthast, P. Rosso, B. Stein, and A. Eiselt. Corpus and Evaluation Measures for Automatic Plagiarism Detection. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2010).

Table B.2: Overview of publications in conferences. The information includes related chapter in the thesis, conference CORE level, and number of citations (with self citations) including: (*i*) journals, (*iii*) conferences, and (*iv*) theses (in (*i*) to (*iv*) no self-citations included).

publication	**chapter(s)**	**CORE**	**citations (self):**	journals	conferences	theses
1	6	A	3 (1)		2	
2	5, 6, 7, 9	A	2 (0)		2	
3	4, 7	A	24 (4)	4	16	
4	10	A				
5	5	B	24 (4)	7	9	4
6	5	B	16 (5)	2	8	1
7	5, 7	B	4 (0)	2	1	1
8	9					
9	4, 7, A		2		2	
10	6, 7, A		1		1	
11	4, 9		7 (2)		4	1
12	4, 7					
13	10	C				
14	10					

10. G. Sidorov, **A. Barrón-Cedeño**, and P. Rosso. English-Spanish Large Statistical Dictionary of Inflectional Forms. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2010).
11. **A. Barrón-Cedeño**, A. Eiselt, and P. Rosso. Monolingual Text Similarity Measures: A Comparison of Models over Wikipedia Articles Revisions. In ICON 2009, pages 29-38. Macmillan Publishers, 2009.
12. **A. Barrón-Cedeño** and P. Rosso. Towards the 2nd International Competition on Plagiarism Detection and Beyond. In Proceedings of the 4th International Plagiarism Conference (IPC 2010), Newcastle upon Tyne, UK, 2010. Plagiarism Advice.
13. E. Flores, **A. Barrón-Cedeño**, P. Rosso, L. Moreno. Towards the Detection of Cross-Language Source Code Reuse. In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems, NLDB-2011, Springer-Verlag, LNCS(6716), pp. 250-253, 2011
14. E. Flores, **A. Barrón-Cedeño**, P. Rosso, L. Moreno. Detecting Source Code Reuse across Programming Languages. Poster at Conf. of Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), Huelva, Spain, 2011

Table B.2 shows the impact of these publications and their related chapters.

B.3 Book Chapters

1. **A. Barrón-Cedeño**, M. Vila, and P. Rosso. Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica. (Jornadas (in)formativas de lingüística forense), chapter Detección automática de plagio: De la copia exacta

a la paráfrasis. Euphonia Ediciones SL., Madrid, Spain, 2010.

This publication is related to Chapter 8.

B.4 Workshops

1. M. Potthast, A. Eiselt, **A. Barrón-Cedeño**, B. Stein, and P. Rosso. Overview of the 3rd International Competition on Plagiarism Detection. In Notebook Papers of CLEF 2011 Labs and Workshops, Amsterdam, The Netherlands, 2011
2. J.A. Silvestre-Cerdà, M. García-Martínez, **A. Barrón-Cedeño**, J. Civera, and P. Rosso. Extracción de corpus paralelos de la Wikipedia basada en la obtención de alineamientos bilingües a nivel de frase. In Proceedings of the SEPLN Workshop ICL: Iberian Cross-Language NLP tasks, CEUR-WS.org, vol. 824, pp. 14-21, 2011.
3. M. Potthast, **A. Barrón-Cedeño**, A. Eiselt, B. Stein, and P. Rosso. Overview of the 2nd International Competition on Plagiarism Detection. In Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy, 2010.
4. M. Potthast, B. Stein, A. Eiselt, **A. Barrón-Cedeño**, and P. Rosso. Overview of the 1st International Competition on Plagiarism Detection. In Proceedings of the SEPLN Workshop PAN: Uncovering Plagiarism, Authorship and Social Software Misuse, pages 1-9. CEUS-WS.org, 2009.
5. **A. Barrón-Cedeño**, P. Rosso, D. Pinto, and A. Juan. On Crosslingual Plagiarism Analysis Using a Statistical Model. In ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008), pages 9-13, Patras, Greece. CEUR-WS.org, 2008
6. **A. Barrón-Cedeño** and P. Rosso. Towards the Exploitation of Statistical Language Models for Plagiarism Detection with Reference. In ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2008), pages 15-19, Patras, Greece. CEUR-WS.org, 2008
7. **A. Barrón-Cedeño** and P. Rosso. Monolingual and Crosslingual Plagiarism Detection: Towards the Competition @ SEPLN09. In Memorias de las III Jornadas PLN-TIMM, Madrid, Spain, February 5-6, 2009, pp.29-32
8. E. Flores, **A. Barrón-Cedeño**, P. Rosso, and L. Moreno. Detección de reutilización de código fuente entre lenguajes de programación en base a la frecuencia de términos. In Memorias de las IV Jornadas PLN-TIMM, Torres, Jaén, Spain, 2011, pp.21-26

Table B.3 shows the impact of these publications and their related chapters.

Table B.3: Overview of publications in workshops. The information includes related chapter in the thesis and number of citations (with self citations) including: (*i*) journals, (*iii*) conferences, and (*iv*) theses (in (*i*) to (*iv*) no self-citations are included).

publication	**chapter(s)**	**citations (self):**	journals	conferences	theses
1	4, 7				
2	9				
3	4, 7	34 (5)	3	23	3
4	4, 7	37 (9)	6	20	2
5	3, 6	16 (10)	3	3	
6	5	1 (0)			1
7	6, 7				
8	10				

Appendix C

Media Coverage

Our research work has received certain attention from media as well, particularly regarding cross-language plagiarism. In this appendix we include just some of the reports. Please note that some of the news are in the tabloid journalism style, where stories are exaggerated in order to make them more sensational: e.g. "a Mexican makes history in Spain" (...).

C.1 News

1. El País. "A la caza del plagio en las traducciones" `http://bit.ly/pais_caza_plagio`. April 5, 2011, Spain.
2. ABC (Tecnología). "A la caza del plagio en las traducciones" `http://bit.ly/abc_caza_plagio`. April 5, 2011, Spain.
3. Levante-EMT. "Cerco informático al plagio en los textos traducidos". `http://bit.ly/levante_cerco-informatico` April 6, 2011. Spain
4. Las Provincias. "Un estudiante de la Politécnica logra detectar traducciones plagiadas". `http://bit.ly/provincias_traducciones_plagiadas`. April 5, 2011, Spain.
5. Actualités, Portalingua, Observatoire Des Langues Dans La Connaissance. "A la caza del plagio en las traducciones" `http://bit.ly/portalingua_plagio_traducciones`. April 8, 2011.
6. El Mundo Edición: C. Valenciana. "Detección automática de plagios de texto, incluso en traducciones". `http://users.dsic.upv.es/~prosso/ElMundo.pdf` April 11, 2011. Spain
7. Excélsior El periodico de la vida nacional. "Un mexicano hace historia en España con detector de plagios". `http://bit.ly/excelsior_espana_detector`. May 4, 2011, Mexico.
8. El Universal. "Mexicano crea método para detectar plagios". `http://bit.ly/universal_mex`. May 4, 2011, Mexico.
9. Noticias De Mérida Yucatán Hoy. "Académico desarrolla método para detectar

plagios" http://bit.ly/yucatan_detectar_plagio. May 4, 2011, Mexico.
10. La Patria. "Un estudiante mexicano logra detectar traducciones plagiadas en fondo y forma" http://bit.ly/patria_plagio_traducciones. May 5, 2011, Bolivia.
11. Reforma (Bucio, Erika P). "Detectan plagio en traducciones". http://users.dsic.upv.es/~prosso/Reforma.pdf. May 7, 2011, Mexico
12. ...

C.2 On Air and TV

Four radio shows interviewed us:

- Radio Nacional de España: http://users.dsic.upv.es/~prosso/RadioNacional.wmv April 6, 2011.
- Radio Nou: http://users.dsic.upv.es/~prosso/Radio9.wmv April 6, 2011.
- Cadena COPE. April 6, 2011.
- Radio Nou: http://users.dsic.upv.es/~prosso/Radio9_bis.wmv April 11, 2011

Additionally, some TV news shows made coverage of our research:

1. Nt9 1a Edició, Canal Nou: "Un arma nova contra qui copie". http://users.dsic.upv.es/~prosso/CANAL9.wmv Valencia, Spain, May 9 2011
2. EFE: "Mexicano inventa método para detectar plagio de textos" http://www.youtube.com/watch?v=SKGi-XIy1O4 May 14, 2011
3. CNN en Español: "Encuentro". May 16, 2011.
4. UPV-TV: "Detección de plagio". http://bit.ly/upv_news_plagio Valencia, Spain, May 2011.

Acronyms

AAPS	Anti-Anti-Plagiarism System
AP	Associated Press
API	application programming interface
BoW	bag of words
CL	computational linguistics
CL!TR	Cross-Language !ndian Text Re-use
CL-ASA	cross-language alignment-based similarity analysis
CL-C3G	cross-language character 3-grams
CL-CNG	cross-language character n-grams
CL-COG	cross-language cognateness
CL-ESA	cross-language explicit semantic analysis
CL-KCCA	cross-language kernel canonical correlation analysis
CL-LSI	cross-language latent semantic indexing
CLEF	Cross Language Evaluation Forum
CLIR	cross-language information retrieval
COPS	COpy Protection System
EM	expectation maximisation
ESA	explicit semantic analysis
ETS	Educational Testing Service
FIRE	Forum for Information Retrieval Evaluation
FL	forensic linguistics
GST	greedy string tiling
HFM	highest false match
IBM	International Business Machines Corporation
ICFL	Informative Conference on Forensic Linguists
IEEE	Institute of Electrical and Electronics Engineers
IR	information retrieval
ISO	International Organization for Standardization
JRC	Joint Research Centre
JWPL	Java Wikipedia Library
KCCA	kernel canonical correlation analysis
KL	Kullback-Leibler
LM	language model
LSI	latent semantic indexing
LTM	lowest true match
M1	IBM model 1
ML	machine learning

MLE	maximum likelihood estimation
MT	machine translation
MCD	multilingual copy detection
METER	Measuring TExt Reuse
NER	named entity recognition
NLE	natural language engineering
NLP	natural language processing
P4P	paraphrases for plagiarism
PA	Press Association
PAN	Uncovering Plagiarism, Authorship and Social Software Misuse
PAN-PC	PAN plagiarism corpus
POS	part of speech
SAER	sentence alignment error rate
SCAM	Stanford Copy Analysis Mechanism
SMS	Short Message Service
SMT	statistical machine translation
Stylysis	style analysis
SVD	singular value decomposition
SVM	support vector machine
TM	translation model
TREC	Text REtrieval Conference
ttr	type/token ratio
T+MA	translation plus monolingual analysis
VSM	vector space model
WSD	word sense disambiguation
XML	eXtensible Markup Language

Index