Semantic Annotation of CESS-ECE: Named Entities annotation criteria

1. Used labels:

Morphological level:

Words tagged with one of the following labels in the PoS are considered Strong Named Entities (SNE). We record a brief explanatory note about the semantics of each of the labels, which will be discussed further on:

- **np0000p** for people
- **np0000o** for organizations
- **np0000l** for locations
- **np0000a** for *others*
- **W** for dates
- **Z** for numbers
- **Zp** for percentages
- **Zm** for coins

Syntactic level:

Phrases tagged with one of the following labels in the PoS are considered Weak Named Entities (WNE) unless they consist of a mere SNE (that is to say, unless there is a 100 % match between SNE and WNE). We record a brief explanatory note about the semantics of each of the labels, which will be discussed further on:

- **snp** for people (stands for *sintagma nominal de persona* –person noun phrase–).
- **sno** for organizations (stands for *sintagma nominal d'organització* –organization noun phrase–).
- **snl** for locations (stands for *sintagma nominal de lloc* –place noun phrase–).
- **snd** for dates (stand for *sintagma nominal de data* –date noun phrase–).
- **snn** for numbers, percentages, coins and magnitudes (stands for *sintagma nominal numèric* –numerical noun phrase–).
- **sna** for *others* (stands for *altres sintagmes nominals* –other noun phrases–).

2. Detection and delimitation: formal characterization of NEs

There are two underlying golden rules in the definition of NE in Spanish and Catalan. On one hand, only a noun phrase can be a NE. On the other, its referent must be unique and unambiguous. Finally, another very strong rule (though not 100 % reliable on) is that only a *definite singular* noun phrase might be a NE. Next, we report the guidelines followed in order to define what should be tagged as a NE:

- All definite noun phrases containing a Proper Noun (PN) are considered NEs, even if it is not the core of the phrase. In the later case, the core should be a Trigger Word.
- All definite noun phrases containing an element with the date (W) or numerical (Z, Zm, Zp) label in the PoS are considered NEs, unless that PoS is a specifier, which is a very common

case for numbers.

- All definite noun phrases whose core is a Trigger Word complemented either by a demonym or a proper noun derived adjective. Examples of this are *Catalan, European* (for demonyms) and *leninist, fascist* (for proper noun derived adjectives).
- Noun phrases which contain a magnitude name.

3. Classification: semantic characterization of NEs

The semantics of a NE are normally determined by the core noun of the phrase. Consequently, the classification proposed here is syntax-centered. We use six categories:

1. Person Named Entities (syntactic label: **snp**; morphologic label **np0000p**), which include:
   - Real people (*Bruce Springsteen, Karl Marx*)
   - Anthropomorphic fiction characters (*Bilbo Baggins, Sam Spade*)
   - Anthropomorphic religious and mythological figures (*Moses, Buda, Aphrodita*)
   - Anthropomorphic folkloric characters (*Guy Fawkes, the Sandman*)
2. Organization Named Entities (syntactic label **sno**; morphological label **np0000o**), which include:
   - Societies and Enterprises (*Microsoft, General Motors*)
   - Political and Social agents (*the Republican Party, Amnesty International*)
   - Institutions (*the Church, the Senate, the House of Lords*)
   - Associations, groups, teams (*Chelsea FC, the Ramones*)
3. Location Named Entities (syntactic label **snl**, morphological label **np0000o**), which include:
   - Geographical features (*Montblanc, the Thames*)
   - Political divisions of territory (*Spain, Barcelona*).
   - Places of human activities (*Wembley Stadium, the University of Barcelona*)
   - Imaginary places (*Atlantis, the Garden of Eden, Middle-Earth*)
4. Date Named Entities (syntactic label **snd**, morphological label **W**), which include:
   - Days, months, years, centuries (*October 23rd 1999, the 20th century*)
   - Hours (*14:30 pm, April 23rd at 10:00 am*)
5. Alphanumerical Named Entities (syntactic label **snn**), which include:
   - Numbers, whenever they are used as nouns and are the core of a noun phrase. They have **Z** as the morphological label.
   - Percentages, even if they are not part of a *definite* noun phrase (*20 % of the population*). They have **Zp** as the morphological label.
   - Coins, specially if specified by a number indicating an amount of money (*20 million pounds*). They have **Zm** as the morphological label.
   - Magnitudes, specially if specified by a number indicating an amount of (*20 feet, 45.2 Db*). They do not have any specific morphological label.
6. Miscellaneous Named Entities (syntactic label **sna**), which include:
   - Existing or fiction animals (*Lassie, Snowflake*)
   - Fiction non-anthropomorphical beings (*Kraken, Treebeard*)
   - Publications, artistic or literary opera, Internet adresses (*www.ub.edu*, *the Gioconda*)
   - Juridical or Physical laws and principles (*Archimedes' principle, the Constitution*)
   - Artifacts (*Playstation 2, Apollo XII*)
   - Ideologies, theories (*the Perestroika, the Welfare State*)
   - Events and commemorations (*Spanish Civil War, the International Day of AIDS*)
   - Competitions and Awards (*Nobel Prize, the NBA league*)

- Degrees and qualifications
- Heavenly bodies (*Alpha Centauri, the Moon*)

4. Codification of the Named Entities in the column format: the SemEval-2007 dataset for Task#9

When translated from the constituency tree to the column format, the named entities are codified with the open-close format and using a unified set of type labels (irrespective if they are strong entities coming from the morphological level or weak entities coming from the syntax level). The set of labels used is:

- PER (for person; corresponds to the labels: 'np0000p' and 'snp')
- LOC (for location; corresponds to the labels: 'np0000l' and 'snl')
- ORG (for organization; corresponds to the labels: 'np0000o' and 'sno')
- DAT (for date; corresponds to the labels: 'W' and 'snd')
- NUM (for numerical expression; corresponds to the labels: 'Z', 'Zp', 'Zm', and 'snn')
- OTH (for others; corresponds to the labels 'np0000a' and 'sna')

In order to make the recognition task of strong entities not trivial, when converting to the column format, we remove the last symbol from the 'np0000*' POS labels, where * stands for any member of the set ('p', 'o', 'l', 'a'). With the same motivation, the 'sn*' labels corresponding to named entities are reduced to 'sn' when generating the syntax column for the training/test datasets. This corresponds to wash out semantic information from syntactic labels in the constituency tree.

Important note

The main goal when enriching syntactic trees with named entities in the CESS-ECE project was to stick to the syntactic constituents, instead of generating new artificial nodes or to tag an entity fragmented in multiple nodes. Since there is no perfect matching between named entities and syntactic constituents, sometimes the above criteria for named entity annotation (weak entities are tied to SN constituents) lead to undesirable segmentations. For example in the case of SNs including relative clauses, there is no way to mark the entity without including the relative clause. This way, given the following tree (this is, precisely, the example in the website for task#9):

```
(
 (S
   (sn-SUJ-Arg1-TEM
     (espec.fp
       (da0fp0 Las el))
     (grup.nom.fp
       (ncfp000 conclusiones conclusión 01207975n)
       (sp
         (prep
           (sps00 de de))
         (sno
           (espec.fs
             (da0fs0 la el))
           (grup.nom.fs
             (ncfs000 comisión comisión 01207975n)
             (snp
               (grup.nom
                 (np0000p Zapatero Zapatero)))
             (S.F.R
               (Fc , ,)
               (relatiu-SUJ-Arg0-CAU
```

```
              (pr0cn000 que que))
            (gv
              (vmif3s0 ampliará ampliar-a1))
            (sn-CD-Arg1-PAT
              (espec.ms
                (da0ms0 el el))
              (grup.nom.ms
                (ncms000 plazo plazo 01207975n)
                (sp
                  (prep
                    (sps00 de de))
                  (sn
                    (grup.nom.ms
                      (ncms000 trabajo trabajo 01207975n))))))
            (Fc , ,)))))))
(gv
  (vmip3p0 quedan quedar-b3))
(sp-CC-ArgM-TMP
  (prep
    (sps00 para para))
  (sp
    (prep
      (spcms después_del después_del))
    (sn
      (grup.nom.ms
        (ncms000 verano verano 01207975n)))))
(Fp . .)))
```

the weak named entity (ORG) covering “la comisión Zapatero” has to be attached to the ‘sno’ constituent including also the relative clause (S.F.R) and covering the entire “la comisión Zapatero, que ampliará el plazo de trabajo,”. This is, of course, an incorrect segmentation for the entity. There are a few other examples of mismatching situations between syntax and named entities that produce weird NE segmentations.

This is an issue to be fixed in the future and that it is currently under study. For the SemEval-2007 competition we have decided to maintain the current annotations, thus we suggest participants to rely on the syntactic information to do named entity identification. This will presumably be a simple task, being the semantic part of NE classification the most difficult and interesting one.