

## Semantic Annotation of CESS-ECE: WordNet annotation of nouns for Spanish and Catalan languages

The main criterion for selecting the nouns to be annotated has been the frequency. We have only annotated the most common nouns, leaving aside the less frequent, as well as verbs, adjectives and adverbs. We have used a steady version of WordNet 1.6.

Each noun matches one synset or, in some cases, a label indicating a circumstance which lead us not to use the corresponding synset. These synsets or labels are shown at the end of the noun lemma, and are as follows:

- **Synsets:** 8 numerical digits followed by an “n” which stands for “noun”.
- **C1S:** “Word exists in dictionary but not its sense”.
- **C2S:** “Word does not exist in dictionary”.
- **C3S:** “Word is part of a Multiword Lexical Unit or a lexicalized inflected form”.
- **C4S:** “Word is part of a Named Entity”.
- **C5S:** “The tagger is strongly uncertain”.
- **C6S:** “Word was improperly lemmatized or PoS-Tagged”.
- **C7S:** “Word is wrongly used: misspelling”.

Examples:

POS	WORD	LEMMA	SYNSET
	(ncmp000	milions	milió 09902865n)
	(ncfs000	zona	zona 06395720n)
	(ncms000	móvil	móvil C1S)
	(ncmp000	recursos	recurs C2S)
	(ncmp000	serveis	servei C3S)

**NOTE:** For the purpose of Noun Sense Disambiguation in SemEval-2007 task#9 all the ‘C[1-7]S’ labels have been unified in a single ‘CS’ label. Tus, participant systems Hill be required to discriminate among all different synsets for a particular noun, plus an extra ‘CS’ label indicating that the sense for the noun occurrence does not match any of the WordNet synsets for that noun.