

# The NOMAGE Project

## Coding the semantic features of French nominalizations

Antonio Balvet\*      Pauline Haas\*      Richard Huyghe\*\*  
Anne Jugnet\*      Rafael Marín\*

\*Univ. Lille 3/CNRS UMR STL 8163      \*\*Univ. Paris 7

### 1 From verbs to nouns

From the work of (Lees, 1960), through (Chomsky, 1970) and (Grimshaw, 1990), to more recent studies, nominalizations have occupied a central place in grammatical analysis, with a focus on morphological and syntactic aspects. More recently, researchers have begun to address a specific issue often neglected before, *i.e.* the semantics of nominalizations, and its implications for Natural Language Processing applications such as electronic ontologies or Information Retrieval. We focus on precisely these issues in the research project NOMAGE (ANR-07-JCJC-0085-01), a young researchers' project funded by the French National Research Agency (ANR). The present submission outlines this ongoing semantic annotation project. Several previous works on the semantics of deverbal nouns (Haas & Huyghe, to appear), (Haas, Huyghe, & Marín, 2008) and (Huyghe & Marín, 2007), indicate that nominals inherit certain aspectual properties from their associated verbs. This is particularly clear when comparing telic verb (accomplishment and achievement) *vs.* atelic verb (state and activity) nominal derivations. For example, telic verb nominal derivations may appear in an N position such as in (1a) and (1b) while this position is not available to stative verb nominal derivations.

- (1) a. Le N a eu lieu à tel endroit / à tel moment  
The N occurred in that place / at that time
- b. Le lieu du N  
The place of the N

The behavior of activity verb nominal derivations, in contrast, is not as clear. At least two different subclasses should be considered: one composed of nominals such as *natation* ‘swimming’ or *jardinage* ‘gardening’ which, like states, are not acceptable in such contexts; and a second group with cases such as *manifestation* ‘demonstration’ or *discussion* ‘discussion’ which, surprisingly, are indeed acceptable in such contexts. This distinction is correlated with another feature: its countability; nominalizations such as *natation* are mass nouns, while those such as *manifestation* are count nouns.

## 2 Annotating semantic properties in context

Based on the aforementioned works, we may note that while some aspectual features are indeed inherited by the nominal form of the verb, this is not always the case, and the effects are not always the same as in the original verb form. Detecting such discrepancies between verbal and nominal domains, together with finding the most useful linguistic tests to account for the semantics of deverbal nouns, are two of the main issues the NOMAGE project aims to address. Based on these preliminary findings, we have just started a manual semantic annotation task on the French Treebank (Abeillé, 2003) –a 1 million-word tagged, parsed and functionally annotated corpus– centred on aspectual properties. The first nominal candidates annotated so far are those ending in *-ment* and *-age* suffixes, which represent 2154 items (out of 9853 candidates), respectively 1611 *-ment* and 543 *-age* nouns, of which only 1409 and 382 were deemed relevant. For this annotation task, one of the main challenges has been to adapt the semantic tests used (transformations) to *real-life* sentences. Based on the data gathered so far, we have proposed a higher-level semantic classification for those nouns, as shown in the table below, thus allowing us to analyse the distribution of aspectual classes among different suffixes, and to compare stative *vs.* non-stative interpretations of nominalizations.

Suffix type	<i>-age</i>	<i>-ment</i>
Stative	0%	3.4%
Dynamic	46.33%	59.76%
Concrete	7.85%	26.96%
Uncategorizable	45.8%	7.88%

As the table above shows, the majority of the processed nominalizations fall into the Dynamic category; Stative nouns are very rare (3.4% of the *-ment* nominalizations), to the extent that no occurrence of a *-age* noun

could be found. For *-ment* nouns, around 30% of the cases are Concrete interpretations, e.g.: *bâtiment* ‘building’. As for the *-age* nouns, over 45% are not categorizable based on the annotations gathered so far: in half of the Uncategorizable cases, a truly underspecified reading seems to apply, while in the rest of the cases, the interpretation of the utterance does not depend on the fine-grained semantic categorization considered above. An example of such an underspecified reading is *emballage* ‘packaging’, as in *leader européen et troisième mondial de l’emballage* ‘european first and world-class third positions in packaging’, with both a Dynamic and Concrete readings.

### 3 Perspectives

In the course of the NOMAGE project, subsequent annotation campaigns will be undertaken. Our aim is to collect enough consistent data to train a semantic tagger, based on tools developed by members of the project’s consortium. Preliminary experiments on functional tagging described in (Moreau et al., submitted) show the need to supplement the French Treebank and other usable corpora with basic information such as: headedness for nouns; predicative nature for verbs, nouns and adjectives; and the adjunct/argument distinction for NPs. The data obtained will form the groundwork of a semantic lexicon of nominalizations for French, a resource intended for both human and machine use in NLP and IR applications. For example, the high-level distinctions discussed above, namely Stative, Dynamic and Concrete interpretation could prove useful in tasks such as automatic Word Sense Disambiguation (e.g. ‘construction’ in a Concrete *vs.* a Dynamic interpretation) and Topic Detection and Tracking (e.g. spotting different instances of a given Event).

### References

- Abeillé, A. (2003). *Treebanks, Building and Using Parsed Corpora*. Dordrecht, Kluwer.
- Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in english transformational grammar*. Waltham (MA): Ginn and Co.
- Grimshaw, J. (1990). *Argument structure*. Cambridge/ London: The MIT Press.
- Haas, P., & Huyghe, R. (to appear). Les propriétés aspectuelles des noms d’activités. *Cahiers Chronos*.
- Haas, P., Huyghe, R., & Marín, R. (2008). Du verbe au nom: calques et décalages aspectuels. In J. Durand, B. Habert, & B. Laks (Eds.), *Congrès mondial de linguistique française CMLF’08* (p. 2051-2065). Paris.

- Huyghe, R., & Marín, R. (2007). L'héritage aspectuel des noms déverbaux en français et en espagnol. *Faits de Langues*, 30, 265-273.
- Lees, R. B. (1960). *The grammar of english nominalizations*. The Hague: Mouton de Gruyter.
- Moreau, E., Tellier, I., Balvet, A., Laurence, G., Rozenknop, A., & Poibeu, T. (submitted). Learning function tags in a french treebank using CRF. In *Proceedings of the eacl'09 conference*.