Looking for Syntax/Aspect Mappings: a Case Study on the French Treebank

Antonio Balvet, Pauline Haas, Anna Kupsc, Rafael Marin

The aim of this study is to investigate the interaction between syntactic and semantic properties of French verbs, based on usage data extracted from a reference corpus. Specifically, we try to assess whether notions such as the telic/internal argument dependence have any verifiable support. We also want to assess whether aspectual distinctions can be annotated in a consistent way, based on syntactic construction and semantic/syntactic tests.

TreeLex (Kupść & Abeillé, 2008), a syntactic lexicon of 1912 verbs automatically extracted from the French Treebank, FTB (Abeillé et al. 2003), was used for this study. TreeLex uses a rich syntactic representation: every frame (syntactic argument structure) contains both the function of an argument and its phrasal realization. Yet, no semantic information is available. In line with other FTB-based projects, e.g., Nomage (Balvet et al. 2012), we have produced a TreeLex version augmented with verbal aspect information. In order to achieve this, considerable manual annotation work was needed to determine each verb's aspectual properties in context: *ca.* 40% of TreeLex verbs have multiple frames (Kupść & Abeillé, 2008); thus, they are potentially polysemous, and probably polyaspectual too.

Aspectual Information

Our annotation process was twofold. First, we collected examples from the FTB to illustrate how each frame is instantiated. We restricted our study to TreeLex verbs with a single frame only, excluding Multiword Verbal Units, as well as low-frequency frames (e.g., reflexive "se" constructions). This reduction resulted in 1019 single-frame verbs, with their respective FTB examples. Then, aspect information for these verbs was added manually by two experts in semantics, adopting the four major (Vendlerian) classes: States (ETAT, 84 verbs or 8.2%), Activities (ACT, 198 verbs or 19.4%), Accomplishments (ACC, 222 or 21.8%) and Achievements (ACH, 515 or 50.5%). The aspectual properties were determined using the main standard tests on aspect (Dowty, 1979; Rothstein, 2004), e.g., the progressive form (*être en train de* in French), which is deemed incompatible with stative predicates, or the *in x time* test, only compatible with telic predicates (accomplishments and achievements). The two experts gave the same judgement for 750 verbs (73.6%), which is an indication that a full-fledged annotation on the whole set of frames would yield reasonable inter-annotator agreement scores.

First Results (Correspondence Analyses)

As can be seen in figure 1, all 4 aspects seem clearly separated: each aspect appears in a distinct quadrant of the plane. Telic (ACH or ACC) verbs seem correlated with a direct object (OBJ), while ETAT or ACT verbs are not. While ACT verbs appear correlated with intransitive constructions ("NO" tag: the argument is not realized), this does not necessarily mean that we have a clear indication of a strong telicity/internal argument (Argument2) association. Since only surface construction data is available so far, unaccusative vs. unergative constructions will have to be further distinguished.

In figure 2, the main conclusion seems to be that Argument3, whether realized as a P-OBJ (an indirect object other than A-OBJ or DE-OBJ) or absent (NO), fails to discriminate between the 4 different aspectual classes. Therefore, the influence of Argument3 on aspect does not seem decisive, here.

Conclusion and Perspectives

Despite a considerable aspectual annotation effort, we must emphasize that the distribution of aspect classes with respect to construction frames is heavily skewed in the FTB, which is attributable to both the nature of the corpus (extracts from only one

source: the *Le Monde* newspaper), and to its size (*ca.* 1 million words, and *ca.* only 1000 verbs studied here). In other words, more data, from varied sources, and annotated along the same protocol, are clearly needed to determine aspectual/syntactic dependencies with more confidence.

References

Abeillé, A., L. Clément, and F. Toussenel. 2003. Building a treebank for French, in A. Abeillé (ed.), *Treebanks*, Kluwer, Dordrecht, pp.165-187

Balvet A., L. Barque, M.-H. Condette, P. Haas, R. Huyghe, R. Marín, A. Merlo. 2012. La ressource Nomage, Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus, *TAL*, Volume 52 – n° 3/2012, pp.1-24.

Dowty, D. 1979. Word meaning and Montague Grammar. Dordrect: Reidel.

Kupść, A. and A. Abeillé. 2008. Growing TreeLex, in A. Gelbukh (ed.), *9th International Conference CICLing*, Lecture Notes in Computational Linguistics, 4919, pp.28-39.

Rothstein, S. 2004. *Structuring Events.* Oxford: Blackwell.

Figures

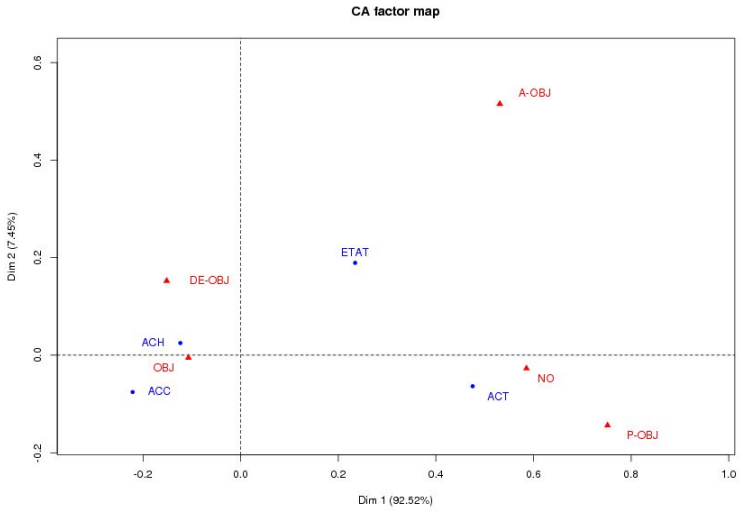

Figure 1: Correspondence Analysis, **Aspect x Argument2** as factors

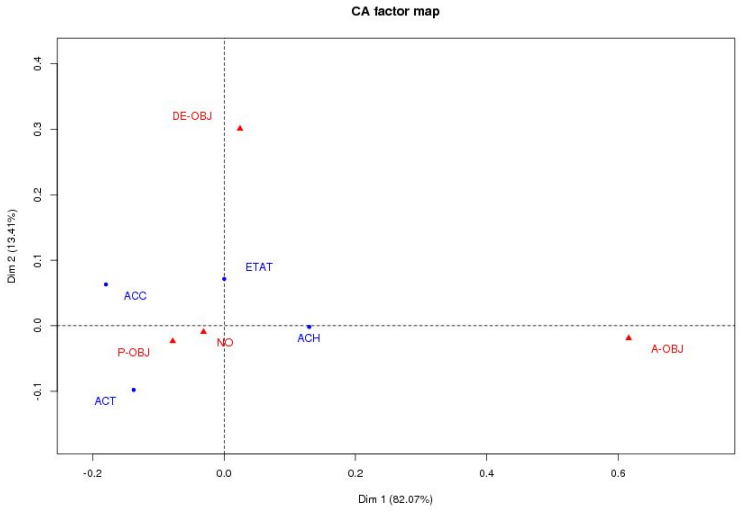

Figure 2: Correspondence Analysis, **Aspect x Argument3** as factors