Growing TreeLex

Anna Kupść[1] and Anne Abeillé[2]

[1] Université de Bordeaux, ERSSàB/SIGNES and IPIPAN;
Université Michel de Montaigne, Domaine Universitaire, UFRL
33607 Pessac Cedex, France
[2] Université Paris7, LLF/CNRS
UMR 7110, CNRS-Université Paris 7, Case 7031, 2, pl. Jussieu
75251 Paris Cedex 05, France
akupsc@u-bordeaux3.fr, anne.abeille@linguist.jussieu.fr

Abstract. TreeLex is a subcategorization lexicon of French, automatically extracted from a syntactically annotated corpus. The lexicon comprises 2006 verbs (25076 occurrences). The goal of the project is to obtain a list of subcategorization frames of contemporary French verbs and to estimate the number of different verb frames available in French in general. A few more frames are discovered when the corpus size changes, but the average number of frames per verb remains relatively stable (about 1.91–2.09 frames per verb).

Key words: Verb valence, subcategorization, treebank

1 Introduction

The paper presents TreeLex, a subcategorization lexicon for French, automatically extracted from a syntactically annotated corpus.

Information about the combinatory potential of a predicate, i.e., the number and the type of its arguments, is called a subcategorization frame or valence. For example, the verb *embrasser* 'kiss' requires two arguments (the subject and an object), both of them realized as a noun phrase, whereas the predicative adjective *fier* 'proud' selects a prepositional complement introduced by the preposition *de*. This kind of syntactic properties is individually associated with every predicate, both within a single language and cross-linguistically. For example, the English verb *miss* has two NP arguments but the second argument of its French equivalent *manquer* is a PP (and semantic roles of the two arguments are reversed). This implies that subcategorization lexicons which store such syntactic information have to be developed for each language individually.[3] In addition to their importance in language learning, they play a crucial role in many NLP

[3] Work on mapping theory has revealed partial correlations between lexical semantics and subcategorization frames, see for example [10] for linking relations of verbs' arguments. We are not aware of any similar work done for other types of predicates, e.g., adjectives or adverbs.