The TALANA annotated corpus for French: some experimental results

Abeillé Anne, Clément Lionel, Kinyon Alexandra, Toussenel François
TALANA-UFRL. Université Paris 7
{abeille, clement, kinyon, ftoussen}@linguist.jussieu.fr

This paper presents the first linguistic results exploiting the new annotated corpus for French developed at Talana-Paris 7 (Abeille & al 00). The corpus comprises 1 million words fully annotated and disambiguated for parts of speech, inflectional morphology, compounds and lemmas, and partially annotated with syntactic constituents. It is representative of contemporary normalized written French, and covers a variety of authors and subjects (economy, literature, politics, etc.), with extracts from newspapers ranging from 1989 to 93.

After explaining how this corpus was built, we present some linguistic results obtained when searching the corpus for lexical or syntactic frequencies, for lexical or syntactic preferences, and explain why we think some of these results are relevant both for theoretical linguistics and psycholinguistics.