

A Quantitative Study of German Compound Nouns and their Polish Equivalents

We present preliminary results of a quantitative, contrastive study of the structural and semantic relations that arise between German compound nouns (GCNs) and their Polish equivalents. The main premise has been to use only automatic tools for the collection of the data as well as for the analysis of the collected parallel pairs. We show that interesting linguistic insights can be obtained despite of the use of generally error-prone methods as long as the analysed data set is large enough and as long as it remains large even after various filtering techniques have been applied.

The topic of GCNs and their Polish equivalents has been investigated before, the most extensive study so far being Jeziorski (1983). Jeziorski relies on a set of roughly 3000 GCNs and corresponding Polish phrases that were manually extracted from handbooks on German word-formation and bilingual German-Polish dictionaries. Mainly aspects of word formation and syntax are contrasted, semantic aspects have not been investigated.

Our study is based on a large parallel corpus — the German-Polish part of the third release of the JRC-Acquis parallel corpus (Steinberger et al. 2006) which is a subset of the body of law of the European Union ranging from 1956 to 2006. We collected 2,163,620 GCN tokens which correspond to 144,207 GCN types. For about 50,000 compound noun types we were able to identify their Polish equivalents with an precision of approximately 93% using statistical alignment models and additional linguistical knowledge for filtering.

We describe our methods of automatic data analysis for both languages, including splitting and semantic interpretation of the GCNs, syntactic parsing of the Polish equivalents, bracketing of GCNs using structural evidence from their Polish counterparts, and mutual semantic disambiguation using a parallel German-Polish thesaurus. All methods of analysis are evaluated against a manually annotated test set.

An overview of structural patterns identified for both language is given. We contrast part-of-speech structures of the GCN segments, bracketing structures and automatically identified semantic relations between the GCN segments with the syntactic and semantic structure of their Polish equivalents. All results are ordered by statistical significance.

The impact of errors introduced due to the application of unsupervised methods is discussed and examples how these error can be minimised using evidence from parallel data are given.

The perhaps greatest advantage of using fully automatic methods — once they have been developed and tested — is the ease of reapplying them to other sources of data the moment they become available. We show how our results improve after adding parallel data from the European Union published in 2007 which is not yet included in the third version of the JRC-Acquis.

References

- Jeziorski, J., 1983. *Substantivische Nominalkomposita des Deutschen und ihre polnischen Entsprechungen*. Wydawnictwo Polskiej Akademii Nauk.
- Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga, 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058. Informal publication.