Lexis-in-context: Methods and techniques for register discrimination

Elke Teich

TU Darmstadt

The research reported on is placed in the context of functional linguistic variation (register) and focuses on contrasts and commonalities in scientific registers. The data set taken as a basis is the *Darmstadt Scientific Text Corpus* (DaSciTex), which is composed of English academic articles from nine scientific disciplines comprising around 17 million words. The texts are taken from four interdisciplinary fields of research (computational linguistics, bioinformatics, digital construction, microelectronics), their corresponding disciplines of origin (linguistics, biology, mechanical engineering, electrical engineering) and computer science.

The research questions posed are of the following kind: What are the linguistic reflexes of two registers coming into contact with one another (as, for instance, in computational linguistics)? To what extent are the linguistic conventions of the discipline of origin (e.g., linguistics) retained and to what extent are the ones of computer science adopted? Or, alternatively, are there new, distinctive registers emerging? Approaching these questions is essentially an exercise in corpus comparison and involves working out the linguistic differences and commonalities between (comparable) corpora. Given that register variation is manifested in the relative frequency of occurrence of particular lexico-grammatical features according to situational context, the central issue is to find those features that are good discriminators between registers. While lexis is clearly the prime distinctive feature between registers, lexical differences do not allow for much of an interpretation beyond field of discourse ("two (sets of) texts t1 and t2 *are about* different things"). In contrast, if lexis is considered in its (grammatical) context, the discriminatory power of any differences found may not be as high, but the conclusions that can be drawn are potentially more revealing ("two (sets of) texts t1 and t2 construe *x* in different ways").

Presenting selected analysis of *lexis-in-context* carried out on the DaSciTex corpus, we discuss different ways of approaching register variation:

- *analysis approach*: univariate vs. multivariate
- *initial disposition*: hypothesis vs. no-hypothesis
- *basis for analysis*: shallow (linguistically uninterpreted) vs. higher-level (linguistically interpreted)
- *interpretation of analysis*: token level vs. aggregated level

We conclude with some reflections on the implications for modeling *lexis-in-context* for the purpose of discriminating between registers.

Acknowlegments

The project 'Linguistische Profile interdisziplinärer Register' was funded by DFG from 2006-2009. My collaborators in the work presented are Monica Holtz, TU Darmstadt, and Peter Fankhauser, L3S Hannover.

References

Teich, E. and M. Holtz, 2009. Scientific registers in contact: An exploration of the lexico-grammatical properties of interdisciplinary discourses. *International Journal of Corpus Linguistics* 14(4): 524—548

Teich, E. and P. Fankhauser, forthcoming. Exploring a corpus of scientific texts using data mining. In: S. Gries, S. Wulff, M. Davies (eds), *Corpus linguistic applications: Current studies, new directions.* Rodopi, Amsterdam