Abs-91

Sabine Bartsch (Technische Universität Darmstadt), Elke Teich (Universität des Saarlandes), and Christoph Tragl (Technische Universität Darmstadt)

Patterns of cohesion in informationally dense texts

Scientific writing is often perceived to deliver a large amount of information in a compact manner. This impression is commonly attributed to features at the level of lexico-grammar, such as domain-specific terminology, complex nominal groups, and a high lexical density. However, there are features at the level of text that contribute to (high) information density, too. One such feature is cohesion, i.e. the choices that make a text hang together in terms of reference, ellipsis, substitution, lexical cohesion, and conjunction (cf. Halliday & Hasan, 1976).

In this paper we examine patterns of cohesion in scientific abstracts. Abstracts are chosen because they present information in an extremely aggregated manner, thus exhibiting a particularly high information density (cf. Swales 1990; Biber 2006). The corpus of abstracts under study is a subcorpus of the Darmstadt Scientific Text Corpus (DaSciTex) (Teich & Holtz 2009) consisting of around 2.000 texts from nine scientific disciplines (17 million words). We have investigated abstracts from four disciplines included in DaSciTex (computer science, linguistics, biology and mechanical engineering), taking samples of ten abstracts per discipline. The corpus is automatically pre-annotated by means of the tool Little Cohesion Helper (LCH) (Bartsch et al., 2009), which was developed for the purpose of cohesion annotation. LCH automatically identifies lexical cohesive chains within a text on the basis of lexical semantic relations represented in the Princeton WordNet (Fellbaum et al. 1998). Since automatic annotation does not achieve a hundred percent accuracy, the annotation needs to be manually corrected. The other types of cohesion (reference, ellipsis, substitution and conjunction) have been manually annotated by means of MMAX2 (Müller & Strube 2006). Since the output of LCH is mapped into the XML-format used by MMAX2, the integration of annotations is straightforward.

The aim of our study is twofold. First, we are interested in possible differences and commonalities in the usage of cohesive devices across disciplines (types of cohesion, density and length of cohesive chains). The second aim is to investigate whether there are any differences in patterns of cohesion in scientific abstracts vs. other, less technical texts. To this end, we have carried out a comparative study using parts of the FLOB corpus. In terms of the usage of cohesive devices, we expect that abstracts exhibit a relatively frequent use of lexical cohesion, rather little use of reference and rather few occurrences of cohesive conjunction. In the case of relative uniformity of cohesive patterns across disciplines and relative distinctness to the texts taken from FLOB, we can conclude that the abstract constitutes a text type/genre on its own (independent of register); in the case of relative diversity across disciplines, we conclude that we encounter a case of register (i.e. domain-specific) variation (with no discrete text type/genre). In the paper we present the results of both studies and their interpretation in terms of register vs. genre attribution.

References:

Bartsch, Sabine et al. (2009): “ObamaSpeeches.com. Building and Processing a Corpus of Political Speeches. A Student Project.“ Poster presentation at the Herbsttagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL) 2009 an der Universität Potsdam, September 2009.

Biber, Douglas. (2006): University language : a corpus-based study of spoken and written registers. Amsterdam, Philadelphia: John Benjamins.

Fellbaum, Christiane (1998, ed.): WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Halliday, MAK & Ruqaiya Hasan (1976): Cohesion in English. Harlow: Longman.

Müller, Christoph & Michael Strube (2006): Multi-Level Annotation of Linguistic Data with MMAX2. In: Sabine Braun, Kurt Kohn, Joybrato Mukherjee (Eds.): Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3).

Swales, John. (1990): Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Teich, Elke & Mônica 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.