Understanding Science

A German popular science corpus

Uli Held[1], Karin Maksymski[2], Oliver Čulo[2] and Silvia Hansen-Schirra[2]

[1] Albert Ludwigs University of Freiburg
uli.held@germanistik.uni-freiburg.de

[2] Johannes Gutenberg University Mainz
maksymsk@uni-mainz.de

Funded by: VolkswagenStiftung

Principal investigators: Prof. Dr. Peter Auer[1], Prof. Dr. Silvia Hansen-Schirra[2], PD Dr. Lars Konieczny[1]

Project – Background to corpus construction

Motivation and Goals
Scientific knowledge becomes ever more important in understanding and solving the problems of our modern world; comprehensible and engaging reporting is therefore crucial. However, whether popular science writers are successful in conveying complex topics in a comprehensible way, remains an open question.

Research questions
- Where do readers have difficulties and how can these difficulties be avoided?
- In which respect do good and bad articles differ?
- How much and which knowledge is transferred in which way?

Method
- Readers perception of the texts is measured through questionnaires.
- Additional eye-tracking studies help us reveal problematic areas in the texts.
- This data is then correlated with measures of linguistic complexity and characteristic features of popular science writing.
- According to the results recommendations to optimize the comprehensibility will be formulated and tested.

Corpus

Main Corpus
ca. 500,000 words, texts from 2005 to 2011

long texts (~450,000 words)

newspapers	science-oriented	lay-oriented
(FAZ, SDZ, ZEIT)	(BDW, SDW)	(GEO, PM, WDW, ZW)
~130,000 words	**~150,000 words**	**~170,000 words**

short texts (~50,000 words)

newspapers	science-oriented	lay-oriented
~16,000 words	**~14,000 words**	**~20,000 words**

Covered domains: biology (incl. genetics), chemistry, engineering, geology, mathematics, medical science, physics (incl. astronomy), psychology

Variation in: length, domain, authors, publication format, sources (target audience, style)

General Language Corpus
Tiger, Zeit, FAZ

Science Language Corpus
domains as in main corpus

Diachronic Corpus
1975 to 1985

Annotation

Almost any linguistic level of description can play an important part in the process of understanding a written text. Although there are always simpler measurements like length or frequency, we want to get to the root of the problems readers might have. With limited resources the only way to produce such a high number of different annotations is to use a processing pipeline which combines manual and automatic methods.

Word

Word complexity
Especially in German, complex nominal compounds are an indicator of a nominal style, which is typical of scientific writing.

Vocabulary
Foreign words (from Latin, Greek, English), technical terms, named entities, word frequencies

Metaphors
We are especially interested in metaphors which serve as an explanatory tool.

Morpho-syntactic categories
Part of speech, tense, number, case, voice, gender, comparative

Sentence

Syntax
All texts in the main corpus are syntactically annotated with a dependency structure. The annotation scheme is derived from the Tiger annotation scheme.

Direct and indirect speech
The quotation of scientists is a popularization strategy often used by authors.

Text

Cohesive structure
To capture the cohesive structure of a text, we annotate lexical chains, coreference, and anaphora (planned).

Information structure
Shifts in information structure are not observed directly but through relevant features like definiteness, word order, new / old entity.

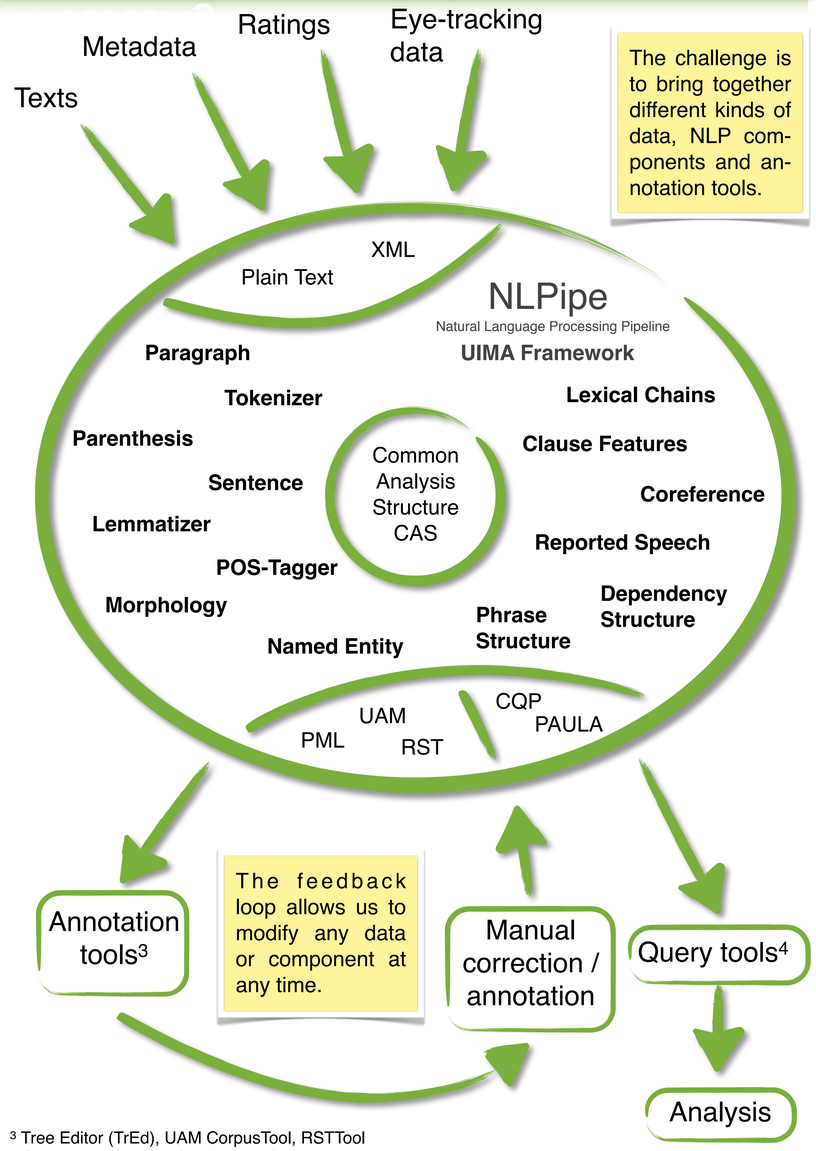
Rhetorical structure
To describe the rhetorical structure we use rhetorical structure theory (RST) with a reduced and modified relation set.

Discourse strategies
In the popular science register a limited number of discourse strategies can be found, such as descriptions of a workday life, the history of science or the explanation of scientific facts.

Article

Macro structure
A popular science article does not simply consist of its main text. It is structured through headings, passages, figures, pictures, captions and other text elements.

Processing



The challenge is to bring together different kinds of data, NLP components and annotation tools.

[3] Tree Editor (TrEd), UAM CorpusTool, RSTTool
[4] Corpus Workbench (CWB), ANNIS2

Literature

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Brants, S. & S. Hansen (2002). Developments in the TiGer Annotation Scheme and their Realization in the Corpus. In: Proceedings of LREC. Las Palmas. 1643-1649.

Hansen-Schirra, S., S. Hansen, L. Konieczny & S. Wolfer (2009). Fachkommunikation, Popularisierung, Übersetzung: empirische Vergleiche am Beispiel der Nominalphrase im Englischen und Deutschen. In: Linguistik online 38 2/2009.

Niederhauser, J. (1996). Wissenschaftliche Fachsprache und populärwissenschaftliche Vermittlung. Linguistische Untersuchungen zur fachexternen Wissenschaftskommunikation. Dissertation. Bern.