Vitor Lécio Lacerda Fontanella[1,2], Tom Bleckmann[2]
Lukas Dieckhoff[2], Gunnar Friege[2], Christian Wartena[1]
([1]Hannover University of Applied Science and Arts; [2]Leibniz University Hannover)

TeCoPhy: A Text Corpus of German Physics Texts

Keywords: *Corpus construction, German, Physics, Textbooks*

To learn a subject, the acquisition of the associated technical language is important (Diethelm & Goschler, 2014; Pineker-Fischer, 2017; Poupova, 2018). Despite this widely accepted importance of learning the technical language, hardly any studies are published that describe the characteristics of most technical languages that students are supposed to learn. This might largely be due to the absence of specialized text corpora to study such languages at lexical, syntactical and textual level. In the present paper we describe a corpus of German physics text that can be used to study the language used in physics. The composition of such a corpus faces three major challenges:

1. We have to deal with OCR and the complicated layout of textbooks;
2. Physics texts contain a large number of symbols and formula.
3. Due to copyright restrictions, a corpus of texts from textbooks cannot be published.

Our primary goal was to have a large collection of German texts on physics covering various topics and different levels of proficiency, including at least some texts intended for secondary school students. Thus we included Wikipedia articles from the category physics (excluding articles about institutions and biographies of physicists), articles on school physics from the website `https://www.leifiphysik.de/`, as well as many (printed) textbooks (at secondary school and university level) and a few scientific books. Initially 264 books were scanned. Books with severe OCR problems were just removed from the collection. 221 books could be used for further processing.

We extracted the text from the scanned books using PDFMiner[1]. To avoid problems with footers, page numbers, captions, etc., we determined the main fonts used in each book and extracted only text blocks using these fonts. After sentence splitting, only sentences having at least 50% alphabetical characters are kept. Finally, we removed English sentences, appearing e.g. in quotes. Thus, the corpus is a collection of sentences rather than a collection of coherent texts. On average 47% of the text could be extracted, resulting in $2.36 \cdot 10^5$ sentences or $5.3 \cdot 10^6$ tokens.

According to the German copyright laws we are allowed to distribute (still with restrictions) at most 15% of the text of each book. Thus we have to make a subselection of the texts. To guarantee the presence of enough terminology, we extracted a list of nouns occurring more than 5 times in the corpus and having a higher relative frequency than the word has in the German Reference Corpus DeReKo (Kupietz & Lüngen, 2014). We added moreover around 600 words that occur in typical collocations. This results in a list of 30.681 words. Half of the small corpus was constructed by selecting between 5 and 10 example sentences for each noun. The other half

[1] `https://github.com/pdfminer/pdfminer.six`

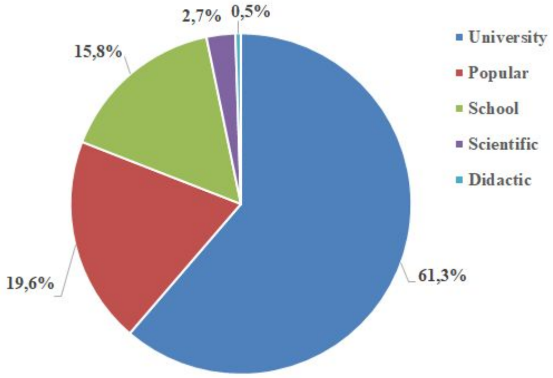

Figure 1: Composition of the corpus.

was selected by random sampling from the remaining sentences. It was guaranteed that at most 14% of each book was included.

The selection consists of $2.36 \cdot 10^5$ sentences and $5.32 \cdot 10^6$ words. The composition of this selection from different types of sources is given in Figure 1.

In order to see how representative the selection is for the large corpus, we extracted from the large corpus a list of words that occur at least 10 times. This list contains $45,332$ word forms, $39,317$ of which are also found in the small corpus. The Pearson correlation of the frequency values in both corpora is $0,999$. If we extract from both corpora the 1000 word forms with the highest relative frequency (compared with the DeReKo data), 904 words are included in both lists.

The large corpus cannot be distributed. The smaller variant of TeCoPhy can only be distributed for research on text and data mining and is available on request from the authors.

References

Diethelm, I., & Goschler, J. (2014). On human language and terminology used for teaching and learning cs/informatics. In *Proceedings of the 9th workshop in primary and secondary computing education* (p. 122–123). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/2670757.2670765` doi: 10.1145/2670757.2670765

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. In (p. 2378-2385). Reykjavik: European Language Resources Association (ELRA). Retrieved from `https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31353`

Pineker-Fischer, A. (2017). Von der Alltags- zur Bildungs- und Fachsprache. In *Sprach- und fachlernen im naturwissenschaftlichen unterricht: Umgang von lehrpersonen in soziokulturell heterogenen klassen mit bildungssprache* (pp. 41–82). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from `https://doi.org/10.1007/978-3-658-16353-2_5` doi: 10.1007/978-3-658-16353-2_5

Poupova, J. (2018). Biological terminology: an opportunity for teaching in tandem. In *International conference new perspectives in science education* (pp. 382–385).