CorTeDo. The first representative corpus of technical documentation

Franziska Zellner

Abstract: This essay presents the M.A. project *CorTeDo* (Corpus of Technical Documentation), in which a corpus of technical documentation was compiled. With the aid of the corpus, claims about technical documents, such as their limited lexical and syntactic variation, were revisited. Additional characteristics were unearthed with an array of methods such as keyword analysis, n-gram analysis and collocation analysis.
This research is a contribution to the fields of technical documentation, genre-based teaching and linguistics. It promotes the understanding of different genres and uses of language, especially those that have traditionally been overlooked. In the future, *CorTeDo* will be expanded to form a group of corpora that can be utilised as the basis for further analysis in all of the above fields and beyond.

About the author: Franziska Zellner studied and taught English Linguistics at the University of Regensburg. The following essay is based on her MA thesis. Supervisor: Dr. Thorsten Brato and Prof. Dr. Jakob Leimgruber

Keywords: technical documentation; corpus linguistics; lexical richness; syntactic complexity; keyword analysis

> Do not fail to follow the instructions.
> (Manual for an infant incubator included in the corpus)

Depending on the type of technical product, *failing to follow the instructions* may have more or less serious consequences. Obviously, an infant incubator is very different from an electric pet clipper, a milk frother or an industrial shredder. Regardless of their purpose or intended group of users though, all of these products need to be accompanied by technical documents that provide information and instructions on their installation, use and maintenance. As more and more such products enter the global market every day, there is a huge demand for technical documentation. Given its importance, it is surprising that this subgenre of technical writing has been somewhat neglected in linguistics so far. In fact, as Lassen (2003: 145) notes, technical documentation accounts for 70 % of worldwide technical communication, but most research focuses on other subgenres such as lab reports or scientific publications, which have a very different purpose and audience compared to, for example, an instruction manual. This research gap needs to be filled by linguistic analysis on the basis of corpora that represent the subgenre of technical documentation.

The project *CorTeDo* (Corpus of Technical Documentation) aims to compile a representative corpus of procedural texts, which are the most central type of technical documents (Saint-Dizier, 2014: 1), to test existing claims about characteristics of the subgenre and to unearth new characteristics. *CorTeDo* is beneficial not only for linguistic research in understanding a genre and use of language that has traditionally been overlooked, but can also be used to facilitate practical technical writing and to train new technical writers with genre-based teaching.

Technical documentation

Before delving into technical documentation, it must be distinguished from its related fields by disentangling various definitions. Often, terms such as *technical communication*, *technical writing* or *technical documentation* are not clearly defined, but for the purpose of this project, I devised a framework to categorise them, which is illustrated in Fig. 1.

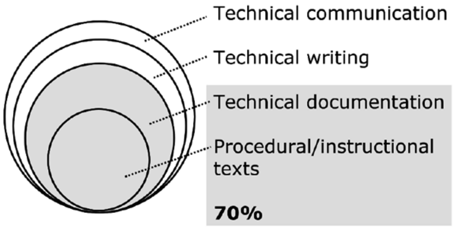

Fig. 1: Overview of technical communication, technical writing, technical documentation and procedural/instructional texts

Technical communication, according to Dobrin et al. (2014: 4), is the "communication about complex, highly detailed problems, issues, or subjects in the professional world". This type of communication is a complex process with a heterogeneous group of actors, not only the communicators and their audience, but also subject matter experts, illustrators or translator, just to name a few (Byrne, 2006: 61).

When technical matters are communicated in written form, it is called technical writing. This includes a range of artefacts, for example manuals, lab reports, articles for a technical audience or company newsletters (Byrne, 2006: 60).

When it comes to technical documentation, internal and external technical documentation must be differentiated. While internal documentation remains within the company that manufactures the product or provides the documentation, external documentation is intended for the customer or other actors. The external type is the focus of *CorTeDo*, as it encompasses carefully constructed texts with which various target audiences come in contact.

Regarding the characteristics of external documentation, most definitions share the following commonalities.

- A particular, very clear organisation and structure (Chagheri / et al., 2011: 808; Dobrin et al. 2014: 5; Saint-Dizier 2014: 1), which is signalled by headings, subheadings or lists.
- A limited variation on different linguistic levels, especially regarding "lexical realisation and grammatical and style constructions" (Saint-Dizier, 2014: 1).
- A high precision and unambiguity that make documents user-friendly (Blake & Bly, 1993: 3 f:).
- Multimodality, as technical documents usually combine text with graphic elements like screenshots, graphs (Saint-Dizier, 2014: 2, Povolná, 2020: 230), technical drawings or photos.

What most of the previous linguistic approaches to technical documentation share is their focus on procedural (or instructional) texts, which are the most central type of document (Saint-Dizier, 2014: 1). These texts instruct on how to do certain tasks, which may be decomposed into subtasks. To make the instructions easy to follow, even complex actions and issues in the real world should be made linear and unambiguous in procedural texts (Saint-Dizier, 2014: 10).

In terms of target group, many studies distinguish between technical documents for laypersons and for trained staff, but there are no widely agreed upon terms for these different types. For the project *CorTeDo,* the technical documents under investigation are called *procedural texts*, as this stresses both their function and their being a textual subgenre that can be studied linguistically. These procedural texts may be aimed at either lay users or experts.

Design and compilation of *CorTeDo*

The theoretical and methodological basis on which the project *CorTeDo* is built is Corpus Linguistics. A corpus is a structured, computer-readable collection of linguistic data. This data may be written texts or transcriptions of oral communication. In Corpus Linguistics, such corpora are examined for linguistic patterns, either to test existing hypotheses or to make new observations. The former represents the corpus-based approach, while the latter is the corpus-driven approach. For these analyses, so-called concordance software and related tools are used to filter out linguistic patterns from the corpus or to carry out statistical analyses. To ensure that the results of these analyses do not only apply to the corpus itself but can also be generalised, corpora are designed and compiled with great care in order to be representative. Ideally, they represent a specific language, a selected genre or another application in which language is used.

Since technical documentation as a genre has been neglected in linguistics so far, there are no representative corpora of it, either. This means that before any analyses could be carried out, a new corpus needed to be designed and compiled. Since it is the first of its kind, it is called the *Corpus of Technical Documentation* – in short, *CorTeDo*. Then, both corpus-based and corpus-driven analyses were carried out, as will be illustrated later.

CorTeDo aims to represent typical technical documentation. Accordingly, the focus lies on procedural texts in their classic print or PDF format, which despite current trends such as augmented reality (AR) or apps still remains the most central and widespread medium. In fact, traditional instructions are the basis for any of these more elaborate formats (Huber / Lierheimer, 2022: 15) and therefore will continue to be relevant.

To find the best structure for *CorTeDo*, the industries in which procedural texts are usually found were narrowed down: Machinery and plant engineering, the automobile industry, consumer electronics and IT plus a few more (tekom, 2022). Each of these will be represented in a subcorpus of *CorTeDo*, as shown in Fig. 2.

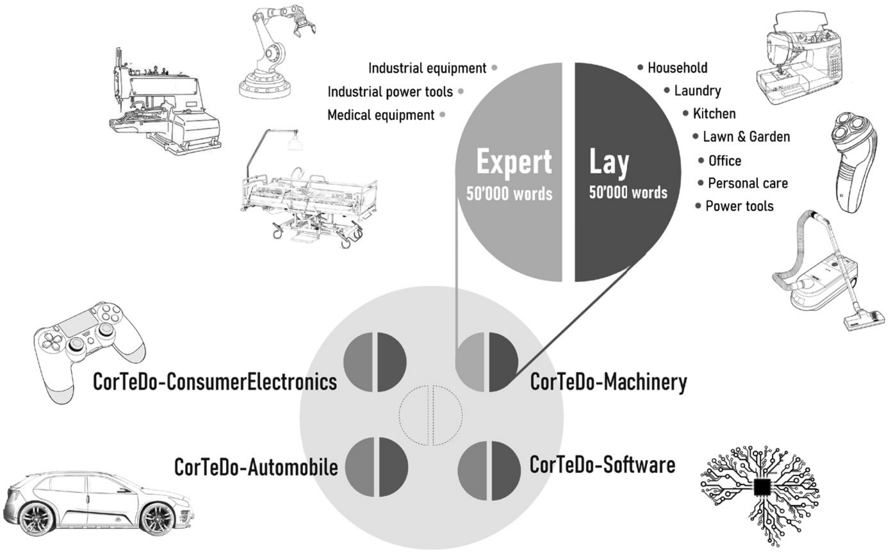

Fig. 2: Structure of CorTeDo, including its subcorpora

As a starting point and model for further subcorpora, the industry of machinery and plant engineering was selected. For the corresponding subcorpus *CorTeDo* -Machinery, procedural texts that are aimed at experts and laypersons were collected from open-access online sources. Industrial machines are represented alongside those used in the home: The industrial sewing machine used by professional tailors is juxtaposed with the sewing machine used by hobby sewists, the large industrial saw used in the woodworking industry complements the small jigsaw that DIY enthusiasts have in their toolboxes. In short, *CorTeDo* -Machinery provides a picture of the wide range of products and their instructions. Random extracts of approx. 1,000 words were taken from almost 100 manuals and saved in txt format. Thus, *CorTeDo* -Machinery currently comprises approx. 100,000 words, half of which come from expert documentation and the other half from lay documentation. This corpus size already allows for robust analyses of sentence structure and vocabulary by contrasting expert and lay documentation.

In addition to the collection of the raw data in txt format, there is a second version of the corpus that is enriched with part-of-speech tags (POS tags). This means that each individual word is automatically labelled according to its word form, for example whether it is a noun, a verb or an adjective. This labelling is carried out by software that can assign tags from a standardised tag set with a high degree of accuracy. The resulting version of the corpus with POS tags allows for a more precise analysis of sentence structure or types of words that tend to occur together.

Syntactic complexity and lexical richness

As mentioned above, low syntactic complexity and limited lexical richness have been reported by diverse sources to be two main characteristics of technical documentation. Unfortunately, the literature remains rather vague here: What exactly is *low syntactic complexity* and *limited lexical richness* in this context? Finding satisfying answers to this question constituted the corpus-based component of the analysis. To get a bigger picture, *CorTeDo* -Machinery was compared to two reference corpora: *COCA*, a general language corpus of English, and the *IULA Technical Corpus*, which represents different types of technical writing.

Regarding syntactic complexity, which measures "how varied and sophisticated the production units or grammatical structures are" (Lu, 2010: 474), several measures were selected, for example the length of sentences or the number of clauses per sentence. The following pattern emerges: *CorTeDo* -Machinery has shorter sentences overall, fewer complex clauses and less coordination than the two reference corpora. For example, an average sentence consists of 15 words in *CorTeDo* -Machinery, 22 words in the technical writing corpus and almost 25 words in the general English corpus, as illustrated in Fig. 3. Within *CorTeDo* -Machinery, the expert documentation appears to be slightly more complex in its sentence structure than the non-expert documentation, but not significantly so. Further research is needed to make concrete statements on this.

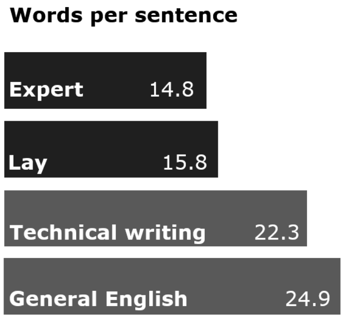

Fig. 3: Number of words per sentence in CorTeDo-Machinery, IULA Technical Corpus and COCA

At the same time, the results also show that the digital tools that are commonly used for syntactic analysis in linguistics have their difficulties with technical documentation. The reason lies in its frequent use of sentence fragments in headings, figure descriptions or bullet points as well as the use of special characters and numbers in the text. Since these are the structures that are characteristic of the genre, it may make more sense to concentrate on analyses other than sentence structure.

Thus, the analysis turned to the vocabulary to understand what *limited lexical richness* looks like in technical documentation. The results revealed that the number of different words in the expert and lay corpus is similar and overall rather low. However, there is a difference in the types of words that are used. The expert documentation contains more words that describe very specific concepts, as well as longer compound words. While compounds in lay documentation rarely consist of more than two components, compounds with three or four components are not uncommon in expert documentation. It can be concluded that expert documentation seems to be characterised by a more sophisticated vocabulary, albeit not at the level of word diversity in the overall text, but rather at the level of the terminology itself. In other words: the same words are used repeatedly, but in expert documentation these words are linguistically more complex and sophisticated from the outset.

Lay and expert users

In addition to revisiting existing claims about lexical and syntactic characteristics, i.e. corpus-based research, additional corpus-driven analyses were carried out to unearth new characteristics of technical documentation. Specifically, this included the analysis of keywords, n-grams and collocations. Keyword analysis extracts those words that are unusually frequent in one corpus as compared to another. These words are called keywords and are a strong indicator for linguistic and thematic idiosyncrasies of a corpus. n-grams, on the other hand, are multi-word units that are especially frequent. *n* stands for the number of consecutive words: A trigram or 3-gram consists of exactly three words. Lastly, collocation analysis is concerned with words that commonly appear near each other. Both n-grams and collocations are indicators for linguistic patterns in a corpus and serve as a good starting point for more detailed analyses.

The results show additional differences between the lay and the expert subcorpus, especially in terms of the interaction with the users. The interaction with lay users is very direct and highlights their personal experience. Not only are the personal pronouns *you* and *your* keywords in the lay subcorpus, but they also collocate with words that express a close relationship between the author and the user, such as *we, thank, recommend,* or words that refer to the body and the personal experience of the user, such as *skin, hand, personalizing, feel* and *touch.* In the expert documentation, these patterns are absent. *You* and *your* are rather rare there, and if *your* does appear, it is usually followed by the product itself, as in *your vehicle* or *your machine.* Thus, the interaction is much more impersonal. Additionally, expert documentation often makes use of the passive, while lay documentation has a strong preference for active verb forms.

At the same time, expert documentation is characterised by a sense of urgency – which is adequate for the heightened danger of larger machines. Many n-grams are concerned with possible dangers, for example the 4-gram *severe injury or death* – which (luckily) is missing in lay documentation. Also, the auxiliary verb *must* is very frequent in expert documentation and seems to further emphasise urgency. Thus, the user is made aware of dangers not only through the use of signal words, pictograms or colours, but also through word choice and verb forms.

A look into the future of *CorTeDo*

The analysis showed that *CorTeDo* -Machinery, the subcorpus for machinery and plant engineering, is a valuable basis for a number of linguistic analyses. Once other industries are represented with their own subcorpora, as well, the beneficial effects will be even greater. As Fig. 2 showed, these subcorpora will be *CorTeDo* -ConsumerElecronics, *CorTeDo* -Automobile and *CorTeDo* -Software. This is not an exhaustive list, however: Other important industries can be added in the process. By doing so, the overall number of words will increase and allow for more robust analyses. Also, comparisons between industries will become possible. What are the differences between procedural texts for machines and those for software? Are there linguistic patterns that technical writers may have to adapt to when they move from one industry to another?

Additionally, since technical writing is inherently multilingual, *CorTeDo* should be compiled in languages other than English. Doing so opens the possibility of comparing these languages. How do English instructions differ from German ones? Are there hidden challenges for translation? A very common scenario: What do technical writers need to be aware of when they write in a language that is not their first language?

CorTeDo can help both new and experienced technical writers to get a better understanding of technical documentation as a genre, including its differences between industries and target groups.

To complete this big project, further research and cooperation with companies and professional associations are necessary. In turn, multiple areas will benefit from it: practical technical communication, genre-based teaching and of course linguistics.

Literaturverzeichnis

Blake, Gary / Bly, Robert W. (1993): *The elements of technical writing,* New York: McMillan.

Byrne, Jody (2006): *Technical translation: Usability strategies for translating technical,* Dordrecht: Springer.

Chagheri, Samaneh et al. (2011): „Technical documents classification“, in: *Proceedings of the 2011 15th International Conference on Computer-Supported Cooperative Work in Design.*

Dobrin, Sidney et al. (2014): *Technical communication in the twenty-first century*, Columbus, OH: Pearson.

Huber, Bernhard / Lierheimer, Gerhard (2022): „Kompetenz: Welche Kompetenz erfordert die Dokumentation in der Zukunft (aus Sicht der Redaktionsmitarbeiter)“, in: Huber, Bernhard et al. (Hrsg.): *Technische Dokumentation und Informationsmanagement: Intelligent – Wertschöpfend – Visionär* (Eine Publikation des VDMA Arbeitskreises „Technische Dokumentation und Informationsmanagement), S. 14 f.

Lassen, Inger (2003): *Accessibility and acceptability in technical manuals: A survey of style and grammatical metaphor*, Amsterdam: John Benjamins Publishing Company.

Lu, Xiaofei (2010): „Automatic analysis of syntactic complexity in second language writing“, in: *International Journal of Corpus Linguistics*, 15(4), S. 474–496.

Povolná, Renata (2020): „Persuasion in technical discourse: The role of interpersonal metadiscourse markers in user manuals“, in: Dontcheva-Navratilova, Olga et al. (Hrsg.): *Persuasion in specialised discourses*, Cham: Palgrave Macmillan, S. 229–262.

Saint-Dizier, Patrick (Ed.) (2014): *Challenges of discourse processing: The case of technical documents.* Newcastle upon Tyne: Cambridge Scholars Publishing.

Tekom (2022): „Der Beruf des Technischen Redakteurs: Branchen und Arbeitgeber“, in: *Tekom* / tekom.de/technische-kommunikation-das-berufsfeld/arbeitswelt/der-beruf-des-technischen-redakteurs.