


# An Annotated German-Language Medical Text Corpus as Language Resource

Joachim Wermter      Udo Hahn

 Text Knowledge Engineering Lab  
Freiburg University  
Werthmannplatz 1  
D-79098 Freiburg, Germany  
<http://www.coling.uni-freiburg.de/>

## Abstract

We describe the structure of a German-language corpus which contains a variety of medical text genres. Clinical documents (discharge summaries, pathology, histology and surgery reports) are distinguished from non-clinical ones (textbook articles and consumer health care documents from a Web portal). After introducing a medical extension of the general-language STTS tagset which accounts for unique features of the medical sublanguage encountered in these documents, we discuss some of the quantitative properties of the annotations (e.g., distribution patterns of part-of-speech tags).

## 1. Introduction

Linguistic corpora play a key role for empirically grounded NLP, language technology in particular. On the one hand, corpora are used for calibrating and fine-tuning linguistic specifications such that they reflect the quantitative dimension of language use and, thus, help to tame different forms of ambiguity (e.g., in terms of probabilistic parsers and lexical resources). At the same time, they help to enhance the level of robustness when linguistic specifications are lacking or incomplete. On the other hand, corpora are a crucial piece of infrastructure for automated learning of linguistic and conceptual knowledge so that system building in our field does no longer rely on hand-crafted grammars, lexicons and domain ontologies. In any case, the quality of meta-information by which plain linguistic utterances are enriched (i.e., the kind of annotation language being used) is essential for both forms of corpus utilization.

In our lab, research efforts are directed at the implementation of information extraction systems for the medical field (Hahn et al., 2002). In order to meet clinical requirements, both a large coverage of medical domain knowledge and robustness of medical language analysis are key issues in rendering systems for routine use. After setting up a very large medical ontology for the fields of anatomy and pathology (Hahn and Schulz, 2002), we have recently begun to address the robustness issue. In order to adapt our parsing facilities (Hahn et al., 2000) to cope with medical jargon in a more fault-tolerant manner, we created a training and test environment with a major medical language resource component, *viz.* a large annotated medical text corpus.

In Section 2. we describe the structure of this corpus which contains a variety of medical text genres: clinical documents (discharge summaries, pathology, histology and surgery reports) are distinguished from non-clinical ones (textbook articles and consumer health care documents). After introducing a medical extension to the general-language STTS tagset which accounts for unique features of the medical sublanguage encountered in these documents, we discuss some of the quantitative properties of the annotations in Section 3.

## 2. Corpus Description

FRAMED, the FREiburg Annotated MEDicine text corpus, combines a large variety of German-language medical text genres with a focus on clinical reports, and runs about 100,150 tokens in size. The clinical text genres, taken from Freiburg University Hospital, cover discharge summaries, pathology reports, histology reports and surgery reports. The non-clinical ones consist of medical expert texts (taken from a medical textbook, *Manual der Diagnostik und Therapie* (MSD, 1993)) and health care consumer texts taken from a health-centered Web portal.<sup>1</sup>

A quantitative account in terms of the number of sentences, text tokens and types distributed over the different genres is given in Table 1. Table 2 shows the corresponding average sentence lengths and the normalized token/type ratios. For purposes of comparison, we also included these measures for a random sample of approximately 100,000 tokens taken from NEGRA, a 355,000 token-sized annotated German newspaper corpus (Brants et al., 2003).

Text Genre	# sentences	# tokens	# types
discharge summaries	513	7138	2076
pathology reports	1522	20734	3815
histology reports	881	15022	2821
surgery reports	1303	17003	3123
textbook	1222	24347	5372
consumer texts	1053	15906	3522
FRAMED total	6494	100150	20729
NEGRA newspaper	5254	100139	18954

Table 1: Text Genre Distribution and Quantitative Data of the FRAMED Medical Text Corpus plus a random sample of the NEGRA newspaper corpus.

The availability of the clinical texts depends on the successful completion of anonymization processes. Lack of manpower for this task is the reason why one text genre, *viz.* discharge summaries, is much smaller in size than the other ones. Ongoing work is seeking to balance this out.

<sup>1</sup>These texts were obtained from <http://www.netdoktor.de/>, last visited on Feb. 23, 2004.

Text Genre	Average Sentence Length	TTR norm
discharge summaries	12.9 (11.1)	3.4
pathology reports	12.6 (8.6)	3.6
histology reports	16.1 (13.8)	4.8
surgery reports	12.7 (7.4)	3.7
textbook	18.9 (11.7)	3.3
consumer text	14.1 (8.6)	3.6
FRAMED	14.4 (10.8)	3.7
NEGRA newspaper	20.4 (11.5)	2.7

Table 2: Average Sentence Length (standard deviation in parentheses) and normalized token/type ratio for the various text genres. The TTR value for FRAMED is averaged.

A look at the individual clinical text genres (discharge, pathology, histology and surgery reports) reveals that there is a great deal of variability in sentence length, as witnessed by the enormous standard deviations. Except for surgery reports, the standard deviation always amounts to at least two thirds of the average sentence length. This may very well have to do with the way we defined the notion of *sentence*, viz. as a consecutive array of tokens delimited by a period, a question mark or an exclamation mark. Thus, we refrained from including any more specific linguistic criteria (e.g., the presence of a verb).<sup>2</sup> The non-clinical text genres (textbook material and consumer texts) show a less dramatic deviation. This does not come as a surprise, because these medical text genres are of a more standardized and carefully produced sort, both linguistically and stylistically. Especially textbook texts seem to be in line with the newspaper material, both in terms of the average sentence length and the less pronounced standard deviation.

The token/type ratio is a measure in corpus statistics that usually indicates the variety of vocabulary in a text.<sup>3</sup> It is well-known that this measure becomes less reliable when comparing texts of different sizes. Therefore, we normalized the medical texts and the newspaper material by taking a random 7138-token sized sample (the size of the discharge summaries) of each genre. Of the medical texts, histology reports exhibit the lowest vocabulary diversity (4.8 tokens per type), whereas the textbook material shows the highest one (3.3 tokens per type). Of all clinical genres, discharge summaries show the most variation (3.4 tokens per type). This can be attributed to the fact that they are the most articulate and prosaic of all clinical document genres, both in terms of their linguistic form and contents.

Interestingly though, comparing the TTR values of the FRAMED document types against the newspaper one, their sublanguage character becomes evident: the newspaper material shows substantially more variation in vocabulary than any of the medical text genres or FRAMED as a whole.

<sup>2</sup>Especially for clinical texts (i.e., discharge summaries, pathology, histology and surgery reports), this definition is justified since they are often marked by a telegraphic style containing a lot of pseudo-sentences just composed of noun or prepositional phrases, e.g., ‘*Chronic recidivating bronchitis with distinctive bronchietasia*’. Overall, they are characterized by a high degree of linguistic para-grammaticality and stylistic deviation.

<sup>3</sup>Unlike its counterpart type/token ratio, the *token/type* ratio is always greater than 1; a smaller number indicates a more varied vocabulary repertoire.

### 3. Annotating the FRAMED Text Corpus

The manual linguistic annotation of text corpora is a prerequisite for the development of standard NLP tools, such as part-of-speech taggers, phrase chunkers, syntactic parsers, grammar and lexicon learners. Up until now, the creation of these kinds of resources has almost exclusively focused on general-language newspaper and newswire genres. The annotation of FRAMED is meant to fill this gap for a particular sublanguage domain, viz. German medical language. While we focus on the part-of-speech (POS) annotation of a medical corpus, (Vintar et al., 2002) add a semantic annotation level to their corpus based on the UMLS and EUROWORDNET.

For our annotation purposes, we took STTS (Thielen and Schiller, 1996), a standard general-language tagset for German comprised of 54 tags, which is described in Section 3.1. Medical language, as used in clinical documents, has some unique properties though. This is reflected in a medical language extension of the STTS, which is described in Section 3.2.

The annotation itself was carried out by three student annotators under the supervision of the first author. The inter-annotator consistency (mean: 98.4%, standard deviation: 0.6) of our annotation group was slightly below the numbers reported for the NEGRA group (98.6%), which consisted of two human annotators (Brants, 2000).

#### 3.1. The standard part-of-speech tagset

STTS is a general-purpose tagset developed for the POS annotation of German newspaper and newswire text corpora, such as the NEGRA newspaper corpus (see Table 3 for a selection of some of the tags being used).

POS tag	Definition	German Examples ( <i>English</i> )
ART	article	eine ( <i>a</i> ), der/die ( <i>the</i> )
ADJD	adverbial modifier	zunehmend ( <i>increasingly</i> )
ADJA	prenominal adjective	fiebrige ( <i>febril</i> )
NN	common noun	Krankheit ( <i>disease</i> )
NE	proper noun	Aspirin, Pfizer
VVINF	verbal infinitive	untersuchen ( <i>to examine</i> )
APPR	preposition	in, auf ( <i>on</i> ), mit ( <i>with</i> )
PPER	personal pronoun	es ( <i>it</i> ), er ( <i>he</i> ), ihn ( <i>him</i> )
KON	coordination	und ( <i>and</i> ), oder ( <i>or</i> )
KOUS	subordinating conjunction	weil ( <i>because</i> )
PTKZU	infinitive marker	zu ( <i>to</i> )
FM	non-German word	English, French words
XY	non-words	H <sub>2</sub> O, P02.7, Q61.3
\$.	sentence-final marker	.!?
,\$	comma	,
\$(	sent.-internal marker	'-''()[]

Table 3: Fragment of STTS Part-of-Speech Tags

A look at the ranking based on the frequency of occurrence of some of the standard STTS POS tags shows which tag types are more or less prominent across the different medical genres and in comparison with the NEGRA newspaper material (see the rows 2 to 8 in Table 4).

Clearly, the common noun (NN) is the most widely used part of speech in all text genres, both medical and non-medical. In three out of four clinical genres (pathology

Standard STTS POS tag	ranking of POS tags for different text genres							
	pathology	histology	surgery	discharge	textbook	consumer	FRAMED total	newspaper
NN	1	1	1	1	1	1	1	1
ADJA	2	2	3	2	3	4	3	4
XY	18	16	28	17	41	43	26	34
ADJD	6	5	5	6	6	9	5	13
NE	12	19	15	7	13	21	13	5
KOUS	31	33	20	28	17	16	22	17
PTKZU	38	37		39	20	25	29	24
Extended STTS-MED	pathology	histology	surgery	discharge	textbook	consumer	FRAMED total	newspaper
LATIN	14	20	13	22	29	42	19	
ENUM	15	10	14	26	33	15	14	
FDSREF		15			40		33	
size of POS tagset	57	57	57	57	57	57	57	54

Table 4: Ranking of some selected STTS POS tags. (Blank cells indicate that a tag is not used.)

and histology reports, discharge summaries), the prenominal adjective (ADJA) occupies the second rank, whereas it is on rank four in the newspaper genre and, thus, resembles more consumer health documents. An introspection into some of the clinical texts illustrates its prominence:

- Peritonealkarzinose mit multiplen/ADJA, disseminierten/ADJA, perikolischen/ADJA Tumorinfiltrationen des Dünndarmes  
(*peritoneal carcinosis with multiple, disseminated, pericolonic tumor infiltrations of the small intestine*)

The ranking of the adverbial modifier (ADJD) also marks a difference between medical and newspaper texts. Whereas it is on ranks 5 or 6 in clinical and textbook texts, it only occupies rank 13 in the NEGRA corpus (again, similar to consumer health texts). Its prominence can be explained by its relationship to ADJA, which is illustrated below:

- minimal/ADJD verbreiterten/ADJA Septen  
(*minimally broadened septums*)
- vermehrt/ADJD schaumzellig/ADJD transformierte/ADJA Makrophagen  
(*increasingly foam-cell-like transformed macrophages*)

As can be seen, these adverbial modifiers modify prenominal adjectives. This is a linguistic phenomenon which mostly occurs in the clinical document types, which in turn explains their prominent ranking.

The ranking of the POS tag for non-words (XY) indicates their prominence particularly for the clinical text types. Whereas it only occupies rank 34 in the newspaper genre, it is much higher ranked in discharge summaries (17), pathology (18) and histology reports (16). This contrasts sharply with the ranking this tag holds for textbook (41) and consumer health texts (43). The clinical text types contain various disease and health-related classification keys, such as the following classification codes:

- T3C3 N1C3 M1OSSC2 (= TNM classification code for malignant tumors, in this case *prostate carcinoma*)
- K22.3 (= ICD-10 International Classification of Diseases, in this case *perforation of oesophagus*)

As can be seen, these codes are usually made of alphanumeric characters, and hence have to be tagged as non-words according to the STTS tagging guidelines (Thielen and Schiller, 1996).

The ranking of the POS tag for proper names (NE) actually shows that there is a clinical text genre, viz. discharge summaries (rank 7), which is, in this respect, quite similar to the newspaper text genre (rank 5), whereas it is ranked much lower in all the other medical genres. This can be attributed to the fact that, both in discharge summaries and in newspaper articles, proper names (e.g., of persons, pharmaceutical products, places) are mentioned quite often. Whereas histology and pathology reports deal with samples of tissues to be examined (and hence there is no need to mention the patient's name in the first place), it is not necessary in to include patients' names in surgery reports since this kind of information is available as (HL-7-style) metadata in most hospital information systems. Discharge summaries, on the other hand, are a more prosaic and personalized form of medical narrative which describe the patient's case and medical history and thus usually include various person names (of the patient, acting physicians, etc.). As a matter of fact, discharge summaries are usually embedded in a letter format, which is also evidenced by the German term 'Arztbrief', in English ('*doctor's letter*'). Moreover, they quite often contain a list of the patient's prescription drugs, another source of proper names.

Two other standard STTS POS tags, the one for subordinate conjunctions (KOUS) and for the German infinitival marker 'zu' (PTKZU) in subordinate clauses, are more of a syntactic nature, i.e. their ranking indicates how much syntactic *subordination* one may observe in a corpus. The rankings for both newspaper texts and for the non-clinical text genres (textbook and consumer material) are quite similar, viz. between 17 and 25. For the clinical genres, however, the picture looks different: KOUS is ranked up to 16 positions lower (in histology reports) and PTKZU up to 15 positions lower (in discharge summaries) compared with newspapers. In surgery reports, PTKZU does not even occur. This clearly indicates that clinical reports in general are characterized by a paratactic sentence style, whereas non-clinical ones and also newspaper texts exhibit a more hypotactic style.

### 3.2. Extending the standard tagset

An introspection into the variety of clinical and non-clinical texts indicates that medical language has some unique properties. Among them are the use of Latin and Greek terminology (sometimes also mixed with the host language, here German), various *ad hoc* forms for abbreviations and acronyms, a variety of (sometimes idiosyncratically used) measure units, enumerations, and some others.

The question arises whether these are characteristic or just marginal sublanguage properties. Thus, for our tagging purposes, we enhanced the standard STTS tagset with three novel tags (see Table 5) which are intended to capture some of the ubiquitous properties in medical texts not covered by a general-purpose tagset: the extended STTS-MED tagset.

Tag	Description	Examples
LATIN	Latin nominatives or genitives in medical terms	<i>Arteria pulmonalis dextra</i> <i>Ulcus ventriculi</i>
ENUM	Enumerations	<i>1., 2., a., (b), i., ii.</i>
FDSREF	Reference patterns w.r.t. formal document structure	as described under 2. as mentioned in <i>1.a.</i>

Table 5: Medical Extension of the STTS Tagset

A look at the frequency ranking of these medical tags underlines their relevance for annotating medical corpora (see rows 10 to 12 in Table 4) and shows that they are not just marginal sublanguage properties. Among the 57 tags which make up the entire STTS-MED tagset, the first tag, LATIN, is ranked 19 in FRAMED. It refers to Latin nominatives and genitives, which occur almost exclusively in anatomical and pathoanatomical terms (see the first and second example for LATIN in Table 5, respectively). This explains why it is ranked highest in pathology (which deal with macroscopic anatomical structures) and surgery reports, whereas it is less prominent in histology reports (which deal with microscopic anatomical structure on the cellular layer) and discharge summaries.

The other two new medical tags, ENUM and FDSREF, refer to the formal structure of a particular document. As can be seen, the tag for enumerations is the most prominently ranked (rank 14 in FRAMED) of all the newly introduced ones. As ordinal numbers (e.g., 1., 2., 3., etc.) and letters (a., b., 2.c., etc.) they are particularly widespread in three clinical text genres (histology, pathology, and surgery reports). Their rather prominent presence in health care consumer texts is in form of bullet points, which is a characteristic feature of texts taken from the Web. FDSREF, the tag which basically references ENUM, is less prominent and only has some impact in histology reports.

Finally, it should be noted that we cover the other unique medical sublanguage properties mentioned above (abbreviations, acronyms, measure units) by exploiting the original STTS tagset. This is in accordance with basic design principles which require to keep a tagset as small as possible because each additional tag introduces new potential decision ambiguities, which have to be resolved by a POS tagger. We justify the new tags for formal document structure, ENUM and REF, by the fact that these properties were only inadequately covered by the standard STTS tagset, viz. by the tag for adverbs (ADV) and car-

dinal numbers (CARD), which is, obviously, wrong. For the Latin nominatives and genitives, a standard tag for foreign language material (FM) would have been available. Anatomical and pathoanatomical terms, however, are quite ubiquitous in some of the clinical text genres so that their subsumption under “foreign language material” in general would not adequately reflect their real linguistic status.<sup>4</sup>

## 4. Conclusion

We described the structure and the quantitative properties of FRAMED, an annotated German-language corpus which contains a variety of medical text genres. On a broader scale, such specialized language resources seem to be a valuable asset for re-training and newly developing robust and effective domain-targeted medical language technologies. Among these are vital applications such as routine information extraction from clinical documents (Hahn et al., 2002), clinical document routing or medical question answering services.

**Acknowledgements.** This work was supported by Deutsche Forschungsgemeinschaft (DFG), grant KL 640/5-1, and by the Faculty of Medicine at Freiburg University, grant KLA231/03.

## 5. References

- Brants, T., 2000. Inter-annotator agreement for a German newspaper corpus. In *LREC 2000 – Proceedings 2nd Intl. Conference on Language Resources and Evaluation*.
- Brants, T., W. Skut, and H. Uszkoreit, 2003. Syntactic annotation of a German newspaper corpus. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*. Kluwer, pages 73–87.
- Hahn, U., N. Bröker, and Peter Neuhaus, 2000. Let’s PARSETALK: Message-passing protocols for object-oriented parsing. In H. Bunt and A. Nijholt (eds.), *Advances in Probabilistic and Other Parsing Technologies*. Kluwer, pages 177–201.
- Hahn, U., M. Romacker, and S. Schulz, 2002. MEDSYNDIKATE: A natural language system for the extraction of medical information from finding reports. *International Journal of Medical Informatics*, 67(1/3):63–74.
- Hahn, U. and S. Schulz, 2002. Towards very large ontologies for medical language processing. In *LREC 2002 – Proceedings 3rd International Conference on Language Resources and Evaluation*. pages 2137–2144.
- MSD, 1993. – *Manual der Diagnostik und Therapie [CD-ROM]*. Urban & Schwarzenberg, 5th edition.
- Thielen, C. and A. Schiller, 1996. Ein kleines und erweitertes Tagset fürs Deutsche. In H. Feldweg and E. Hinrichs (eds.), *Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*. Niemeyer Verlag, pages 193–204.
- Vintar, Š., P. Buitelaar, B. Ripplinger, B. Sacaleanu, D. Raileanu, and D. Prescher, 2002. Towards very large ontologies for medical language processing. In *LREC 2002 – Proceedings 3rd International Conference on Language Resources and Evaluation*. pages 2137–2144.

<sup>4</sup>In Latin, e.g., unlike in German or English, attributive adjectives follow their nouns (see the first example in Table 5)