Optimizing and Translating LSP texts

1. Introduction
LSP texts comprise texts that are as different as a textbook chapter written for a first semester medical student or the latest journal article written by a frontline-researcher. Yet the main purpose of an LSP text is always to convey factual information (Gläser 1995:153). Therefore, when it comes to writing or translating an LSP text, its readability should be of highest priority:


In order to write a readable text, one has to know what characteristics influence readability. For text translation, it is not only important to know what makes a text easy to understand, but also to take the different conventions and styles of the languages into account (Göpferich 1993; 1995; 1998; Nord 1998). Therefore, the translator has to aim at high readability but also meet the stylistic requirements of the target language. These two criteria may or may not be compatible. The better the translator’s knowledge about criteria that influence readability and their traditional distribution in source and target language LSP texts, the more he or she will be able to make the right translation decisions.
2. Readability of English and German LSP texts

English LSP texts are usually perceived to be much easier to comprehend than German LSP texts. Significant differences have been described at the macrostructural level where English texts are considered to be more reader-oriented while German texts are more content-oriented (Kaplan 1966; Clyne 1984; 1987; 1996; House 1996; 1997; Gerzymisch-Arbogast 1993; 1997; Buhl 1999). Yet no differences are described on the microstructural/lexical level. However, it is a widely felt impression that the lexis of German LSP texts is particularly difficult and complicated. The present work addresses lexical parameters that might influence the comprehensibility of English and German LSP.

The lexical inventory of an LSP text consists of the subject-specific terminology and of ‘other’ words.

\[
\text{LSP-text} = \text{subject-specific terminology} + \text{remaining text (RT)}
\]

In cooperation with Jan Alexandersson and Paul Buitelaar from the German Research Center for Artificial Intelligence (DFKI) I analyzed written English and German medical texts, such as textbook texts and journal articles with regard to their proportion of medical terminology. It has been described that in these texts, the medical terminology makes up 20-25% of the texts in both languages (Beier 1980:40; Sieper 1980:3; also confirmed by own data, see below). Thus, the medical terminology cannot account for the difference in readability.

So far, the lexis in the LSP context has usually been reduced to subject-specific terminology, strongly neglecting the remaining 75-80% of the texts (cf. Roelke 1999:50). It is, therefore, conceivable that precisely this part of LSP texts causes the subjective differences in comprehensibility. To address this possibility, we focused our study on the ‘remaining text’.

The remaining text consists of basic terms (BT) and non-basic-terms (non-BT).

\[
\text{RT} = \text{basic terms} + \text{non-basic terms}
\]

It is conceivable that basic terms, which comprise the most frequent and the most important terms (Langenscheidt 2000:VII), are easier to understand than non-basic terms.
3. **Hypothesis**
While it is certainly true for all LSP texts that they consist of subject-specific terminology, basic and non-basic terms, and the proportion of terminology does not differ in the languages under investigation, nothing is known about the distribution of basic- and non-basic terms in different languages.

It can therefore be hypothesized that English LSP texts have a higher proportion of basic terms in their RT than German LSP texts, and that this difference accounts for the subjective impression of higher comprehensibility of the English texts.

4. **Corpus analysis**
What is common to many of the studies in the field of LSP research is that they usually remain at the qualitative level (Vihla 1998:74). In our study, we aimed at combining qualitative and quantitative aspects by testing our hypothesis on a large computer corpus. We wanted to show whether or not there is a difference in the relative number of basic terms in the RT of German and English medical LSP texts. Our study was conducted in three steps:

- Manual analysis of a small control corpus
- automated analysis of the same control corpus & comparison of results
- automated analysis of a large corpus (book corpus and paper corpus)

4.1. **Corpus**
When designing the corpus, we chose the following criteria for the texts we wanted to include:

- Size: we aimed at several hundred thousand words
- Extracts from books were accepted
- Research articles were to be included in full
- Subject: texts from textbooks for medical students and medical journals
- Texts had to be written by native speakers
- Publication date: 1990 and later
The small control corpus for the manual and automated control analysis contains 7 English texts and 8 German texts randomly selected from different textbooks for medical students, each approximately 400 words long. The large book corpus for the automated analysis contains excerpts from 15 different authors for each language. The excerpts are between 240 and 46,606 words long and were randomly chosen by the publishing houses that provided them.¹ The paper corpus for automated analysis contains full text articles of five and 13 different native authors, respectively.

<table>
<thead>
<tr>
<th>Size [words]</th>
<th>Control corpus</th>
<th>Book corpus</th>
<th>Paper corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng</td>
<td>2.479</td>
<td>123.149</td>
<td>52.617</td>
</tr>
<tr>
<td>Ger</td>
<td>2.512</td>
<td>108.622</td>
<td>8.632</td>
</tr>
<tr>
<td>Total</td>
<td>4,991</td>
<td>231.771</td>
<td>61,249</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of texts</th>
<th>Control corpus</th>
<th>Book corpus</th>
<th>Paper corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eng</td>
<td>7</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Ger</td>
<td>8</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>30</td>
<td>18</td>
</tr>
</tbody>
</table>

Tab. 4.1: Corpora. For further information on corpus design and compilation see Bowker & Pearson (2002: 45ff.)

4.2. Analysis

A manual and an automated analysis was conducted: The manual analysis was performed to initially test our hypothesis on the small control corpus. This control corpus was subsequently analyzed by the FURORE Software developed by Jan Alexandersson especially for this purpose, and the performance of the program was evaluated. Following this, the same automated analysis was performed evaluating the large (book and paper) corpus.

Manual analysis

Texts were analyzed with respect to their total number of words, average sentence length and number and proportion of medical terms

¹ German texts were provided by the Springer Verlag, Abteilung Medizinische Lehrbücher; English texts were provided by Garland Science Publishing.
(medical terms were identified by specialists). In the ‘remaining text’, basic terms\(^2\) were counted and their proportion was calculated.

**Automated analysis**

First, the documents were annotated with multiple layers of linguistic information, including part-of-speech tagging, morphological analysis, and the identification of medical terms using the MeSH (Medical Subject Heading) subset of the Metathesaurus of the Unified Medical Language System UMLS\(^3\). Annotations were performed by Paul Buitelaar. The FURORE software determined the total number of words, average sentence length and number and proportion of medical terms. Medical terms were subsequently deleted to prevent an overlap in word counts. In the remaining text, basic terms were counted and their proportion was calculated:

\(^2\) We combined the information from four dictionaries of basic terminology to create a list of 5000 basic terms for each language.

\(^3\) For linguistic processing we used ShProT (Buitelaar 2003) a shallow processing tool that consists of four integrated components of which we used the first three: the SPPC tokenizer, TnT for part-of-speech tagging, and Mmorph for morphological analysis. ShProT is used in the scope of the MUCHMORE project at the DFKI.
Loading basic voc ' $/GWSe.txt'...done
reading file: 'ceb3.xml'...done
Processing......................done


Number of sentences: 19
Number of words: 379
Average sentence length: 19
Number of concepts: 54 (14.248021%)
Number of basic terms: 229 (60.422165%)
Number of rest terms: 96 (25.329817%)

Fig. 4.2: Example of a result file produced by the FURORE software: File 'ceb3.xml' is read and processed by the program, basic vocabulary such as ‘for’, ‘example’, ‘be’, ‘very’, is followed by medical terms such as ‘spatial’, ‘discrimination’, ‘retina’, ‘sensory’, ‘receptor’, followed by non-basic words such as ‘intensity’, ‘determine’, ‘amplitude’, or ‘historic’ (sample words marked in bold). At the end of the file, the program calculates the data.
5. Results

5.1. Manual analysis

The manual analysis of the small control corpus confirmed the hypothesis:

![Manual analysis of control corpus](image)

The proportion of medical terminology was as described in the literature between 20 and 25% with no difference between the two languages. The sentence length (not shown), too, did not display any differences. However, we found a significant difference in the proportion of non-BT with 31% in German as compared to 14% in English texts, supporting our hypothesis that understandability of the texts correlates with the portion of basic terms.

5.2. Manual vs. automated analysis

In order to evaluate the correctness of the data obtained with FURORE, we next compared the manual analysis with the automated analysis performed by FURORE on the same corpus.
The automated analysis reveals a similar proportion of medical terms in English and German texts with a statistically not significant tendency to fewer medical terms in English (10% vs. 17% respectively). The overall percentages, however, were lower in the automated than in the manual approach.

The analysis of non-BT revealed higher proportions in the automated approach (but similar tendencies in both approaches) with 43% in German vs. 31% in English texts. Further analysis reveals that, in the automated countings, the higher numbers of non-BT and the lower number of medical terms are a result of misgrouping of a small portion of the medical terms into the non-BT group. However, this misgrouping did not affect the overall tendencies among the groups. Finally, the sentence lengths display high agreement in both approaches (not shown). Thus, FURORE appears to be a reliable tool for performing an automated analysis on large corpora.
5.3. Automated analysis: textbook vs. paper

The analysis of textbook texts and paper texts lead to the same results as the above manual analysis:

![Automated analysis book vs. paper](image)

Both groups display slightly more medical terms in the English texts and much more non-BT were found in the German texts. This is an important finding since it shows there is no connection between the use of basic terminology and information density. While authors are very free in their choice of form and style when writing books, the structure and length of a journal article is usually strictly regulated. As shown in Figure 5.3, book and journal texts are, however, almost identical with regard to the characteristic proportion of basic terminology in both languages.

6. Discussion

The present study reveals that the proportion of basic terms is higher in English than in German medical texts. This supports our hypothesis that there are language-specific differences in the use of basic and non-ba-
sic terms, and that the comprehensibility of medical texts increases with the proportion of frequently used, standard-language words. A special software tool, FURORE, was written to evaluate a large medical corpus, and we could show that this leads to similar results to a manual analysis, confirming its validity. This offers the possibility to subject large medical text corpora to automated lexical analysis that might reveal differences between English and German texts that have not been detected so far. We believe that these analyses will provide the basis for simple approaches on a lexical level to improve the comprehensibility of specialized texts.

Differences between the manual and automated approaches
The analysis of medical terms reveals differences between the manual and automated approaches. This was expected since the Medical Terminology (MT)-reference list for automated analysis, as any lexical reference list, is incomplete, and it is even less comprehensive for German MT. Thus, some medical terms are not recognized by the program, whereas in a manual analysis, medical language specialists can identify all medical terminology. This explains the lower values for medical terms in the automated vs. manual analysis, but also the higher absolute values for the non-BT groups: Due to the evaluation scheme, unrecognized medical terminology accumulates in this group. This effect is of similar extent for both languages and thus does not significantly influence the results.

Influence of the BT-proportion on text comprehensibility
In the German texts, more than a third of the non-MT are not basic terms, i.e. complicated and unusual expressions that are more difficult to recognize and to understand, but only 14% of the English non-MT belongs to this group. This may contribute to the subjective impression that English LSP texts are easier to understand than German texts.

A multidisciplinary point of view
Whether a word is considered a ‘basic term’ is determined by two criteria: (1) The frequency with which it occurs in spoken and written language, and (2) how basic/fundamental it is (Langenscheidt 2000:VII). The latter corresponds with the age at which a word is learned. This so-called “Age of Acquisition” and the “Word Frequency” are the most wide-
ly investigated variables in word recognition, processing and memory research in the field of cognitive science (Ghyselinck & Lewis & Brysbaert 2004, 43ff.). Studies in this field have revealed, for example, that frequently encountered words are processed more easily than less common words (Grainger 1990; Jescheniak & Levelt 1994). Semantic decisions about high-frequency words can be made faster (Chee & Westphal & Goh 2003) and with greater ease (Chee & Hon & Caplan 2002) than decisions related to low-frequency words. This so called “Frequency effect” can be clearly distinguished from the “Age-of-Aquisition (AoA) effect” (Fiebach & Friederici & Müller 2003; Ghyselinck & Lewis & Brysbaert 2004, 43 ff.) that states that early learned words are processed faster than those words learned at a later stage (Ghyselinck & Lewis & Brysbaert 2004). The AoA proved to be a significant variable for a wide range of word processing paradigms such as lexical decision (Brysbaert & Lange & van Wijnendaele 2000; Gerhand & Barry 1999b; Morrison & Ellis 1995; 2000), semantic categorization (Brysbaert & Lange & van Wijnendaele 2000), picture naming (Barry & Morrison & Ellis 1997; Bonin & Chalard & Meot 2002), speeded word naming (Gerhand & Barry 1999a), and auditory lexical decision (Turner & Valentine & Ellis 1998).

With these two factors, word frequency and AoA, having a large impact on the ease of several levels of word processing, it can be predicted that basic terms will most likely show similar effects, and that the relative amount of basic terms in a text strongly influences its readability and comprehensibility.

From the field of translation theory, it has been argued that the attention of the reader is distracted from the text content if text type conventions are not met properly, and therefore this should be avoided in LSP translation (Reiß & Vermeer 1984:189; Biere 1989; Göpferich 1998:62f.). It would be important to verify whether this has a negative effect on text comprehensibility and would therefore counteract the above mentioned beneficial effects of AoA and word frequency.

Concluding remarks
Increasing the comprehensibility of specialized text based on lexical parameters might prove to be useful for optimizing newly written medical texts as well as for their translation. The main purpose of a specialized text is to impart information. Thus, optimization of this process should
be a major goal, and increasing the comprehensibility for the audience (target group) is surely one important aspect in this context. In contrast, the implementation of a scientific style of writing as a convention for the generation and translation of specialized texts does, at least on the German side, not necessarily act in synergy with the specification of optimal information transfer. Also, studies in the field of cognitive science show that the ease of processing positively correlates with the simplicity of texts. We therefore propose that the criterion of comprehensibility should be given priority over scientific style conventions.

To this end, our future research will focus on optimizing the corpus analysis tools, and collecting further evidence to show that specific lexical criteria such as the proportion of BT-words are causal factors for the comprehensibility of a specialized text. In particular, we will expand the corpora to be analyzed and establish comprehensive English and German biomedical text corpora, improve the term recognition, investigate the AoA and frequency-proportions seperately, and extend the analysis to other levels such as syntax and concepts. Parameters revealed by these studies will then be used to test our hypothesis on human subjects. These studies may involve both behavioral memory tests and functional brain imaging. In this context, we will also evaluate whether trespassing classical conventions in the style of scientific writing interferes with comprehensibility.

7. References
7.1. General references


Bonin, Patrick/Chalard, Marylène/Méot, Alain/Fayol, Michel 2002: The determinants of spoken and written picture naming latencies. In *British Journal of Psychology* 93, 89-114.


Chee, Michael W. L./Westphal, Christopher/Goh, Joshua/Graham, Steven/Song, Allan W. 2003: Word frequency and subsequent memory effects studied using event-related fMRI. In *NeuroImage* 20, 1042-1051.


Gläser, Rosemarie 1995: *Linguistic features and genre profiles of scientific English.* Frankfurt am Main u.a.: Lang.


Grainger, Jonathan 1990: Word frequency and the neighbourhood frequency effects in lexical decision and word naming. In *Journal of Memory & Language* 29, 228-244.


7.2. Dictionaries


Langenscheidt 2000: *Basic German Vocabulary*. Berlin u.a.: Langenscheidt.


Add a
dimension to
your
sociology research...

sociological abstracts

Comprehensive, cost-effective, timely coverage of current ideas in sociological research

Abstracts of articles, books, and conference papers from nearly 2,000 journals published in 35 countries; citations of relevant dissertations as well as books and other media.

Available in print or electronically through CSA Illumina (www.csa.com).

Contact sales@csa.com for trial Internet access or a sample issue.

Now featuring:
• Cited references
• Additional abstracts covering 1963-1972

www.csa.com