Quality indicators of LSP texts – selection and measurements

Measuring the terminological usefulness of documents for an LSP corpus

Jakob Halskov[1], Dorte Haltrup Hansen[2], Anna Braasch[3], Sussi Olsen[4]
[1]The Danish Language Council, [2,3,4]University of Copenhagen, Centre for Language Technology
Njalsgade 140, DK-2300 Copenhagen S
E-mail: [1]jhalskov@dsn.dk, [2]dorteh@hum.ku.dk, [3]braasch@hum.ku.dk, [4]saolsen@hum.ku.dk

Abstract

This paper describes and evaluates a prototype quality assurance system for LSP corpora. The system will be employed in compiling a corpus of 11 M tokens for various linguistic and terminological purposes. The system utilizes a number of linguistic features as quality indicators. These represent two dimensions of quality, namely readability/formality (e.g. word length and passive constructions) and density of specialized knowledge (e.g. out-of-vocabulary items). Threshold values for each indicator are induced from a reference corpus of general (fiction, magazines and newspapers) and specialized language (the domains of Health/Medicine and Environment/Climate). In order to test the efficiency of the indicators, a number of terminologically relevant, irrelevant and possibly relevant texts are manually selected from target web sites as candidate texts. By applying the indicators to these candidate texts, the system is able to filter out non-LSP and "poor" LSP texts with a precision of 100% and a recall of 55%. Thus, the experiment described in this paper constitutes fundamental work towards a formulation of 'best practice' for implementing quality assurance when selecting appropriate texts for an LSP corpus. The domain independence of the quality indicators still remains to be thoroughly tested on more than just two domains.

1. Introduction

The work presented in this paper is carried out within the framework of the Danish CLARIN project (DK-CLARIN; http://english.dkclarin.ku.dk/) the aim of which is to create a research infrastructure for the humanities in Denmark, focusing on written and spoken language resources and multimodal resources including their integration into a web-based environment. The DK-CLARIN infrastructure will comprise a variety of corpora of old and modern Danish, parallel texts, spoken language, dictionaries, video sequences etc.

The project runs from 2008 until the end of 2010 and is funded by The Danish Agency for Science Technology and Innovation. The project participants come from several Danish universities as well as from other cultural institutions including the National Library and the Danish Language Council.

2. The DK-CLARIN LSP corpus

In one of the work packages, The Danish Language Council and Centre for Language Technology, University of Copenhagen cooperate on compilation of a corpus of LSP (Language for Special Purposes) texts from broadly selected domains. The DK-CLARIN LSP corpus will comprise 11 M tokens from the period 2000-2010, complementing the general language corpora of DK-CLARIN.

After a phase of specification of the project's common metadata set, the LSP corpus group is now in the process of collecting texts from different suppliers, developing the methods and tools for selecting, processing and storing the texts.

The corpus specifications state that the *domains* covered by the corpus should be of widespread public interest, i.e. domains where much terminological work is carried out and where terms are likely to become part of the general language over time. Though, the domains selected should not be too close to general language, which excludes domains like sports and culture. Domains the texts of which contain many formulae, etc., such as mathematics and chemistry, are also disregarded because of the difficulties in treating these appropriately in the process. Accordingly, the following domains have been selected:

- Health and Medicine
- Environment and Climate
- Information Technology (IT)
- Nanotechnology
- Construction
- Agriculture

The corpus should be composed of texts in the *communicative settings* expert/(semi)-expert to (semi)expert/layman since these are expected to have the appropriate degree of specialization (Pearson 1998: 35ff). Furthermore, the corpus should comprise *text types* with different communicative aims: informative (reports, product descriptions), normative (standards, regulations), instructive (textbooks, manuals), etc.

The DK-CLARIN LSP corpus is meant for various linguistic and terminological research purposes; this requires a thorough, *quality-based selection* of texts for the corpus. To this end, we have defined a set of quality indicators, for example estimated term richness, explicit knowledge density, use of passive voice and use of English, all concerning text internal properties and relevant to including a text into the corpus or leaving it out, as discussed in the section on Dimensions of quality. Other, text external, quality indicators, such as text source reliability, time and modus of publication, relevance of the topic, original or translation, etc., will not be discussed further here.

3. Corpus compilation

Due to copyright issues, web-sites are manually selected, and there is thus no need to automatically classify the documents they contain as belonging to a particular domain (this information is already present as a selection criterion for the source) or to employ reranking mechanisms to the search results returned by commercial search engines (as is done in Agbago & Barrière, 2005). On the other hand, it is highly important to continually evaluate the terminological quality (or usefulness) of the corpus which is being compiled.

A large amount of documents for the corpus are harvested automatically from web-sites for relevant institutions, government departments, etc. Especially, the homepages of government departments are regarded as reliable web-sites with easily accessible and freely available material. Furthermore, these texts are written by field experts for (semi-)experts and laymen, thus their communicative settings are in accordance with our selection principles. Additionally, the texts are very often also checked by a language expert or editor in order to ensure proper linguistic quality. Nevertheless, some of the texts harvested are unfit for our purposes and must be filtered out. The unwanted texts mostly deal with web-navigation, administrative information and some legal and political issues. These kinds of texts are usually rather short and of very low term density, in other cases the terminology of the texts just belongs to other domains than the one in question.

The experiments described in this paper constitute fundamental work towards a formulation of 'best practices' for implementing quality indicators for selecting appropriate texts for our corpus.

4. Methodological framework

In the NLP field known as automatic text classification the typical approach is to employ a training set of objects, each labelled with one or more classes which are encoded via a data representation model (Manning and Schütze, 1999: 575-576). While the data representation model is typically the vector space model known from information retrieval, approaches differ in their choice of classification model and classifier training procedure.

In traditional corpus linguistics, a prevalent approach to text classification is the multidimensional analysis of register variation. A range of linguistic criteria can be used to categorize a document, and a number of the features in focus are especially important when examining the appropriateness of a document for particular types of corpora.

In computational terminology, candidate texts are often classified by assessing their degree of specialization as well as their density of *explicit knowledge*. This has been done by estimating the density of so-called knowledge-rich contexts via an automatic identification of selected "knowledge patterns" in the texts (Meyer, 2001). In fact, this approach has been combined with techniques from the so-called Web as Corpus (WaC) field (Kilgarriff & Grefenstette 2003) in applications like TerminoWeb (Barrière and Agbago, 2006).

Our approach combines methods from the above three fields. In particular the following three goals are defined for the LSP corpus. It should

- be rich in specialized terminology
- have a high density of explicit knowledge
- reflect a versatility of the domain(s) in terms of topics and text types

On the one hand, specialized terms represent the domain specific concepts which act as nodes in domain specific ontologies; on the other hand, knowledge-rich contexts are needed to provide information on the conceptual relations which link these nodes so that they may form a coherent network. The ideal corpus is thus not only rich in specialized terminology but also rich in explicit knowledge.

5. The experiment

The experiment is based on three elements: reference data, dimensions of quality and test documents. The reference data consists of an LGP corpus (i.e Language for General Purposes) and an LSP corpus, and the dimensions of quality are represented by a number of concrete linguistic features thought to be characteristic of the sort of texts we wish to include in the DK-CLARIN LSP corpus. Both the reference data and the quality indicators will be described below.

As a first step, average values for the quality indicators are induced from the reference data. This should elucidate the usefulness of each feature as an indicator of terminological quality and at the same time provide us with threshold values for each indicator.

The second step in the experiment is to select a number of texts known to be relevant, irrelevant or potentially relevant, and to compute values for all quality indicators across all texts. The results are evaluated by means of precision and recall.

6. Reference corpora

The reference corpus for LGP is the DK87-90 corpus (Bergenholtz, 1988) which is a corpus of 4 M words, comprising a balanced selection of books, magazines and newspaper articles.

The reference corpora for LSP employed in these experiments come from two domains, health/medicine and environment/climate.

For both domains, the texts have been collected manually from selected web-sites with clearly declared informative and/or instructive aims. Most of the health/medicine texts are collected from the National Board of Health and the Capital Region of Denmark. The texts from the environment/climate domain have been collected from the National Environmental Research Institute, the Ecological Council and a popular science journal.

The texts are partly web-texts written in an expert to layman communicative setting, partly technical reports or surveys written in an expert to (semi)-expert communicative setting. It has been endeavoured to cover the different communicative settings for the two domains but it has proved difficult to distinguish between expert to expert and expert to semi-expert communication when it comes to real texts, where neither the author (except for the name) nor the expected readers are known. That is why we have chosen to work with a communicative setting called expert to expert.

We have ensured that the amount of LSP texts is comparable for the two domains in that we have tried to have comparable amounts of expert-expert texts and expert-laymen texts.

Corpus type			Tokens	Words> 8 letters (%)	Compounds (%)	Average sent. length	Passive verbs (%)	Personal pronouns (%)	Top 1000 A, N, V tokens (%)	Top 1000 all word tokens (%)	KPs per 1000 tokens	OOV types per 1000 tokens	English OOV types per 1000 tokens
Non-LSP	Bergenholtz LGP	Fiction	2,385,912	8.04	0.95	13.63	0.10	9.30	21.85	53.61	0.20	3.81	0.09
		Magazine	1,225,267	10.29	1.30	14.94	0.20	7.20	22.88	52.29	0.40	4.65	0.07
		Newspaper	1,183,849	16.98	2.50	18.45	0.30	2.80	21.69	48.69	0.50	6.61	0.08
LSP	Exp -layman	Environment	569,595	21.86		19.52	0.51	0.50	16.39	45.71	2.60	9.52	0.43
		Health	496,322	22.13		17.48	0.74	0.30	14.79	47.35	2.90	9.63	0.46
	Exp -exp	Environment	1,005,796	24.06	5.50	20.39	0.48	0.10	12.36	38.08	1.80	13.03	1.72
		Health	1,235,515	31.89	6.19	18.29	0.71	0.10	14.23	38.57	2.40	15.07	2.17
Tentative threshold values for target texts			>500	>20		>17	0.50-1	0-0.5	<17	<50	>1.5	>7	>0

Table 1. The reference corpora with average values for the quality indicators

Table 1 shows the average values for the quality indicators discussed in the following sections. As can be seen from the numbers in table 1, there are significant differences between the non-LSP and the LSP corpora for all the quality indicators. The values in the last row are used as thresholds in the experiment that attempts to distinguish between target and non-target texts.

7. Dimensions of quality

Multiple dimensions of quality could be considered relevant when assessing the terminological value of a candidate corpus text.

Biber (1998: 144ff.) states that “Linguists have long recognized that groups of co-occurring linguistic features are instrumental in distinguishing among registers”. He develops further that “academic prose [...] has a different set of co-occurring features: frequent nouns and attributive adjectives, nominalizations, and other specialized vocabulary items [....] passive constructions and extraposed constructions.” and concludes: “Each set of co-occurring features is called a ‘dimension’ of variation.”

In our experiment, we talk about dimensions of quality rather than dimensions of variation since our objective is a normative, not a descriptive one.

Based on a literature survey, including Biber’s observations, introspection and personal experience with terminology work, we reduced the number of potentially relevant dimensions to the following two, absolutely crucial, ones: readability/formality and density of specialized knowledge. In the graph below (Fig. 1) we have tried to visualize how various text types presumably position themselves in this two-dimensional space, and we have also added the actual linguistic features used in the experiment as indicators of each quality dimension. The most neutral one of all text types is thought to be newspaper articles which supposedly are moderately readable/formal and moderately low in specialized knowledge. As mentioned earlier in this paper, there are two types of specialized knowledge, namely that which is only implicitly present in the text (i.e. term density as measured by the OOV feature) and that which is explicitly present (i.e. semantic relations between specialized concepts instantiated by knowledge patterns).

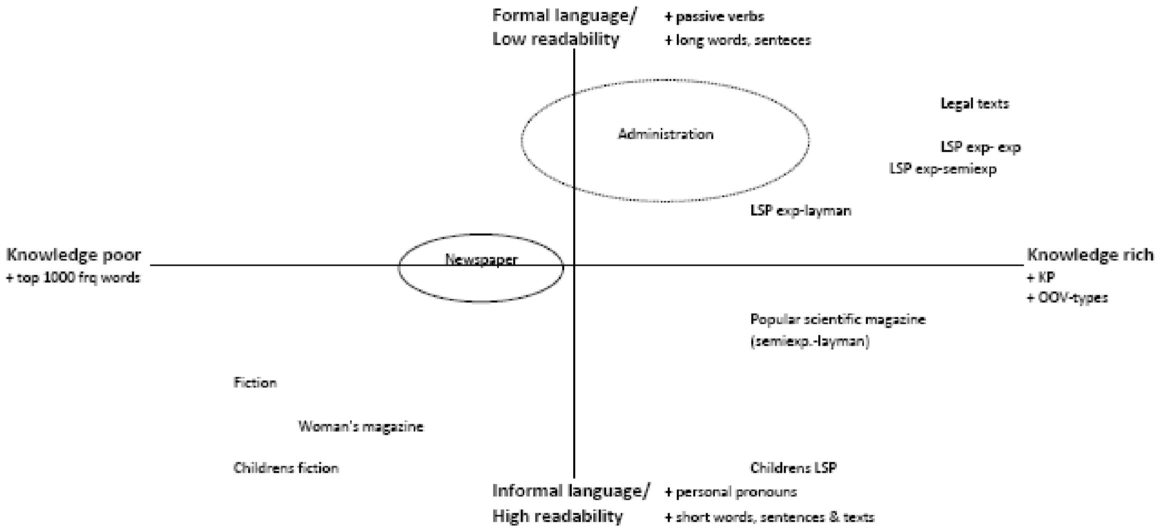

Fig. 1 Text types in relation to the two dimensions of quality

8. Methodology

The two dimensions of quality, readability/formality and density of specialized knowledge, are represented by the following linguistic features.

8.1 Readability/formality

8.1.1 Sentence length, word length and compounds

The Danish readability figure, called lix, is a measure that comprises the number of words, sentences and long words

lix = words/sentences + long words/words

The figure is usually employed in pedagogical environments to measure the degree of readability of a text for a child of a certain age or for people who suffer from dyslexia. Since our objective is quite different, the lix figure is not an appropriate indicator, but still we think that parts of the lix measure might prove useful for our purpose. We therefore split the lix measure into sentence length and the number of long words (> 8 characters).

Danish, like German, constructs compound words by joining the parts together, consequently compound words are often long words. Compounds are typically used to describe more specific concepts and are thus a prominent feature of specialised language. Obviously, there is a correlation between the number of compound words and the number of long words in a text, which is evidenced by the figures in table 1. Since identifying compounds in texts can be a computationally demanding process, we calculate word length as an approximation of compound density.

8.1.2 Use of passive voice and personal pronouns

Biber mentions passive voice and the relatively low occurrence of personal pronouns as observable features of academic texts. We employ this observation in our experiment of measuring the 'LSP'ishness' of a text.

The passive voice can be constructed in two ways in Danish: the infinitive, present and past tense adds *–(e)s* to the stem (the so-called –s passive), or it is formed with an inflected form of the auxiliary *blive* (be) and the past participle of the verb in question. The use of passive verbs signals an objective and impersonal tone with absence of agent that is characteristic of technical reports, legal and administrative documents and scientific papers. The investigation referred to in this article was concerned only with the *–s* passive since these are easily recognized by a PoS-tagger.

The use of personal pronouns in the grammatical function of a subject, on the other hand, shows a personal and subjective engagement in the matter expressed. Fiction is therefore expected to contain a large number of personal pronouns and correspondingly a low degree of passive verbs compared to non-fiction. We have counted the number of personal pronouns in the texts – leaving out the ones which are not exclusively used in the subject function.

8.1.3 Density of common vocabulary

In order to see how rich (or poor) the vocabulary of the text is, we measure the percentage of the 1000 most frequent Danish words in the text. We also measure the percentage of the 1000 most frequent content words (verbs, nouns and adjectives). High percentages of these measures signal a high readability. The lists of frequent

Danish words come from the 28M word Korpus2000, that comprises texts from 1998 – 2002 (Skovgaard Andersen et al. 2002).

8.2 Density of specialized knowledge

8.2.1 Density of out-of-vocabulary (OOV) items

OOV items are often considered a nuisance to NLP applications because they are difficult to process automatically. However, in this context OOV items are indeed exactly what we are looking for because they will often turn out to be domain-specific terms. Since the reference vocabulary is non-specialized, texts that have a high OOV ratio are very likely to also have a high density of specialized terms. Indeed, the OOV ratio can be expected to approximate the term density.

In order to calculate the number of OOV types in a document, the system employs a database table containing approximately 2 million general language word forms from the official dictionary of Danish orthography (*Retskrivningsordbogen 2001*), the largest Danish LGP corpus (*KorpusDK*) and a monitor corpus containing more recent newspaper articles from www.infomedia.dk.

8.2.2 Density of knowledge patterns

For the identification of knowledge patterns (henceforth KP), translations of English patterns empirically induced in Halskov & Barrière (2008, Terminology 14:1) were used. Only patterns representing the semantic relation *synonymy* and the conceptual relation *ISA* were included since the same study suggested that other relation types are less universal and more dependent on the domain in question. The following 16 KPs were used: *så som* (such as), *refererer til* (refers to), *også kendt som* (also known as), *også kaldet* (also called), *kaldes* (called), *er defineret som* (is defined as), *defineres som* (defined as), *fx* (e.g.), *for eksempel* (for example), *omfatter* (includes), *dvs* (i.e.), *det vil sige* (that is), *er en slags* (is a kind of), *er en type* (is a type of), *og andre* (and other), and *i form af* (in the form of). This inventory of KPs is by no means exhaustive, and it could be expanded to include also paralinguistic markers such as parentheses.

8.2.3 Use of English

In countries where English is not the official language, English is often used to invoke an air of expertise or learnedness, and scientific articles typically have an English abstract even if the text proper is written in Danish. Furthermore Danish often lacks domestic equivalents for specialized vocabulary. Consequently, (moderate) use of English in a candidate corpus text may very well be an indicator of LSP'ishness. In the experiment, English usage is assessed by looking up all OOV types in the British National Corpus (BNC).

9. Results

In this section we assess the ability of the quality indicators to distinguish terminologically useful texts from terminologically irrelevant texts. In order to do this, we compute values for each of the indicators in a total of 28 manually selected texts. Four of the texts are known to be terminologically useful, three are terminologically irrelevant newspaper articles, and the remaining 21 are potentially irrelevant documents harvested from terminologically relevant websites. The results of the computation are listed in table 2.

As we have defined two dimensions of quality, namely readability/formality and density of specialized knowledge, the elimination of non-target texts is, in fact, a two-step procedure. First, we eliminate texts that are not LSP because they either have a too high readability or use too formal language. Second, we eliminate texts that are potentially LSP but have an insufficient density of specialized knowledge as measured by the two indicators, namely OOV and KP density.

The first observation, which can be made on the basis of the figures in table 2, is that step one filtering (using the threshold values from table 1) eliminates 18 irrelevant texts, leaving six[1] potentially relevant texts and four target texts. In other words, 75% of the irrelevant texts are identified as being non-LSP in this step. The navigational pages[2], for example, are rejected at a very general level, simply because they are very short (< 500 tokens) and hardly contain any sentences (the average sentence length is lower than five words). As regards the irrelevant texts of an administrative or political nature[3], these do have a very high density of long words (often exceeding 30%) and a low density of the most common words (< 15%), but they are rejected because their density of passive constructions is too high (the threshold was set to 0,5 – 1 % on the basis of the reference corpora).

The second step eliminates six additional texts[4], leaving three target texts and one potentially relevant text. Taking a closer look at the latter text[5], we discover that this text is, in fact, perfectly useful with lots of specialized terminology and a fair level of explicit semantic relations (KP density). In other words, system precision is 100%, because the two-step filtering leaves only four texts all of which are terminologically relevant.

Unfortunately, texts 16 and 17 were eliminated already in step one because of their exceedingly high density of passive constructions (>1%). On closer inspection, these two texts are, in fact, terminologically relevant, and this suggests that the threshold value of this particular indicator should perhaps be raised somewhat to allow for a higher, more appropriate density of passive constructions. Also, a single target text (text 27) was eliminated in step two because of a low KP density. Accordingly, system recall is 55.2% (4/7).

[1] texts 9, 11, 19, 22, 23, 24

[2] texts 4-7

[3] texts 13-17

[4] texts 9, 11, 19, 23, 24, 27

[5] text 22

Content type of articles	Texts	Step 1: LGP vs. LSP							Step 2: 'Poor' LSP vs. 'good' LSP		
		Tokens	Words> 8 letters (%)	Av. sent length	Passive verbs (%)	Personal pronouns (%)	Top 1000 A, N, V tokens (%)	Top 1000 all words tokens (%)	KPs per 1000 tokens	OOV types per 1000 tokens	English OOV types per 1000 tokens
News articles	1 Art.	542	18.94	15.66	0.35	0.17	13,40	42.21	0.00	14.76	1.85
	2 Art.	628	11.09	18.03	0.15	1.06	19,42	54.01	0.00	6.37	0.00
	3 Art.	547	13.38	16.24	0.00	0.17	15,73	47.52	0.00	1.83	0.00
Navigation pages	4 Health	235	29.19	4.27	0.43	0.00	8,09	25.53	0.00	17.40	0.00
	5 Health	348	33.93	4.46	0.29	0.00	8,33	23.56	0.00	16.60	0.00
	6 Health	261	33.12	3.67	0.00	0.00	8,81	20.31	0.00	0.00	0.00
	7 Health	298	35.26	4.80	0.00	0.00	5,37	15.10	0.00	0.00	0.00
Adm. articles	8 Health	1122	24.69	17.05	1.07	1.87	18,36	54.28	0.00	3.67	0.00
	9 Health	2570	32.27	18.77	0.93	0.47	10,27	37.94	4.09	8.18	0.00
	10 Health	7397	22.61	21.36	1.01	3.60	15,14	55.12	3.19	2.63	0.00
	11 Env.	2047	31.19	17.62	0.54	0.24	13,48	45.48	3.65	11.49	0.00
	12 Env.	3838	25.76	17.52	0.91	0.49	16,46	50.29	2.74	3.84	0.00
Adm. / political articles	13 Health	4543	32.35	17.44	1.96	0.00	11,14	41.21	0.22	5.89	0.00
	14 Health	16,951	28.99	17.73	2.05	0.03	15,49	50.65	1.29	6.98	0.05
	15 Health	58,167	34.08	22.14	1.78	0.00	10,74	39.28	2.87	7.40	0.02
	16 Env.	4393	30.37	19.79	1.16	0.30	14,34	45.64	1.18	19.06	0.00
	17 Env.	3212	33.23	18.46	1.00	0.19	10,09	37.42	3.74	23.26	0.37
Borderline articles	18 Health insur.	411	28.90	30.10	0.73	1.22	10,22	30.41	0.00	0.00	0.00
	19 Health insur.	1017	22.18	18.36	0.89	0.00	13,37	44.15	3.28	5.44	0.00
	20 Health insur.	867	27.41	18.81	0.35	2.19	11,88	34.72	6.05	3.03	0.00
	21 Health insur.	379	31.06	22.00	0.00	0.26	12,93	32.19	3.15	3.15	0.00
	22 Health legal	2259	26.22	21.77	0.53	0.00	15,27	49.14	1.70	15.20	0.44
	23 Health legal	1200	23.13	20.90	0.42	0.00	16,83	46.08	1.60	11.50	0.00
	24 Health legal	1201	23.89	16.40	0.17	0.00	14,99	46.55	0.80	11.70	0.84
Target articles	25 Health exp-laym	1904	22.09	19.72	0.32	0.58	15,44	48.89	1.49	25.20	0.50
	26 Env. exp-laym	29,722	26.51	17.93	0.30	0.56	11,18	41.32	4.21	7.70	0.60
	27 Health exp-exp	34,030	30.11	29.00	0.79	0.14	10,27	37.34	0.65	13.61	4.17
	28 Env. exp-exp	111,904	24.23	20.01	0.60	0.07	12,55	44.24	4.10	8.27	0.69

Table 2 Applying quality indicators to test documents

In conclusion, applying all quality indicators to the test data with the threshold values listed in table 1 proves to be very useful to filter out 'poor' LSP and non LSP texts. All of the four target texts[6] survive the first step of the filtering process and are indeed identified as LSP texts. One of them, however, has an insufficient KP density and is not formally recognized as being a "good" LSP text. This might indicate that the tentative thresholds probably are too restrictive.

10. Perspectives and future work

Given this evaluation, we believe that the quality indicators suggested in this study have proven useful in the task of automatically assessing the terminological value of candidate corpus texts. The indicators need to be evaluated on a wider range of target, non-target and borderline texts so that the threshold values can be

[6] texts 25-28

more accurately optimized and fixed. For example, it proved difficult to set an upper limit on the density of passive verbs as this particular feature seems to be characteristic of both administrative texts and academic texts. Arguably, we need to find an additional feature with which administrative texts can be reliably eliminated without “collateral damage” so to speak.

Also, the range of text source domains needs to be expanded in order to assess to what extent the quality indicators (and their threshold values) can be regarded as domain independent.

A domain like Information Technology (IT), for example, may very well be characterized by a much lower OOV type density because many terms in this particular domain are coined by sense extension of lexemes of the general vocabulary (e.g. "window", "mouse", "desktop", etc.). Also, this particular domain is characterized by a very pronounced use of English loans, and this may very well reduce the number of long words in the texts because compounds are not rendered as continuous strings in English. For these reasons, IT will be the next domain on the project agenda.

Having tested (and adjusted) the quality indicators on additional domains, the next step is to implement the methods described in this paper as a quality assurance system in our corpus processing pipeline.

11. References

Agbago, A. and C. Barrière (2005) "Corpus construction for terminology" In: *Proceedings from the Corpus Linguistics Conference Series,* Vol. 1, no. 1, University of Birmingham, ISSN 1747-9398.

Barrière, C. and Agbago, A. (2006) "TerminoWeb: A Software Environment for Term Study in Rich Contexts" *International Conference on Terminology, Standardisation and Technology Transfer* (TSTT 2006), Beijing, China.

Bergenholtz, H. (1988) ”DK87: Et korpus med dansk almensprog” In: *Hermes - Journal of Language and Communication Studies,* Vol. 1, pp. 229-237, Aarhus University, Denmark.

Biber, D., S. Conrad and R. Reppen (1998) *Corpus Linguistics* Cambridge University Press.

Halskov, J. and C. Barrière (2008) "Web-based extraction of semantic relation instances for terminology work" In: *Terminology* 14:1, pp. 20-44, John Benjamins, Amsterdam.

Kilgarriff, A. and G. Grefenstette (2003) "Introduction to the special issue on the web as corpus" In: *Computational Linguistics,* Vol .29, no. 3, pp. 333-347.

Manning, C. D. and H. Schütze (1999) *Foundations of statistical natural language processing*. MIT Press.

Meyer, I. (2001) "Extracting knowledge-rich contexts for terminography" In: Bourigault, D; C. Jacquemin; M.-C. L'Homme (eds.) *Recent Advances in Computational Terminology*. John Benjamins.

Pearson, J. (1998) *Terms in Context*. John Benjamins.

Skovgaard Andersen, M., Asmussen, H. and Asmussen, J. (2002) “The Project of Korpus 2000 Going Public”. In: Braasch A. and Povlsen, C. (eds.) *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen. Pp. 291-299.

Electronic resources:

Korpus 2000
http://korpus.dsl.dk/korpus2000/indgang.php?lang=uk

KorpusDK
http://ordnet.dk/korpusdk_en/front-page/view?set_language=en

Retskrivningsordbogen (2001) at
http://www.dsn.dk/ro2001/