Adapting and evaluating a generic term extraction tool

Anita Gojun, Ulrich Heid, Bernd Weissbach, Carola Loth and Insa Mingers

Institute for Natural Language Processing
University of Stuttgart
Dräger Safety AG & Co. KGaA
Lübeck, Germany
gojunaa,heid@ims.uni-stuttgart.de, Bernd.Weissbach,Carola.Loth,Insa.Mingers@draeger.com

Abstract

On the basis of the term candidate extraction tools under development in the EU project TTC, we designed an application for German and English data that serves as a first evaluation of the approach and of the techniques for monolingual term candidate extraction used in the project. The application situation highlighted, among others, the need for tools to remove incomplete word sequences from multi-word term candidate lists, as well as the fact that the provision of German citation forms requires more morphological knowledge than TTC's slim approach can provide. In the detailed evaluation of our extraction results, we profited from interaction with domain experts and from the fact that the same texts were used for both manual and automatic term extraction.

Keywords: Terminology extraction, comparable corpora, metadata

1. Introduction

We present the extraction of term candidates from German and English texts from the domains of chemical protection suits and of alcohol and drug detection provided by Dräger Safety AG & Co. KGaA[1].

Our extraction tools are based in part on the tools under development in the EU project TTC[2]: these are aimed at genericity and at the use of slim linguistic knowledge. The present assessment serves as a test of these tools.

A major source of noise in multi-word term (MWT) extraction is the fact that often only parts of longer MWTs are extracted, which are either not terms or even formally incomplete. Therefore, we designed a tool to find the longest text-specific word sequences among those extracted by a set of terminologically relevant part-of-speech (POS) patterns. We produce XML output with text-related metadata, and we correct lemma and citation forms for both languages.

Section 2 gives an overview of the term extraction techniques used; we also describe the tool for removing incomplete sequences from the term candidate lists. Section 3 is devoted to the interplay between term extraction and computer–assisted translation (CAT), with emphasis on devices for producing morphologically correct data, such as lemma forms and citation forms. The latter problem clearly shows limitations of linguistically slim approaches. In section 4, we report on the evaluation of our tools and discuss general questions concerning the evaluation methodology.

2. A slim approach to term extraction

We make use of the term extraction approach adopted in the TTC project (cf. e.g. (Weller et al., 2011)). The extraction consists in two steps: (i) extraction of single-word (SWTs) and multi-word term (MWTs) candidates based on part-of-speech tags, and (ii) filtering of term candidate lists in order to identify domain-specific terms. For the second step, we use *weirdness ratio*, the domain specificity value defined by (Ahmad, 1992).

In the literature, many other statistical measures were proposed for the second step, e.g. frequency-based filtering (Daille et al., 1994), C/NC value (Frantzi and Ananiadou, 1999), TF/IDF (Paslaru et al., 2005), the log-likelihood ratio test (Rayson and Garside, 2000), etc.

The TTC approach is a hybrid one (cf. the classification provided by (Cabré et al., 2001)); for the first step (POS-based extraction of term candidates), we use the linguistic knowledge provided by POS-taggers, and for the second step, we only rely on (relative) frequency; thus, the tool only use tagging and frequency data. For the weirdness calculation, frequency data from a general language corpus are used, i.e. from texts which are not biased to any particular topic or domain. Thus, we avoid the use of tools requiring more (detailed) linguistic knowledge, as the term extractor is designed to operate analogously for several typologically different languages.[3]

In this paper, we do not propose a novel approach to terminology extraction: we rather explore the appropriateness of a well-known extraction method in a real application and show extensions needed in order to fulfill requirements arising in a practical working context.

2.1. Corpus preprocessing

The corpus preprocessing includes tokenization, lemmatization and POS tagging.

As our extraction tool considers lemmas and their frequency counts (cf. section 2.2.), it is important that lemma annotation is correct. Since we observed that the taggers we use often lemmatize domain-specific words incorrectly, we implemented an additional preprocessing tool which performs the correction of lemmas output by the taggers.

Lemma correction The lemma correction component is

[1] `http://www.draeger.com/GC/en/`

[2] TTC: Terminology extraction, translation tools and comparable corpora, `www.ttc-project.eu`.

[3] TTC deals with German, English, Spanish, French, Latvian, Russian and Chinese. Extensions to other languages should be easy to add, provided POS-tagging and data from a general language corpus are available.

based on (i) word similarity and (ii) a set of simple morphological rules.

Using word similarity allows for language independent grouping of similarly spelled words (i.e. inflected forms of one lemma), e.g. *[Chemikalienschutzanzug (chemical protective suite), Chemikalienschutzanzugs, Chemikalienschutzanzüge]*. Within such groups, the POS tags of the grouped words (containing morphological information: we use the MulText tagset[4]) are searched in order to find the form which corresponds to the lemma. Subsequently, all items of a group are assigned the same lemma.

The rule-based approach uses a set of language-specific morphological rules which map inflected forms to their lemmas, e.g. *cats '-s'* → *cat*. To avoid the generation of wrong lemmas (e.g. *analysis '-s'* → **analysi*), each generated lemma is searched in the corpus and retained only if found in the corpus.

Since the similarity-based approach may lead to an erroneous grouping of words, and thus to the assignment of erroneous lemmas, we use this method as a *backup* method in case the lemma could not be found by means of the available morphological rules. In section 4.3., we show that lemma correction has a positive impact both on the size of the candidate lists, as well as on their quality.

2.2. Extraction

In TTC, we use a list of pre-defined patterns based on POS tags to extract both single-word and multi-word terms. We consider noun and adjectival phrases as *base* terms. These can be easily extended in order to also consider different modifications of a head noun, such as adjectives, prepositional phrases, coordination, etc. These patterns are language-specific and were collected for each of the handled languages separately. An advantage of this method is that base terms and term variants (cf. (Daille, 2005)) can be handled with one and the same type of procedures. To identify the appropriate patterns, we analyzed the POS sequences found in ca. 100 multi-word terms contained in the terminological glossary of Dräger. As the patterns are provided to the tool in a separate parameter file, changes are easy to realize.

In table 1, a few extraction patterns are listed as examples. In addition to the patterns collected within the TTC project, we use a few domain-specific POS sequences, e.g. prepositional phrases with domain-specific prepositions: (EN) *resistance [to the permeation of chemicals]$_{PP}$)*, (DE) *Schutz [gegen flüssige Chemikalien]$_{PP}$ (protection against liquid chemicals)*.

Each pattern is processed separately resulting in a list of pattern-specific term candidates.

2.3. Filtering

Since in the TTC project, we aim at developing tools applicable to a number of different languages and domains, we implemented a rather simple filtering method. The calculation of the domain specificity *ds* of a term candidate is based on the comparison of relative frequencies of a term candidate in a domain-specific text and in a general language corpus, as described in (Ahmad, 1992). The equation for computing ds is given in (1).

$$ds(t) = \frac{\frac{f_s(t)}{size\ s}}{\frac{f_g(t)}{size\ g}} \quad (1)$$

For each extracted phrase t, we count how often its lemmatized form occurs in the domain-specific corpus ($f_s(t)$). Additionally, we count its occurrences in the general language corpus ($f_g(t)$). The frequency counts are normalized by dividing them by the number of extracted phrases for a given pattern. *ds* of t ($ds(t)$) is the ratio of relative frequencies of t in the domain-specific text and the general language corpus.

The extracted term candidates are sorted by descending ds values with domain-specific term candidates at the beginning of the list and more general terms at its bottom.

2.4. English noun sequences

The POS patterns described above contain optional elements. In addition, the tools do not make use of any identification or annotation of phrasal constructs (e.g. noun phrases). Thus, term candidates are extracted by patterns without considering the context they occur in. In this approach, English MWT candidates may pose problems, as incomplete sequences may be extracted, e.g. $^{??}$*rolled-up safety* vs. *rolled-up safety boots*, (cf. e.g. (Vu et al., 2008)). To identify such items, we check their contexts. If in the majority of corpus occurrences the term candidate occurs without surrounding nouns, it is considered to be an independent phrase and thus it is included in the term candidate list. Otherwise, it is discarded.

This kind of term independence definition is also used in the *C-value* calculus proposed by (Frantzi and Ananiadou, 1999). In contrast to (Frantzi and Ananiadou, 1999) who use this information to compute the domain specificity of a term candidate, we use the context information mainly to remove incomplete phrases.

However, this simple approach can also lead to the deletion of shorter complete phrases. The first experiments showed that frequency a ratio of 0.7 already yields good results, which means that only a small number of good term candidates gets omitted. But nevertheless, it still remains to be investigated which frequency ratio threshold provides the best results for identifying incomplete phrases.

The formulation of POS patterns (as shown in table 1) ensures that all sequences of potential relevance are found (recall), and this nestedness filtering ensures precision.

3. Preparing term candidates for CAT

The lists of extracted term candidates were processed to import them into a standard CAT tool. Terminologists are used to have German entries in their terminological database which are in the indefinite nominative singular, as is the case with citation forms in dictionaries. Since our tools operate on lemmatized texts, we had to generate these citation forms. In addition, we needed to derive documentation-related features (encoded as metadata).

[4] http://aune.lpl.univ-aix.fr/projects/Multext/

Pattern	English example	German example
ADJ	gastight	gasdicht
N	flammability	Entflammbarkeit
V	bleach	bleichen
ADJ N	switch-over valve	umschaltbares Ventil
N (PREP ADJ N)+	protection from mechanical stresses	Schutz vor mechanischen Belastungen
N (N)+	drug type	-
N (ART N)+	-	Innenraum des Anzugs (*inside of the suit*)

Table 1: Sample extraction patterns for German and English. Upper part: patterns for SWTs (nouns, adjectives, verbs), lower part: sample of the MWT patterns. Some parts of a pattern can be repeated, which is marked with *"+"*.

3.1. Citation forms

German. For inflecting languages such as German, it is insufficient to provide term candidate lists with lemmatized entries (e.g. *automatisch + Umschalter (automatic switch-over)*). In TTC, we search for citation forms of the term candidates in the set of the extracted inflected forms. But given the rather small size of the specialized texts, we cannot assume that the nominative singular of all terms can be found (in our texts, only for 40% of all terms).

The generation of citation forms requires more knowledge than available in the standard slim approach of the TTC project. Therefore, we implemented an additional function to generate correct German citation forms which is applied to term candidates whose head nouns are modified by an adjective. We resort to a full-scale morphological analysis and generation tool for German (SMOR (Schmid et al., 2004)) to overcome problems of German syncretism and weak vs. strong inflection. In a first step, it is used to derive the gender of the head noun. Secondly, we run SMOR to generate the corresponding adjective form in the nominative from the adjective lemma. In the citation form inventory, optional prepositional or genitive post-modifiers are given in the form most frequent in the corpus.

Given the lemmatized term candidate *beweissicher + Nachweis + von Droge (probative proof of drug)*, the generation of its citation form includes the following steps: (i) derivation of the gender of the head noun *Nachweis* (masculine), (ii) generation of the appropriate adjective form (*beweissicherer*), and (iii) derivation of the most commonly used form of the prepositional phrase modifying the head noun (*von Drogen*$_{Pl}$). The resulting citation form which is then provided to the user is thus *beweissicherer Nachweis von Drogen (probative proof of drugs)*, even if the term was found in the form *den beweissicheren Nachweis von Drogen* (accusative) in the data.

English. For English, there is no need for the use of morphological processing tools. For the head noun and an optional modifier, lemmas are used, while the non-heads are derived in the same way as described for German.

3.2. Term information in XML format

Terminologists and translators are not only interested in lists of term candidates but also in a number of additional types of information concerning each term: morphological information, as well as text type, document type, domain, etc. Such information may be encoded as metadata.

Our extraction tool outputs the required (language-specific) information, e.g. for German, the gender and number of the head noun is provided, derived from the corpus. To support interactive contextual checking of term candidates, we provide two example sentences per candidate, as well as metadata about their source.

The generated XML files, which are compatible with the CAT tool *SDL MultiTerm*, contain all relevant information from the tool output, as well as a feature that informs about their being a result of automatic term extraction. This information allows MultiTerm users to distinguish between terms provided by automatic means and manually collected ones.

4. Experiments and evaluation

4.1. Experimental setup

We tested our term extraction system on texts from the domains of alcohol and drug testing and of chemical protection suits. The alcohol and drug testing corpora contain 110,614 words (DE) and 370,176 words (EN). The corpora on chemical protection suits have 48,664 (DE) and 57,464 words (EN). Most of the texts are part of Dräger's technical documentation.

The extraction process follows the methodology described in section 2.. The English corpora were tagged with TreeTagger (Schmid, 1995), while the German texts were preprocessed with RFTagger (Schmid and Laws, 2008). We built two versions of each corpus: one with corrected lemmas and one with unmodified output of the two used taggers. For the filtering step, we used newspaper corpora as general language corpora. For German, we used *die tageszeitung (1987-1993)* consisting of 97 mio. tokens and for English *British National Corpus* with 117 mio tokens.

The extraction system outputs lists of term candidates ordered by their domain specificity values which contain additional information about the candidates (cf. section 3.2.). A sample term candidate list entry is shown in table 2.

4.2. Evaluation methodology

To evaluate the quality of the automatically extracted terms candidate lists, we compared them with reference lists collected manually by experts of the respective domains. These reference lists were collected independently from our extraction work. The list comparison was implemented as a strict string matching of the list entries.

Feature	Value
absolute frequency	8
relative frequency	0.00007
domain specificity	7058
lemma	analytisch Spezifität
POS	ADJA N
morphology	Nom.Sg.Fem
product group	Alcotest7110
language	DE
Texttyp	competitor literature
most frequent form	Analytische Spezifität
citation form	analytische Spezifität
example sentences	...

Table 2: Output for the extracted term candidate *analytisch + Spezifität (analytical specificity)*

	Alcohol/Drugs	Protective suits
Ref. nouns (RNs)	469	539
Added candidates	120	5

Table 3: German nominal reference terms (top) vs. extraction results

4.3. Evaluation results

4.3.1. Gain with respect to the existing term list

In a first step, term candidates were considered which had been extracted by the tools, but were not part of the reference lists. The top 500 candidates were inspected manually by the domain experts in order to identify gaps in the reference lists and thus candidates for inclusion into the existing terminology collection. The results are shown in table 3.

For the alcohol and drug testing domain, 120 candidates (i.e. 35% of the top 500 extracted non-reference candidates) were added to the existing terminology collection. For the domain of chemical protection suits, this was the case only for 5 terms, one of which is a synonym of an existing reference term: *Chemikalienschutzkleidung (chemical protective clothing)* (existing) vs. *Chemikalienschutzbekleidung* (added). Reasons for this discrepancy are (i) that the alcohol and drug measurement list was created without full access to the texts used for the automatic term extraction, while work on chemical protection suits was based on exactly the same text data in both manual and automatic extraction, and (ii) that the set objective in the manual term extraction work on protective suits was to achieve broad coverage of the domain, while the work on alcohol and drug measurement aimed at the most typical items only.

4.3.2. Comparison with updated reference lists

In a second evaluation, we compared the entire sorted German (cf. section 2.3.) noun candidate lists with the updated noun reference lists. We calculated precision and recall. The results are shown in figures 2 and 3.

Recall. For the two considered domains, our tool did not find 153 (ca. 35%) of the reference terms of the drug testing domain, and 127 (ca. 24%) of the reference terms from the domain of protection suits. Manual checking of the non-extracted reference terms showed that ca. 38% of this silence are due to the fact that terminologists who designed the reference lists included terms in these lists which were not attested in the analysed texts. Another 34% are variants and thus not counted as hits because of the strict string comparison used. The rest is due to erroneous POS tags[5] and text conversion errors[6].

Precision. For the domain of alcohol and drug testing (cf. figure 2), 51% of the extracted reference terms are found in the top 500 candidates. Similar results were obtained for the domain of protective suits, where 48% reference terms were found in the top 500 of the term candidates, cf. figure 3. This result is obviously due to the strong correlation of the domain specificity value (cf. (Ahmad, 1992)) with term frequency. Earlier experiments in the TTC project, with other statistical measures used to identify relevant terms did not suggest alternative measures to be superior to these results as shown in figure 1.

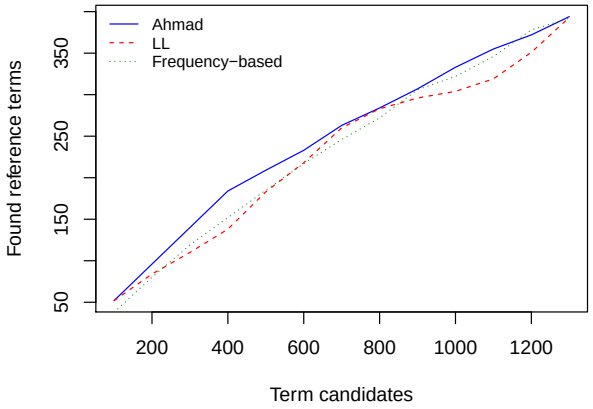

Figure 1: Top n German noun candidates extracted from chemical protective suits texts: comparison of different filtering measures.

The figures also show the impact of the lemma correction (cf. section 2.1.) on the quality of the term candidate lists. The precision of the lists extracted from corpora with corrected lemmas (indicated with solid lines) is consistently higher than that of the lists derived from the corpora with original lemmas (indicated with dotted lines). Although all corpora contain exactly the same set of tokens, the lists with corrected lemmas seem to contain more reference terms. This is due to the fact that without lemma correction, many word forms are erroneously taken to be lemmas in their own right, and that these are identified as term candidates. By "collapsing" all relevant forms into one lemma, we got a smaller lemma list (down from 6,155 without lemma correction to 5,591 nouns with it, i.e. a reduction of 10%); the calculation of domain specificity also profits from lemma correction, as data sparseness is reduced. As a consequence, we get more true positives, e.g. in the top 500 term candidates (cf. table 4).

[5] Some mistagged candidates show up in candidate lists for the wrong POS.

[6] These are mainly due to errors in PDF-to-text conversion.

	top 100	top 200	top 300	top 400	top 500
not corrected lemmas	31	61	97	128	151
corrected lemamas	35	74	112	137	154

Table 4: Number of found reference terms in top n of the German term candidates extracted from the alcohol and drug measurement corpus.

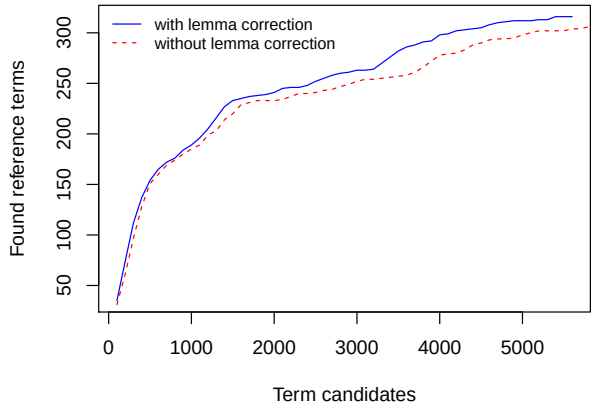

Figure 2: Evaluation results of the German noun candidates extracted from the corpus on alcohol and drug measurement, sorted by their domain specificity vales.

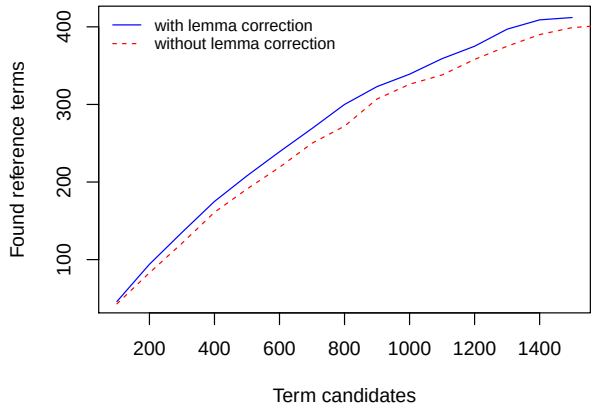

Figure 3: Evaluation results of the German noun candidates extracted from the corpus on chemical protective suits, sorted by their domain specificity vales.

4.3.3. Discussion

Recall. Many of the non-identified terms do not occur in the texts. This is because terminologists often add (generic) terms to their collection in order to make it logically consistent. Furthermore, text-preprocessing steps may lead to errors which affect negatively the extraction process.

Precision. Much noise stems from related domains such as computational control of analytical tools: *Codezahl (code number)*, *Dateneingabe (data entry)*, *Druckerpapier (printer paper)*, etc. Furthermore, there are general-language candidates which are used very often in the domain-specific texts and therefore get a relatively high *ds* value (e.g. *Querrichtung (transverse direction), Temperaturausgleich (temperature compensation)*, etc.). These issues concern all technical domains as most of them rely on basic techniques from other sciences; to identify such "common" scientific terms, contrastive extraction across different domains may be a useful approach (cf. (Fritzinger et al., 2009)).

Methodology. The strict string comparison does not identify a number of equivalent terms which show some variation. For example, the tool may find only inflected forms in the text (e.g. plural, considered to be the lemma), while the reference list contains singular (*gefährliche Stoffe (hazardous substance(s)*). Furthermore, the text may contain orthographical variants of the reference terms, e.g. *Chemikalien-Schutzanzug vs. Chemikalienschutzanzug* or compounds with or without transitional elements: *Anzugmaterial (suit material) vs. Anzugsmaterial*.
There are two ways to overcome this problem. With a simple string similarity measure like *Levenshtein distance ratio* (cf. (Levenshtein, 1966), (Wagner and Fischer, 1974)), such terms can be identified as equal and counted as a hit, e.g. *Schutzanzug-Trocknungsanlage (equipment for drying of chemical protective suits)* and *Schutzanzug-Trockenanlage*. The drawback of this simple method is the possibility to also consider similar, but unrelated words/phrases as hits, e.g. *Analyt (analyte)* vs. *Analyst (analyst)*. A more precise method consists in identifying term variants (e.g. orthographic, syntactic, etc. (cf. (Daille, 2005))) and considering them as hits. This implies (i) that the extraction tools also have to identify the term variants (which is the case for the tool being developed within TTC[7]) and (ii) that the reference lists contain not only reference terms, but also their variants and inflected forms. Such reference lists have been produced within the TTC project (cf. (Loginova et al., 2012)); within Dräger, the main interest in term variants is to include terms as forbidden variants into the terminological database; in the medium term, such knowledge will be used in style and consistency checking, by technical authors and translators.
Finally, reference lists may be more or less detailed, depending on terminologists' needs and preferences which has a significant impact on the evaluation results.

5. Conclusion and Future work

The TTC project develops a generic term extractor whose processing component is language-independent and which is parametrized by means of language-specific patterns and frequency data. This slim approach cannot cater for German citation forms. Therefore, we extended the existing tool and enabled the use of a full morphological system.
The evaluation against manually created reference data showed an acceptable recall (65% and 77%, resp., of the

[7]TTC TermSuite: `http://code.google.com/p/ttc-project/`

reference terms found by automatic means, approx. 50% in the top 500). It also provided inclusion candidates for the domain of alcohol and drug testing: 35% of the top 500 candidates not contained in the reference lists would indeed qualify for inclusion into these lists. This clearly shows that the identification of terminologically relevant data is highly dependant on numerous external influences which are hard to control; all figures of recall and precision of term extraction must thus be interpreted with caution.

Furthermore, our evaluation was based on strict string identity; accepting orthographical, morphological and some syntactic variants in the matching process and signalling the variants to the technical authors (e.g. for inclusion as "unwanted variants") is under way.

The output of the described term extraction tool will be used in experiments towards automatic provision of equivalence pair candidates, by means of term alignment, e.g. from translation memory data and, as planned in the TTC project, from comparable corpora.

Acknowledgement

The research leading to these results has received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248005.

6. References

K. Ahmad. 1992. What is a term? the semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Poechhacker, and K. Kaindl, editors, *Translation studies: an interdiscipline*, pages 267–278. 1st edition.

M. T. Cabré, R. Estopà, and J. Vivaldi. 2001. Automatic term detection: a review of current systems. *Recent Advances in Computational Terminology*, pages 53–87.

B. Daille, E. Gaussier, and J.-M. Lange. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING-94*.

B. Daille. 2005. Variants and application-oriented terminology engineering. *Terminology*, 11:181–197.

K. Frantzi and S. Ananiadou. 1999. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal of Digital Libraries*, 6:145–179.

F. Fritzinger, U. Heid, and N. Siegmund. 2009. Automatic extraction of the phraseology of a legal subdomain. In *XVII European Symposium on Languages for Specific Purporses*.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

E. Loginova, A. Gojun, H. Blancafort, M. Guégan, T. Gornostay, and U. Heid. 2012. Reference lists for the evaluation of term extraction tools. In *Proceeding of Terminology and Knowledge Engineering Conference*.

E. Paslaru, D. Schlangen, and S. Niepage. 2005. Ontology eegineering for the semantic annotation of medical data. In *4^{th} workshop on web semantics – DEXA2005*.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on comparing corpora*.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of COLING 2008*.

H. Schmid, A. Fitschen, and U. Heid. 2004. Smor: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *ACL SIGDAT Workshop*.

T. Vu, A. Aw, and M. Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of the 3^{rd} international joint conference on natural language processing*.

R. A. Wagner and M. J. Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21.

M. Weller, H. Blancafort, A. Gojun, and U. Heid. 2011. Terminology extraction and term variation patterns: A study of french and german data. In *GSCL 2011: Multilingual Resources and Multilingual Applications*.