Simple methods for dealing with term variation and term alignment

M. Weller, Anita Gojun, Ulrich Heid, Béatrice Daille, Rima Harastani

Simple methods for dealing with term variation and term alignment

Marion Weller, Anita Gojun, Ulrich Heid
Universität Stuttgart
gojunaa@ims.uni-stuttgart.de
wellermn@ims.uni-stuttgart.de
heid@ims.uni-stuttgart.de

Béatrice Daille, Rima Harastani
Université de Nantes
Beatrice.Daille@univ-nantes.fr
rima.harastani@etu.univ-nantes.fr

Abstract

In this paper, we deal with bilingual terminology extraction from comparable corpora. The extraction can be seen as a pipeline of processing steps. We will discuss grouping of term variants and describe two methods for bilingual term alignment of neoclassical terms: a knowledge-poor approach using string similarity measures and a linguistically motivated approach which is extended to cover German compound nouns.

1 Introduction: background

The work described in this paper is part of an attempt at term candidate extraction from comparable corpora; single-word terms and multi-word terms are extracted from content-wise related texts of different languages, and items from two languages are grouped into equivalence sets (term alignment). Tools[1] for this task must be aware of term variation, as synonymous or closely related variants of a given language may all be mapped onto a common equivalent of another language.

Variation concerns multi-word expressions (e.g. FR *production électrique* vs. *production d'électricité*), but also complex word formation products, especially German compounds, e.g. DE *Stromproduktion* vs. *Produktion von Strom* (*production of electricity*). Many terms contain neoclassical morphemes, either alone or in combination with native or other neoclassical elements.

[1]This work is part of the EU-project TTC, *Terminology Extraction, Translation Tools and Comparable Corpora*. The research leading to these results has received funding from the European Commission's Seventh Framework Programme (FP-7, 2007-2013), under grant agreement no. 248.005.

1.1 Application domain – domain independent tools

Our experiments concern the domain of wind energy; there are only few monolingual glossaries available for this domain[2], and no bilingual ones for some of the language pairs covered in the project[3]. The domain is characterized by the following tendencies:

– technical interdisciplinarity, with terminology from many fields (e.g. engineering, physics);

– discourses from different viewpoints (e.g. technical, legal, economic);

– rapid development, with many players involved, leading to massive variation in the terminology used and to terminology with a short life span.

Even though we use wind energy as a test domain, our tools are not dependent on particular fields and can thus be used for a range of domains.

1.2 Tool scenario

Translators or terminologists working with new and rapidly evolving domains often need to create their own terminolgical data collections, e.g. from texts available on the Internet. The users should be able, with the tools under development, (i) to collect texts from a given domain using a focused crawler[4], (ii) to process them in order to find term candidates for each language involved, and (iii) to

[2]Glossaries on *wind energy* (DE):
www.strom-magazin.de/energie-lexikon/
www.energieinfo.de/eglossar/
www.energie-lexikon.info/glossar.html

[3]Languages: Spanish (ES), French (FR), English (EN), German (DE), Latvian (LV), Russian (RU) and Chinese (ZH).

[4]At this step, meta-data has to be collected and passed through during the terminology extraction process.

Long papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, pages 87–93
Paris, 8–10 November 2011

align the term candidates of two languages into equivalence pair candidates. In this article we focus on methods for steps (ii) and (iii). A user evaluation is planned for 2012.

1.3 Target of term extraction

Term candidate extraction targets both single-word and multi-word term candidates; single-word terms may be simplex terms or morphologically complex words; in the Romance languages, the latter are mostly derived words (FR: *énergétique, production*), while German provides mainly compounds (DE *Emissionsgrenzwert* (*limit for emissions*)). In our German texts, approximately 52% of all nouns are morphologically complex.

Across typologically different languages, morphologically complex words are often equivalents of multi-word items (e.g. DE *Produktionsstandort* ↔ FR *site de production*, cf. section 4.3).

2 Tools for term extraction and variant recognition

As we aim at domain-independent tools to be used on different languages, we do not rely on very detailed language-specific knowledge: we aim at striking a balance between the genericity of the tools and the precision achievable with "slim" linguistic resources.

2.1 Term candidate extraction based on POS-annotated corpora

One of the project's approaches to monolingual term candidate extraction is based on shallow corpus linguistic annotation (tokenizing, POS-tagging and lemmatization: TreeTagger (Schmid, 1994)).

To find term candidates in texts, we use extraction patterns formulated in terms of POS-tags and lemmata (e.g. of prepositions or other closed-class words). Pattern-based term extraction provides morphosyntactically homogeneous result sets. As a second step, we suggest filtering term candidates against general language data, as proposed by (Ahmad et al., 1992).

2.2 Dealing with unknown word forms

The technology described in the previous paragraph is relatively robust, although obviously POS-tagging tends to degrade when faced with words not contained in the resources of the tagger; to be able to "repair" such cases later in the process, we keep the inflected forms of all words that are unknown to the tagger, and we experiment with two approaches for grouping them. Both approaches take the term list consisting of lemmas and surface forms and compute groups of related terms. The first method is based on a string similarity measure and is nearly language-independent, whereas the second one makes use of rules which model related inflectional affixes.

2.2.1 String similarity

To group inflected forms we use an *adapted Levenshtein distance ratio*. The idea is to compare terms of the same POS shape with each other and then to create groups consisting of terms with similar surface forms. The terms can be both single-word and multi-word terms. Before string comparison is carried out, we lowercase all words. Additionally, only words which begin with the same letter are treated as similarity candidates.

The computation of the term similarity is given in equations (1) and (2). For each component w_i of a term t, the highest Levenshtein similarity $w\_sim(w_i)$ is computed, given k components of the similarity candidate t'. The similarity between t and t' is the ratio of the sum of the maximum component similarities and the length of term t ($len(t)$).

$$w\_sim(w_i) = max[lev(w_i, w_{j \in 1...k})] \quad (1)$$

$$lev(t, t') = \frac{\sum_{i=0}^{n} w\_sim(w_i)}{len(t)} \quad (2)$$

Note that this computation of term similarity is adequate only for term pairs of the same POS shape, and thus of the same length.

The crucial point here is the minimum similarity value which has to be given in order to consider two terms similar. Our experiments showed that this value is language-dependent.

In figure 1, examples of term groups identified for English and German are given. Observing term groups computed by considering different thresholds, we finally set the threshold for German and English to 0.9. A high threshold value allows only rather small string differences like different suffixes, but it provides satisfactory results for German and English.

DE nouns		EN nouns	
Alpha-Strahler		by-product	
Alphastrahler	0.96	byproduct	0.94
Energie-Versorgung		interconnection	
Energieversorgung	0.97	Interconnection	1
photoelektrischer		electric machine	
photoelektrischen	0.94	electrical machine	0.9

Table 1: Example of term groups for German and English derived using Levenshtein distance ratio.

2.2.2 Inflection: rule-based approach

An alternative method to group inflected forms of the same lemma consists in modelling inflectional affixes of nouns and adjectives. Table 2 shows regular expressions for the lemmatization of French and German (relational) adjectives.

DE		FR	
suffix	inflection	suffix	inflection
-bar, -ell	-e	-ique, -oire	
-end, -in	-er	-if, -aire	-s
-isch, -ar	-en	-eur	
-lich, -elt	-es	-é, -al	-e, -es, -s
-ig	-em	-iv → -if	-e, -es

Table 2: Rules allowing to lemmatize French and German adjectives for which TreeTagger failed to find a lemma: inflectional endings are reduced to the respective suffix.

3 Identification of term variants

3.1 Typology of term variants

Our work on term variation takes (Daille, 2005) as a starting point and adapts it to cover the languages dealt with in the project; we see "a variant of a term" as "an utterance which is semantically and conceptually related to an original term". The variant relation is oriented (X is a variant of term Y, where Y is the "base term"). The following are examples of our typology of term variation:

- Graphical variants:
  EN *air flow* ↔ *airflow*,
  FR *rotor multipale* ↔ *rotor multi-pale*
- Morphological variants
  (inflection, derivation, compounding):
  DE *Stromerzeugung* ↔ *Erzeugung von Strom*
  EN *power generation* ↔ *generation of power*
  (cf. also table 3)
- Syntactic variants, e.g. coordination:
  DE *sichere Energieversorgung* vs. *sichere und nachhaltige Energieversorgung*
  EN *safe energy supply* vs. *safe and sustainable energy supply*

3.2 Pattern-based variant identification

Sets of synonymous or related terms are identified via POS-based term patterns. The examples in the upper part of table 3 will be discussed in detail in section 4.3, whereas the rules in the lower part give an impression about further term patterns, but are not dealt with in this work (for example, refer to (Weller et al., 2011)).

$N_1 \mid N_2$ ↔ N_2 *für* N_1	*Emissionsgrenzwert* ↔ *Grenzwert für Emissionen*
N_1 N_2 ↔ N_2 *of* N_1	*energy production* ↔ *production of energy*
N_1 *de* N_2 ↔ N_2 VPART	*production d'énergie* ↔ *énergie produite*
N_1 *de* N_2 ↔ N_1 $ADJ_{relational}$	*source de lumière* ↔ *source lumineuse*

Table 3: Term variation patterns.

3.3 Data-driven derivation of term pattern equivalences

Term pattern equivalences can also be automatically derived by comparing terms of different POS patterns. Term similarity is computed word-wise in a similar way as described in section 2.2.1. Then, POS patterns of identified similar terms are counted, allowing to derive pattern equivalences from the POS counts of similar terms. These frequencies can also provide clues about pattern productivity and point to other aspects of relatedness.

Table 4 shows automatically gained statistics about related patterns for English. The upper part of the table shows results of a comparison of 5,236 terms of the pattern N N with terms of other patterns. There are, for example, 202 identified similar terms of the pattern N of N (e.g. *energy source* ↔ *source of energy*). On the other hand, if the patterns are extended by an adjective (e.g. *alternative energy source*, *alternative source of energy*), significantly less similar terms are identified. This fact indicates that this kind of a pattern equivalence is not productive.

The comparison of N N terms with terms of the pattern N N N showed that a relatively large amount of shorter compounds are related with longer compounds which are composed of three

nouns (complex heads, e.g. *battery pack* → *hydride battery pack*). Moreover, there can be more than one term of the other pattern which is similar to the given term, e.g. *battery pack* → {*hydride battery pack, ion battery pack*}). We identified 1,332 similarity pairs for $N\ N$ and $N\ N\ N$ which means that in some cases, the similarity relationship 1:*n* is given.

We also use these procedures in order to filter out incomplete sequences which are a part of larger terms (e.g. **production of wind* vs. *production of wind energy*).

N + N →	**Total num. of terms**	**Num. of sim. terms**
N + of + N	929	202
N + N + N	806	1332
ADJ + N	7131	148
ADJ + N + N →	**Total num. of terms**	**Num. of sim. terms**
ADJ + N + of + N	199	7

Table 4: Counts of identified similar terms generalized in terms of their POS shapes for English. *Total num. of terms* is the number of terms extracted for a given pattern. *Num. of sim. terms* is the number of term pairs which belong to the compared patterns and which have been identified as similar.

The discussed examples rely on already known pattern equivalences for English, but the method can be used for any language for which the knowledge about variation patterns is still to be collected.

4 Term alignment

Term alignment takes lists of source and target language term candidates as input and returns pairs of term-equivalent candidates[5]. For most term alignment procedures, a bilingual dictionary is required: they largely depend on the quality and size of the used dictionary. While general language dictionaries are available for most language pairs, they are not likely to provide sufficient coverage for terms of technical fields. With the methods presented in this work, we aim at enriching available bilingual dictionaries.

Terms of neoclassical origin play a significant role in scientific texts (Namer and Baud, 2007), (Deléger et al., 2009), they are not very likely to occur in a general language dictionary. Neoclassical terms are usually very similar across different languages (internationalisms, e.g. *calorimétrie* vs. *Kalorimetrie* vs. *calorimetry*). As a first step to enrich a general language dictionary with domain-specific entries, we use two methods to find translations of neoclassical terms without relying on a regular bilingual dictionary. Making use of similar surface forms of neoclassical terms in different languages, the first method is again based on string similarity. In our second approach, we model neoclassical word formation by using a multilingual list of neoclassical roots.

[5]There are several approaches for term alignment, for example (Déjean and Gaussier, 2002) or (Morin et al., 2007).

DE – EN nouns		**EN – LV nouns**	
Biomasse		accumulator	
biomass	0.92	akumulators	0.81
Elektromagnet		aerodynamics	
electromagnet	0.93	aerodynamika	0.75
Polarisation		anemometer	
polarization	0.93	anemometers	0.75
Oszillation		cylinder	
oscillation	0.9	cilindrs	0.75

Table 5: Sample of lemmatized internationalisms identified for German – English and English – Latvian with Levenshtein distance.

4.1 Detecting internationalisms by considering string similarity

Similarly to the approach proposed by (Koehn and Knight, 2002), we aim at extending the bilingual dictionary with internationalisms by computing term similarity across languages. The identified equivalents can also serve as an input to the automatic extraction of rules of neoclassical morpheme pairing for an arbitrary language pair.

Although the method is language-independent, it is still necessary to adapt the threshold to a given language pair. For EN ↔ DE, experiments suggest a relatively high threshold, whereas for EN ↔ LV, a lower threshold is required to capture the string differences, cf. table 5.

4.2 Modelling neoclassical word formation

Neoclassical terms can be decomposed into morphemes which are mostly of Greek or Latin origin, namely *roots*, *transitional elements* and *suffixes*. For equivalent neoclassical words from different languages, we expect isomorphic structures; i.e. decomposing equivalent neoclassical terms into

	R1	tr. el.	R2	suffix
Kalorimetrie	kalor	i	metr	ie
Seismograph	seism	o	graph	
Polygon	poly		gon	

Table 6: Decomposing German neoclassical nouns

	number	DE	FR	EN
R1	474	archä kalor seism	arché calor sism	archae calor seism
R2	174	graph meter oid	graphe mètre oïde	graph meter oid
suffix	27	asma ik ität	asme ique ité	asm ics ity

Table 7: Multilingual list of neoclassical roots.

basic units should lead to the same set of equivalent roots for the respective languages.

In this work, we focus on terms of the structure [*Root*$_1$ · *optional transitional element* · *Root*$_2$ · *optional suffix*], cf. table 6. Of course, neoclassical terms are not restricted to this structure; they may even include native components, as e.g. *Ausgangsparameter* (*default parameter*), cf. section 4.3.

As a basis for translating neoclassical terms, we use a manually compiled list of roots[6] containing the respective equivalents in German, French and English, see table 7. The root components of a source-language term are then separately translated into the target language:
(DE) *seism* · *o* · *graph* → (FR) *sism* · *o* · *graphe*.
The next step consists in checking whether the candidate translation is among the extracted terms of the target language; this guarantees that the candidate is an existing target term, and not a random match, like *ur* · *meter* (DE: *prototype meter*) → **ur* · *mètre*. This method also deals with overgeneration: (FR) *-ique* → (DE) {*-ik, -ikum*}, e.g *thermodynamique* → *Thermodynamik*, ~~*Thermodynamikum*~~. The incorrect form **Thermodynamikum* is discarded for not being in the target language term list.

For the language pairs DE ↔ FR and DE ↔ EN, we carried out an experiment, using all nouns of the respective languages as input. Results are shown in table 8 (corpora and dictionaries are described in section 4.4). In the case of DE→FR, 363 nouns could be decomposed and translated using the list of neoclassical roots. Of the resulting 363 generated translations, 164 were found in the target language term list (TL-terms), and of those, 163 were correct. In an attempt to measure the usefulness of our method, we checked a general language dictionary for these words: while one can already find a dictionary entry for 84 terms, the rest are new translations. Among those terms whose proposed translation was not in the target language term list, roughly half of the translations are correct.

	de–fr	fr–de	de–en	en–de
transl. cand.	363	429	364	315
in TL-terms	164	170	148	155
correct	163	167	147	151
in dictionary	84	86	136	137
not in TL-terms	199	259	216	160
correct	98	203	109	93

Table 8: Results: aligning neoclassical terms.

Generally, we find that the proposed translations which could be found among the set of target language terms are nearly always correct[7], while the precision of the non-verified terms is only 50% - 78%. This is partially due to erroneously decomposed native words (e.g. *Hämmer* (*hammers*) → **häm* · *mer* → **hém* · *mère*), but also to a mismatch of the proportion of neoclassical words in the respective language. For example, French terms like *aéroplane* or *aérostat* (*aircraft*) are expressed by native words in German (*Luftfahrzeug*).

Particularly for DE ↔ FR, a considerable amount of the attested translations is not in our general language dictionary and can thus be considered as an enrichment of the dictionary.

4.3 Translating compound words

The previously presented approaches are restricted to terms which are entirely composed of neoclassical roots. In order to translate terms containing native elements, we need to introduce a dictionary. The method of identifying an alignment between source and target terms follows the

[6] Based on examples for neoclassical word formation on `www.canoo.net` and literature (Béchade, 1989).

[7] Our method can generate several translation possibilities: multiple translations of one source term were gathered and only counted once.

above model: by splitting a complex source term into smaller units, individual translations for each unit can be found either (i) among translated terms consisting only of neoclassical elements or (ii) in a bilingual dictionary covering native words.

This method is applied to native German compounds: productively built compounds do not always occur in a dictionary. By splitting such words into their morphemes and translating these individually, we might be able to find a (multiword) translation which is part of the set of extracted target language terms.

We distinguish the following types:

- terms consisting only of neoclassical components (covered in section 4.2)
- DE: compound words of the type $N_1|N_2$, where N can be either native or neoclassical:
  – neoclassical: *elektronen*$_N$|*mikroskop*$_N$ (*electron microscope*)
  – mixed: *blei*$_N$|*isotop*$_N$ *(isotope of lead)*
  – native: *kupfer*$_N$|*rohr*$_N$ *(copper pipe)*

German compounds of the type $N_1|N_2$ are very common, and are often translated by $N\ PRP\ N$ structures (EN, FR), e.g. *Kupferrohr* ↔ *tuyau de cuivre* or by $N\ N$ structures (EN), cf. table 3.

Compound nouns are split using a statistical splitter based on (Koehn and Knight, 2003), which is able to deal with transitional elements, e.g. the s in *Korrosionsschutz*. For each complex noun of the structure $N_1|N_2$, translations of N_1 and N_2 are individually looked up in the dictionary and then compared with target language terms:

Korrosionsschutz → *korrosion*|*schutz* [8]
korrosion → *corrosion*
schutz → *protection*

Searching in the sets of extracted English $N\ PRP\ N$ and $N\ N$ structures, we can align *Korrosionsschutz* with both *corrosion protection* and *protection against corrosion*. In our current version, we list all attested translation candidates; however, one could simply use the frequency of the target terms as an indicator of how reliable each translation might be.

A similar method of rewriting terms is presented in (Morin and Daille, 2009) where the authors use the relation between relational adjectives and the nouns they are derived from (e.g. *lait* ↔ *laitier*) for aligning French and Japanese terms.

[8] During alignment, words are lowercased.

	de – fr	de – en
in general language dict.	77	630
new translation: neoclassical	9	1
graph. variation: hyphen	0	5
terms with new translation: (1)	152	85
terms with new translation: (2)	-	163

Table 9: Translations for 2000 German nouns. (1) denotes the equivalence pattern $N_1|N_2 \leftrightarrow N_2$ PRP N_1 (used for DE–EN, DE–FR) and (2) the pattern $N_1|N_2 \leftrightarrow N_1\ N_2$ which is used only for DE – EN.

In section 3, we presented methods to group attested monolingual term variants. The results can be used for interactively developing and managing terminological resources. For automatic term alignment, attested monolingual variants can be used: *Energieproduktion* → *Produktion von Energie* → *production of energy*. However, source language terms do not necessarily have variants whose structures match those of the target language terms, e.g. *Windgeschwindigkeit* (*wind speed, vitesse du vent*).

Combining the available knowledge sources, i.e. monolingual variation patterns, compound splitting and equivalence relations between term patterns of different languages, allows to obtain alignments without relying on attested variants: by generating 'potential terms', the equivalence relation between patterns of different languages can be used. This strategy enables us to identify considerably more term alignments than relying only on monolingual variants.

4.4 Experiment

For the 2.000 most domain-relevant German nouns (ranked according to (Ahmad et al., 1992), cf. section 2.1), we tried to find an equivalent term within the set of extracted target language terms, using the above methods. Terms are extracted from web-crawled corpora (Groc, 2011)[9]. We use two general language dictionaries with ca. 30.000 entries (DE – FR) and 820.000 entries (DE – EN)[10].

The first step consists in identifying those nouns which are directly found in the dictionary, and whose translation occurs in the set of extracted terms; verifying the translation proposed by the dictionary also helps to avoid out-of-domain trans-

[9] DE: 1.5 Mio, FR: 1.75 Mio, EN: 1.65 Mio tokens.
[10] Both taken from `www.dict.cc`

lations, e.g. *Strom* (*"electric current"*) $\not\rightarrow$ *torrent*.

For the remaining terms, we search translations based on (i) graphical variants (with/without hyphen, cf. section 2.2.1), (ii) the list of translations produced for neoclassical terms (cf. section 4.2), as well as (iii) by applying the pattern equivalence rules (cf. section 4.3). Results are given in table 9. This test set of 2000 words contains only a fraction of the neoclassical terms from the previous experiment; thus, only few new translations of this category could be found.

Given that the DE-EN dictionary is many times the size of the DE-FR dictionary, it is not surprising that considerably more alignments were identified for DE-EN. As the experiment is ongoing at the time of writing the paper, no full evaluation is yet available; however, the results look promising, as a decent amount of new alignments was found for both language pairs.

5 Conclusion

In this paper, we described methods for terminology extraction from comparable corpora, focusing on the identification of term variants (graphical, inflectional and morpho-syntactic), providing information which will be useful for end users (translators, terminologists), as well as for the task of bilingual term alignment.

As a second step, we investigated a knowledge-poor approach for aligning internationalisms, which is completely language independent and does not need any lexical resources. We then presented a method for modelling neoclassical word formation, relying on a bilingual list of neoclassical morphemes. This approach was expanded to deal with productive German compound nouns.

By combining monolingual variation patterns and equivalence patterns between terms of two languages, as well as using the translations of neoclassical compounds, we increased the amount of German (compound) nouns for which an alignment could be found.

We intend to explore the semi-automatic derivation of monolingual variation patterns, particularly for languages where no set of equivalence rules is available. Also, we only worked on the alignment of German compounds of the type $N_1|N_2$: we plan to investigate alignment strategies for more complex compound words.

References

Khurshid Ahmad, Andrea Davies, Heather Fulford and Margaret Rogers. 1992. *What is a term? The semi-automatic extraction of terms from text*. Translation Studies: An Interdiscipline, John Benjamins, Amsterdam, 1994, pp. 267-278.

Hervé-D. Béchade. 1989. *Phonétique et morphologie du français moderne et contemporain*.

Béatrice Daille. 2005. *Variants and application-oriented terminology engineering*. Terminology, vol. 1, pp. 181-197.

Hervé Déjean and Éric Gaussier. 2002. *Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables*. Lexicometrica, Alignement lexical dans les corpus multilingues, pp. 1-22

Louise Deléger and Fiammetta Namer and Pierre Zweigenbaum. 2009. *Morphosemantic parsing of medical compound words: transferring a French analyzer to English*. I. J. Medical Informatics, volume 78 Suppl 1, pp. 45-55.

Clément de Groc. 2011. *Babouk: Focused web crawling for corpus compilation and automatic terminology extraction*. Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Lyon, France, 2011.

Phillip Koehn and Kevin Knight. 2002. *Learning a Translation Lexicon from Monolingual Corpora*. Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX), Philadelphia, 2002, pp. 9-16.

Phillip Koehn and Kevin Knight. 2003. *Empirical Methods for Compound Splitting*. Proceedings of EACL, 2003, pp. 187-193.

Emmanuel Morin and Béatrice Daille. 2009. *Compositionality and lexical alignment of multi-word terms*. Language Resources and Evaluation, volume 44, pp. 79-95

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, Kyo Kageura. 2007. *Bilingual Terminology Mining – Using Brain, not brawn comparable corpora*. Proceedings of ACL 2007, pp. 664-671.

Fiammetta Namer and Robert H. Baud. 2007. *Defining and relating biomedical terms: Towards a cross-language morphosemantics-based system* in I. J. Medical Informatics, volume 76, pp. 226-233.

Helmut Schmid. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proc. of the Int. Conference on New Methods in Language Processing, Manchester, pp: 114-133.

R.A. Wagner and M.J. Fisher. 1974. *The string-to-string correction problem*. Journal of the Association for Computing Machinery (ACM). vol. 21, 168-173.

Marion Weller, Helena Blancafort, Anita Gojun, Ulrich Heid. 2011. *Terminology extraction and term variation patterns: a study of French and German data*. GSCL 2011, Hamburg, Germany.