Term candidate extraction for terminography and CAT: an overview of TTC

Ulrich Heid & Anita Gojun

Keywords: *Terminology extraction, computer-assisted translation, term alignment*.

Abstract

In this paper, we present a tool chain for terminology extraction and term alignment which is under development in the EU-project TTC.[1] The tool components comprise the crawling of domain-specific text from the internet, in different languages, linguistic pre-processing of the corpus collected in this way, and the extraction of term candidates. Extracted term candidates of two languages are aligned into pairs of source and target term equivalents. This output can be used both in interactive translation setups (e.g. computer-aided translation) and in machine translation.

1. Introduction

This paper is about tools for the extraction of specialized terminology from texts. The tools are being developed in the EU-funded project TTC, *Terminology extraction, Translation Tools and Comparable Corpora*[2], and their output is mainly meant to serve computer-assisted translation (CAT) and machine translation (MT). The tools can however equally well be used to prepare partial input for terminographic work that leads to (electronic) dictionaries or glossaries that contain terms and their variants from domain-specific texts.

We describe the context and motivation of the TTC tools, as well as their purpose in section 2. In section 3, we give more details on the overall architecture and on the current (intermediate) state of the tools. Section 4 will be devoted to the lexicographic data types that are the output of the tools, and to the parametrizability of this output according to different user profiles. We conclude in section 5.

2. Motivation and context

2.1. *Term variation*

In many domains, terminology has been standardized, described in dictionaries and term banks and thus to a certain extent codified. Very prominent examples are chemical substance names, which are governed by IUPAC's nomenclature rules.

In new or upcoming domains, the situation is completely different. Rapidly evolving research areas are multidisciplinary, and they tend to involve numerous, uncoordinated players: scientists, companies, official bodies, etc. Such domains tend to be characterized by massive terminological variation: several single-word or multi-word terms may synonymously denote the same object or phenomenon, or there may be paraphrase-like variants. In addition, the occurrence of a given concept in a text with a certain syntactic "shape" may force writers to deviate from the "standard" form of terms: for example, the noun compound DE *Meeresboden* (ground of the sea) is "divided" into its components in a coordination, such as *am Boden von Meeren und Flüssen (on the ground of the sea and of rivers),* cf. (Weller et al., 2011). For translators, dealing with the sometimes considerable amount of variation is not easy: they need to know about each item's status, the sources it comes from, its linguistic properties, and its translations. (Daille, 2005) found that between 15

and 35% of all term candidates identifiable in technical texts were variants.

As, for upcoming topics, often there are not even handbooks or other "authoritative" texts, translators need to rely also on internet texts; and the use of internet sources contributes to the variation issue, as authors are free to post texts on the Web without any previous terminological verification.

For new and rapidly evolving domains, not only terminology is not standardized, but also very few parallel texts are available; while many of the players involved in new developments produce texts in their own language, typically texts provided in several languages are scarce; thus term candidate extraction must start rather from comparable corpora than from parallel text.

As discussed above, a tool that extracts term candidates from existing text and provides their translation candidate(s) and which can work on web data, needs to be able to detect term variation. Examples of the term variant types we deal with are shown in table 1. For the full picture, see (Daille, 2005).

Table 1. Examples of term variant types.

Type	**Semantic relation**	**Example**
graphical	identical	(EN) byproduct ↔ by-product (FR) énergie éolienne ↔ energie eolienne *(wind energy)*
morphological	related	(FR) production d'énergie ↔ énergie produite (*energy production vs. produced energy*) (FR) synchrone ↔ hypersynchrone *(synchrone vs. hypersynchrone)*
syntactic	synonymous	(DE) Energieversorgung ↔ Versorgung mit Energie (*energy supply vs. supply with energy*) (ES) potencial eólico ↔ potencial del viento *(wind energy potential)*
	related	(EN) renewable energy ↔ renewable and sustainable energy (LV) vertikālā ass ↔ vertikālā rotācijas ass (*vertical axis vs. vertical rotation axis*)
transpositional	synonymous	(ES) tripala ↔ pala con tres hélices *(three-bladed)* (LV) siltumnīcefeka gāze ↔ SEG *(greenhouse gas)*

2.2. *Outline of the TTC tools*

In the TTC project, we aim at designing a tool chain that implements the full pipeline from data search on the internet to bilingual equivalence candidate identification ('term alignment').

Data acquisition starts from a small number of seed words, typically 5 or 6 terms of a given technical domain, which are input to the focused crawler 'Babouk' (de Groc, 2011). This tool finds texts which contain instances of the seed terms and thus may be considered as typical for the targeted domain; it can be run on all TTC languages, and users can define how many texts they wish to be retrieved and how deeply the crawler is supposed to follow links

contained in the documents it finds. If equivalent seed words from two languages are used to search texts on the internet, comparable corpora will be retrieved, i.e. monolingual corpora from the same domain.

The retrieved material is automatically annotated with part-of-speech (POS) tags (e.g. using TreeTagger (Schmid, 1994) or similar tools) and with morpho-syntactic properties, such as number, gender or case (e.g. RFTagger (Schmid and Laws, 2008)).[3] As an alternative to tagging with a POS-tagger, machine learning-based approaches to the annotation of corpus data are also used.

In a subsequent step, term candidates are extracted automatically from the annotated texts: a standard approach used for this purpose is a combination of (i) pattern-based candidate identification (cf. table 2) and (ii) statistical ordering of the retrieved candidates, by decreasing 'domain-specificity' (or better: specificity for the corpus at hand, cf. (Ahmad et al. 1992)). These procedures can again be applied to all TTC languages, as POS patterns have been identified that provide term candidates made up of two or three elements, for all TTC languages. Obviously, patterns that are aimed at finding longer multi-word term candidates retrieve considerably large amounts of noise.

Table 2. Example POS patterns.

POS pattern	Example
N	(LV) ģenerators *(generator)*
ADJ + N	(DE) erneuerbare Energie (*renewable energy*)
N + ADJ	(FR) parc marin *(sea park)*
N + N	(LV) vēja enerģija (*wind energy*)
N|N	(DE) Energieversorgung *(energy supply)*
N + PREP + N	(EN) consumption of energy
N + PREP|DET + N	(ES) imanes del estator *(magnet of the stator)*

To provide bilingual term candidate data, i.e. equivalence pairs L1↔L2, term candidates from the previous step need to be 'aligned', i.e. items from L1 and L2 must be identified as being equivalent.

The term alignment combines lexical approaches and contextual ones. The lexical approach consists in making as much as possible use of general language dictionaries. As many terms will not be contained in freely available online bilingual dictionaries, additional devices are used to enhance the coverage of the tool's dictionary; one is a rule-based component for the translation of neoclassical terms (e.g. *chromatography*, *thermodynamics*, etc.). Another one is compound splitting; its purpose is to identify the morphemes of which Germanic compounds are composed, in order to allow for their individual translation. Further devices include derivational rules to relate morphological and syntactic variants.

The following are monolingual examples for which the above mentioned processing steps provide a basis for equivalent identification:

- Splitting of compounds: (DE) *Rotationsenergie* ↔ *Rotation+Energie* → *rotational energy*
- Relational adjectives: (FR) *source lumineuse* ↔ *source de lumière* → *source of light*

In addition, learning from the term's contexts in the source and in the target language is used to further enlarge the system's bilingual dictionary. The basic assumption underlying this

approach is that terms which are equivalent will co-occur with lexical items of their languages which are again equivalents. Thus we use a dictionary to relate the context partners of two items; pairs of items with "shared" contexts should be equivalents.

2.3. *Application scenario and architecture*

Term extraction is almost never used as a standalone tool; in TTC, it serves to feed into CAT and MT. Thus, the overall TTC usage scenario can be depicted as in figure 1: it consists of two main building blocks: an automatic one (upper part), and the interactive application used by translators (lower part). The automatic part (described above in section 2.2) provides raw material for both, CAT use and machine translation.

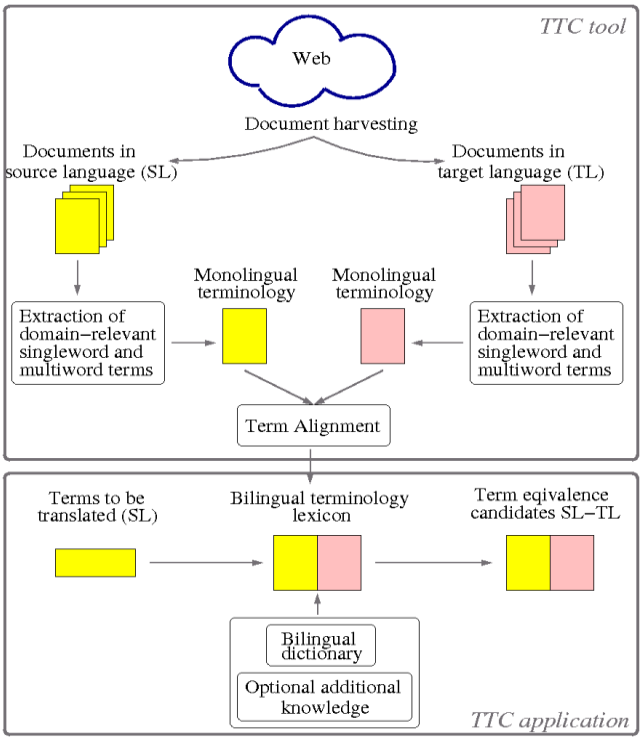

Figure 1: TTC tool chain.

3. Term extraction and its users – TTC on the map of term extraction

3.1. *User needs*

User needs with respect to term extraction differ widely, depending on further tools the users intend to work with. CAT tools and rule-based MT are in the tradition of prescriptive (or even proscriptive) terminography and specialized lexicography: a terminological data collection should only contain recommended terms, or it should mention variants and clearly proscribe those that are not "authorized". As mentioned above, pre-/proscriptive terminology work is hard to carry out in new or upcoming domains. In the TTC output, we thus give all variants found in the texts, along with frequency data and, where possible, metadata about the sources used. As base terms and variants are related in the TTC output, users can quite easily

themselves decide about prescribed and unwanted uses.

Different user groups require different amounts of terminological data: for statistical MT, all found instances of terms and their variants can be integrated. First tests with DE→ ER statistical MT on the basis of the Moses system (Koehn et al., 2007) have shown significant improvements of the SMT output as a result of the inclusion of TTC's term candidates into the system's knowledge sources.

In interactive setups (CAT), translators have different needs than, e.g., domain experts or terminologists. The former require the most compact description of term candidates possible, as they cannot spend much time and effort on the entries of the term collections they work with. Terminologists, revisers etc., on the other hand, typically prefer the most detailed description possible. While a concise translation-oriented term description of TTC just contains a headword term and (up to 5 alternative, automatically derived) equivalence candidates, full-scale entries, as offered to terminologists, include a detailed presentation of term variants (see section 4).

Obviously, to turn the TTC output into raw material for a specialized (online) dictionary, more filtering of the data would be needed, in terms of headword selection, as well as of selection and ordering of linguistic data about terms. However, given that CAT glossaries are the main focus of TTC, we need to adhere to their structures and requirements.

3.2. *State of the Art*

Although in terms of Cabré et al.'s (2001) classification, TTC is an instance of a standard hybrid (symbolic plus statistical) term extractor, the TTC approach to term extraction is characterized by a number of principles which make it different from other tools in this domain.

Firstly, TTC is designed to work with comparable texts, for the reasons discussed in section 2.1. There are no commercial tools for this functionality, and the few research tools under way mainly use less sophisticated term alignment tools.

Secondly, TTC is different from exclusively statistical systems that tend to provide much noise in their output, and from systems based on linguistic procedures only (morphology, syntax, etc.), which tend to be highly language-specific and hard to adapt to other languages than those they are constructed for (cf. Cabré et al., 2001). Linguistic knowledge for the TTC tool is typically provided as a parameter of the generic procedures, or it is learned from large corpora.

A tool that has functions similar to those of TTC is Jaguar (Nazar et al., 2008). It also covers the full pipeline from text crawling to term extraction. However, it is, as of early 2012, still mostly statistics-based, using only little linguistic knowledge. And it is only aimed at monolingual term extraction.

There are more sophisticated statistical devices for term extraction than Ahmad et al.'s (1992) w*eirdness scoring*, which is used in TTC; most such statistics did however not perform better than weirdness scoring in our tests (e.g. (Daille, 1994), (Rayson and Garside, 2000)).

4. Lexicographic and terminographic data types provided by TTC

The TTC tool provides different kinds of output. It annotates terms in the texts crawled by means of the seed words, and it provides both monolingual and bilingual term candidate lists, according to users needs.

4.1. *Linguistic properties and concept-related data*

Output term lists are relatively rich in linguistic and corpus-derived data, compared with simple bilingual glossaries. They contain the lemma and the (conventional) citation form of the term candidate, the term's category or, if it is a multi-word term, its POS sequence, the inflected forms found in the corpus, as well as the graphical, morphological or syntactic variants found in the corpus. All corpus data come with frequency annotations, to allow the user to more easily select appropriate candidates.

4.2. *Variant documentation*

Table 3 provides an example of the above mentioned properties of the TTC output.

Table 3. Example output of the TTTC extraction tools.

Feature	Value
Lemma	Windenergie *(wind energy)*
POS	N|N
Absolute frequency	1254
Relative frequency	0.025
Domain specificity	1572.44
Inflected forms	Windenergie
Citation form	Windenergie
Variant	Energie d Wind
POS	N DET N
Type	Syntactic
Synonymous	yes
Absolute frequency	4
Inflected forms	Energie des Winds, Energie des Windes

It also shows the two strategies used in the tools for distinguishing base terms from variants:

- By means of related POS patterns: one pattern is a base term pattern, the other one a variant pattern. Items following a variant pattern are only part of the output, if a corresponding item of a base term pattern is also found in the data:

Table 4. Variants of the German compound nouns.

Base	Variant	Example
N1|N2	N2 DET*gen* N1*gen*	(DE) Energieproduktion ↔ Produktion der Energie (*energy production* ↔ *production of energy*)
	N2 *von* N1	(DE) Stromerzeugung ↔ Erzeugung von Strom (*power generation* ↔ *generation of power*)

- By means of frequency counts. When the related patterns do not clearly identify one

element as a base term, we tentatively assign base term status to the most frequent one.

4.3. *Metadata*

To fully document the term candidates retrieved from texts harvested from the internet, ideally as many types of metadata as possible would need to be recorded. Given the rather limited possibilities of web documents in this respect, we so far only record the following information: publisher, language, URL, title of web page and date of crawling.

These data need to be provided for all relevant items proposed to the user. As the crawled corpus may provide many instances of a term, an effective and efficient way of abstracting over the metadata of each instance in an equivalent candidate list still needs to be found.

5. Summary and conclusion

In this paper, we have given a general description of the tools term candidates extraction being developed within TTC. The tool chain aims at automatically providing bilingual domain-specific terminology extracted from a comparable corpora. The tools are still under development, but a first UIMA - based version of TermSuite[4], which implements the full pipeline, is already available on the Web.

The screenshots in figures 2 and 3 show the output of the TermSuite. Figure 2 shows a partial list of the term candidates extracted from a French corpus on wind energy. For example, the output of the term candidate *optimisation* contains not only information about its frequency, POS, domain specificity, etc., but also about its variant *optimization* which was found in the corpus.

Figure 3 shows the results of the term alignment run on English and French monolingual term candidates extracted from the texts on wind energy. For example, the French term *parc éolien* is aligned with the English term *windpark*. Multiple translations are also possible: the alignment tool found, for example, several English translations for the French term *énergie renouvleable*, namely *renewable energy, sustainable energy, renewable power.* Additionally to the alignment, on the left side of the picture, the assessment of the tool's alignment proposals is shown, by comparison with a gold-standard term list.

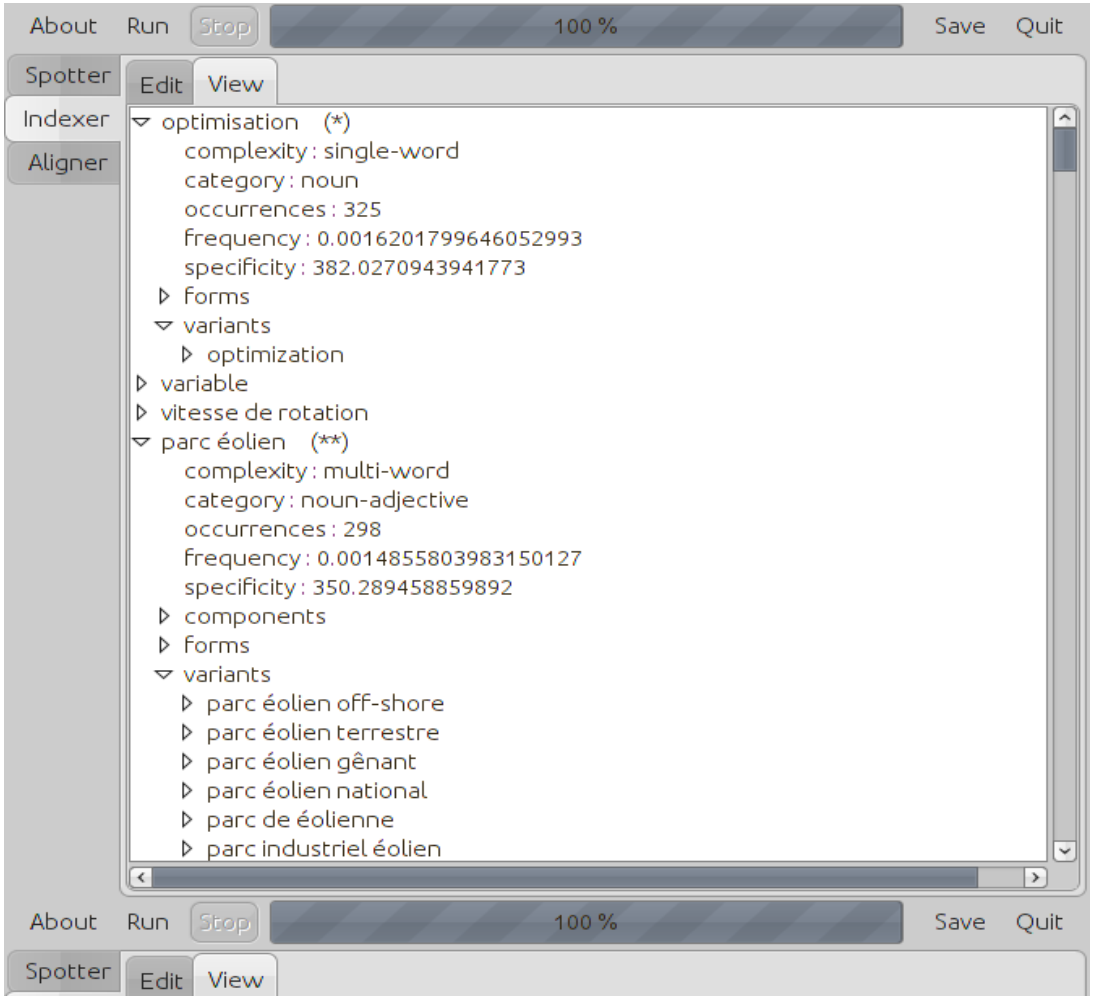

Figure 2. TermSuite: Monolingual term candidates.

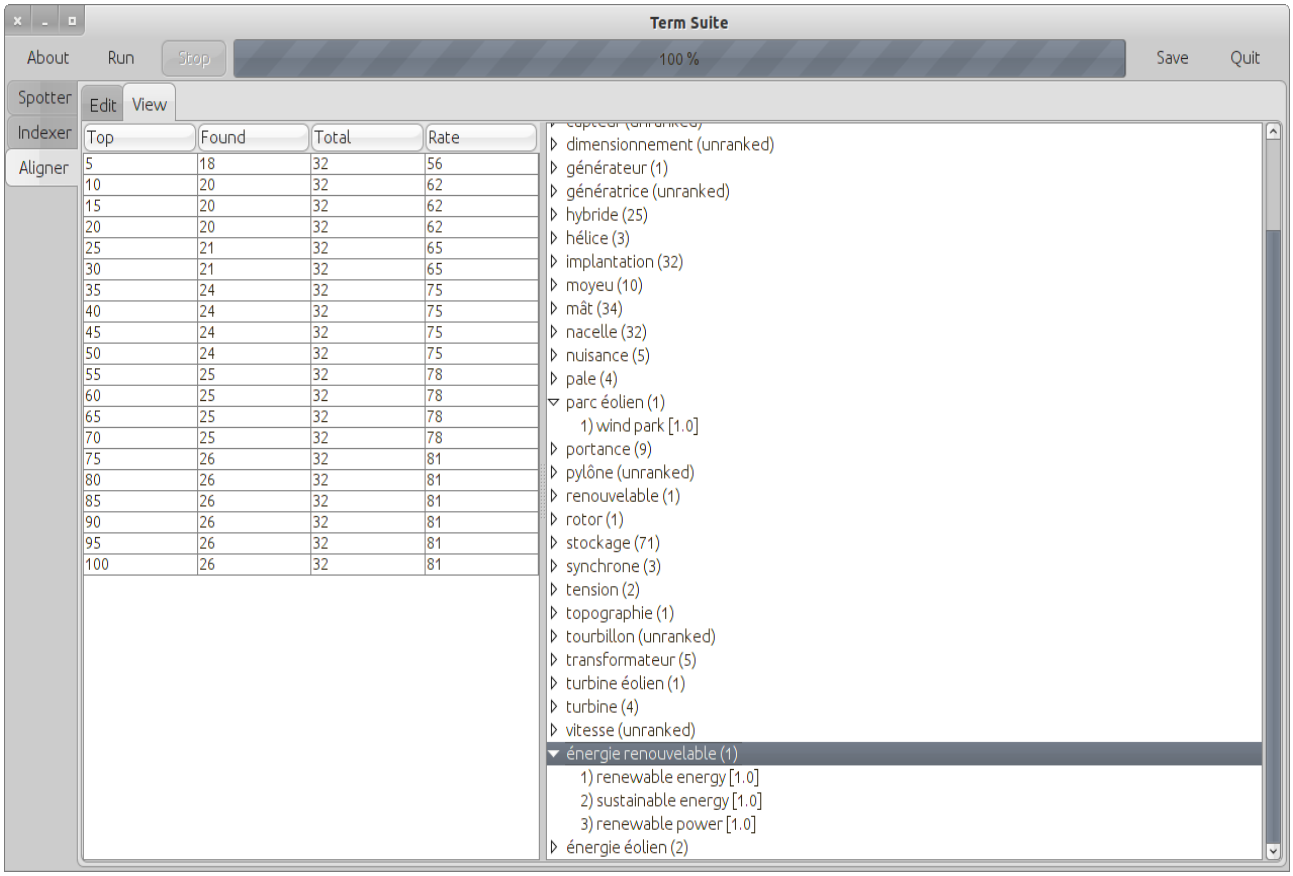

Figure 3. TermSuite: Bilingul term alignment (right) and alignment evaluation (left).

Notes

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248005.
[2] TTC started in January 2010 and will run until end 2012. TTC works with seven languages from five different typological language families: DE, EN, ES, FR, LV(=Latvian), RU, ZH.
[3] For the Latvian language there is no TreeTagger and within the project we use a proprietary tagger developed by the project partner.
[4] http://code.google.com/p/ttc-project/

References

Ahmad, K., A. Davies, H. Fulford and M. Rogers 1994. 'What is a term? The semiautomatic extraction of terms from text.' In Mary Snell-Hornby (ed.): *Translation Studies: An Interdiscipline*. Amsterdam: John Benjamins, 267–278.

Cabré Castellví, M. T., R. Estopà Bagot and J. Vivaldi Palatresi 2001. 'Automatic term detection: a review of current systems.' In D. Bourigault, C. Jacquemin and M.-C. L'Homme (eds.), *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins, 53–87.

Daille, B., E. Gaussier and J.-M. Lange 1994. 'Towards automatic extraction of monolingual and bilingual terminology.' In *Proceedings of the 15th International*

Conference on Computational Linguistics (COLING). Kyoto, Japan.

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

de Groc, C. 2011. 'Babouk: Focused web crawling for corpus compilation and automatic terminology extraction.' In *Proceeding of the International Conference on Web Intelligence 2011*. Lyon, France.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst 2007. 'Moses: Open Source Toolkit for Statistical Machine Translation.' In *Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*. Prague, Czech Republic.

Nazar, R., J. Vivaldi and M.T. Cabré Castellví 2008. 'A Suite to Compile and Analyze an LSP Corpus.' In *6th edition of the Language Resources and Evaluation Conference LREC*. Morocco

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

Schmid, H. 2005. 'Probabilistic Part-of-Speech Tagging Using Decision Trees.' In *Proceeding of the International Conference on New Methods in Language Processing*. Manchester, UK.

Schmid, H. and F. Laws 2008. 'Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging.' In *Proceedings of the 22th International Conference on Computational Linguistics*. Manchester, UK.

Weller, M., H. Blancafort, A. Gojun and U. Heid 2011. 'Terminology extraction and term variation patterns: a study of French and German data.' In *Jahrestagung der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)*. Hamburg, Germany.