FastKwic, an "intelligent" concordancer using FASTR

Véronika Lux-Pogodalla*†, Dominique Besagni*, Karën Fort*

*INIST - CNRS, Equipe SRDI
2 allée de Brabois
54500 Vandoeuvre-lès-nancy, France
{dominique.besagni, karen.fort}@inist.fr

† ATILF - CNRS
44, avenue de la Libération
54000 Nancy
veronika.lux@atilf.fr

Abstract

In this paper, we introduce the FastKwic (Key Word In Context using FASTR), a new concordancer for French and English that does not require users to learn any particular request language. Built on FASTR, it shows them not only occurrences of the searched term but also of several morphological, morpho-syntactic and syntactic variants (for example, *image enhancement, enhancement of image, enhancement of fingerprint image, image texture enhancement*). Fastkwic is freely available. It consists of two UTF-8 compliant Perl modules that depend on several external tools and resources: FASTR, TreeTagger, Flemm (for French). Licenses of theses tools and resources permitting, the FastKwic package is nevertheless self-sufficient. FastKwic first modules is for terminological resource compilation. Its input is a list of terms - as required by FASTR. FastKwic second module is for processing concordances. It relies on FASTR again for indexing the input corpus with terms and their variants. Its output is a concordancer: for each term and its variants, the context of occurrence is provided.

1. Introduction

TermSciences (Khayari et al., 2006) is a large terminological database developped at INIST-CNRS that:

- currently includes 150,000 concepts linked to more than 540,000 terms,
- merges different terminological resources (lexicons, dictionaries, thesauri) produced and maintained by various French public research institutes,
- covers various scientific domains (eg. physics, chemistry, medicine, informatics, philosophy, linguistics, sociology, etc.).

In this database, the linguistic description of the terms is often very limited. In particular, definitions are only available for a few terms. To help users get an idea about a term usage, and since INIST has a very large collection of bibliographical records (17,000,000), we decided to add a concordancer to the TermSciences search engine.
We wanted a concordancer that would show all the occurrences of a given term in a corpus of records, in the usual display (i.e. one concordance per line, centered and highlighted searched term, etc.). The concordancer had to support at least English and French.
Terms in TermSciences are mono- or multi-word terms. They occur in texts as such but also with several variations:

- typographical variations (ex. with or without "-"),
- morphological variations (ex. singular or plural),
- syntactic variations (especially insertion of a modifier, coordination, permutation).

For example, for the term *structural gene*, we would want to find contexts such as *structural erm genes* in which the term occurs with a modifier, or *structural and regulatory genes*, where it occurs within a coordination. For the term *resistance mechanism*, we would want to find contexts such as *mechanism of clarithromycin resistance*.
Christian Jacquemin (Jacquemin, 1994; Jacquemin, 1997; Jacquemin et al., 1997) studied these variations and developped FASTR, a tool that can detect them in texts, given a terminological resource. At INIST, in collaboration with Christian Jacquemin, Jean Royauté (Jacquemin and Royauté, 1994; Royauté, 1999) developped an automatic indexing platform called ILC (Infométrie Langage et Connaissance, i.e. Infometrics Language and Knowledge) that uses FASTR. His colleagues F. Ville-Ometz (Ville-Ometz et al., 2007) and N. Soccol (Soccol, 2001) studied term variations in details and used FASTR meta-rules to model these variations (see section 3.2. for examples). The FastKwic concordancer we introduce here builds on ILC and on this work on term variation. It has to be noticed that FastKwic is freely available.
This article is organized in the following way: first, section 2. provides a summary of our motivations for developping the tool, which is then described with some details in the following section. The last section is about one implementation of FastKwic for TermSciences at INIST. Classically, we conclude with some perspectives about possible improvements.

2. Background

There are many concordancers around and the work of Jacquemin is well known in the community. But we could not find any existing concordancer that take his work into account.

To our knowledge, most concordancers belong to one of the following two types:

- Concordancers that simply provide occurrences in which the surface form of the term exactly matches the term used for search. Such concordancers sometimes allow users to use some special characters such as * (where *cheva** will match *cheval*, *chevaux*, *chevalier*, etc.). Users at INIST, like most non-specialists, like the simple interface of these concordancers but they are critical about the results that "don't even include the plural forms!".

- Concordancers that provide more or less complex query interfaces where the user typically has to describe all forms of the term searched within a particular language, usually based on regular expressions. This is the case, for example, of the concordancers of Le Monde corpus[1], Frantext[2], or ConcQuest[3]. As shown by Olivier Kraiff (Kraif, 2008), request languages can be very complex, especially when the corpus has been linguistically analyzed and elements of the different levels of analysis are used in the request, such as morphosyntactic features and even syntactic dependancies. These concordancers do not suit our users: they are unwilling to invest time in learning a complex language that, in addition, is specific to one system, just to get what seems to them a simple result.

In practice also, most concordancers available on the Internet work on a particular given corpus, but are not available as a plugin for one's own corpus (ex. in English, the Brown Corpus and the BNC corpus, in French, Le Monde corpus, Frantext, Scientext). Of course this restriction also applies to some research work such as (Abbes, 2004) that do take some term variations into account (for example here morpho-syntactic variants for Arabic). The few concordancers that are freely available (for example, a limited version of Context[4]) do not allow for the detection of term variations and their technical features do not suit our needs (i.e. Linux server, large corpus composed of numerous small documents, many concordances).
These observations led us to develop a concordancer using FASTR that, in our opinion, provides a clean model of the most frequent term variations, i.e. just what naïve users expect in a concordance. Building on the work done for the indexing plateform ILC, the concordancer was first developped for our own needs at INIST, to provide an additionnal feature in the TermSciences Web interface, for French and English. We then decided to build a clean, freely downloadable tool called FastKwic[5]: this is just what we had looked for in vain and we hope it will be useful to others.

[1] http://www.bultreebank.org/french/form.htm

[2] http://www.frantext.fr/categ.htm

[3] http://w3.u-grenoble3.fr/kraif/ConcQuest/concquest.php

[4] http://sites.univ-provence.fr/veronis/logiciels/Contextes/index-fr.html

[5] http://www.cnrtl.fr/outils/

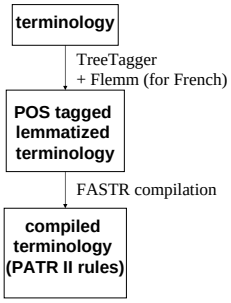

Figure 1: Resource Compilation

3. FastKwic

FastKwic consists of two UTF-8 compliant Perl modules that depend on several freely available external tools and resources, namely FASTR (Jacquemin, 1997; Jacquemin et al., 1997), TreeTagger (Schmid, 1997) and Flemm (Namer, 2000) (for French). The resources needed are those required by FASTR, i.e. a corpus and a list of terms. The two modules are run in sequence, as described below.

3.1. Resource Compilation

The first step of the sequence consists in preparing, i.e. compiling, the list of terms required by FASTR (see figure 1). This list consists in a text file with one term per line, either in English or in French, as those are the languages presently supported by FastKwic.

Example of input terms:

- *Gene amplification*
- *Periodic table*

The list is tagged using TreeTagger. For French, Flemm is applied to improve lemmatization[6]. The result is then given to FASTR for compilation to PATR II rules[7].
At this stage, FastKwic also allows for applying normalization rules, in order to be able to identify, for example, that *beta-lactamase* is a synonym of *β-lactamase*.

3.2. Indexing

The second step concerns the corpus processing, which requires the previously compiled terminological resource (see figure 2).

Example of input corpus:

- "The periodic table is a tabular arrangement of the chemical elements according to their atomic numbers".

[6] FastKwic uses a particular version of Flemm that supports UTF-8. This version should be available soon on the CNRTL Web site. In the meantime, it is included in the FastKwic installation.

[7] PATR II is a unification-based grammar formalism (Shieber, 1986)

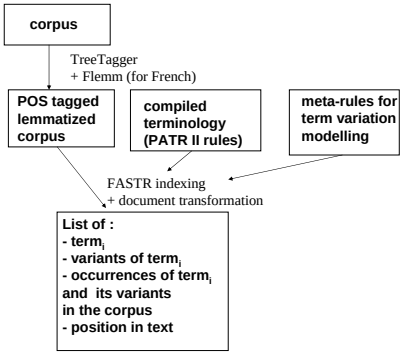

Figure 2: Indexing

```
<?xml version='1.0' encoding='UTF-8'?>
<Concordancer>
  <Term>
    <TotalNumber>2</TotalNumber>
    <Preferential>
      <String>Gene amplification </String>
      <Number>2</Number>
      <Occurrences>
        <Occurrence>
          <Reference>000007</Reference>
          <Position>1:32</Position>
          <Transform>XX,25,Perm</Transform>
          <Context><b>Amplification of the MYC gene is</b> associated with
dmi</Context>
        </Occurrence>
        <Occurrence>
          <Reference>000008</Reference>
          <Position>1:38</Position>
          <Transform>XX,15,Ins</Transform>
          <Context><b>This gene facilitated amplification of</b> a 407-bp DNA
fragme</Context>
        </Occurrence>
      </Occurrences>
    </Preferential>
  </Term>
```

Figure 3: Excerpt of the final output

- "Publication of encyclopaedias, handbooks, periodic tables in electronic form is growing".
- "Amplification of the MYC gene is associated with dmin".
- "This gene facilitated amplification of a 407-bp DNA fragment".

The corpus undergoes the same processing as the list of terms: tagging with TreeTagger, lemmatization using Flemm (for French). Then the result is given to FASTR with the compiled list of terms for indexing and a set of linguistic meta-rules developped for ILC. These meta-rules improve the produced PATR II rules (see figure 2), leading to a finer modelling of term variation.

The output is a file containing a list of terms. Each term comes with its variants and contexts of occurrence.

Thus, for the English term *system design*, several variations are accounted for:

- morphological variation, such as plural as in *systems design*,
- permutation, as in *design of system*,
- insertion of a modifier, as in *design of multi-classifier systems* or *design of the new system*,
- coordination, as in *design of system and components*.

For French, the modelled linguistic phenomena include, for example:

- the insertion of an adjectival modifier after the head noun in a Noun-Adjective term: *système de surface* → *système* ***racinaire*** *de surface*,
- the noun to verb variation that associates a nominal term with a verb phrase: ***variation*** *d'alimention* → *alimentation* ***varie***.

3.3. Results

This output of FASTR is modified and enriched in order to obtain a format that is more adapted to concordances. In particular, the results are given term by term and not sentence by sentence. An excerpt of this file is shown in figure 3. For the term *Gene amplification*, two occurrences were found in the corpus. The reference number of the sentences and the position in the sentence are given (respectively in the fields <Reference> and <Position>), as well as a short context (<Context> field). If a variation of the term is found, this is indicated in the <Transform> field. Here, two variants of *Gene amplification* were identified, one with a permutation and the other with an insertion. It is up to the user to provide access to this file, either directly or through a database, from a Web application.

3.4. Limitations

One limitation of FastKwic comes from the part-of-speech tagging step: a tagger such as TreeTagger is normally used to tag running text and not a list of terms without context. Contextual clues expected by the tagger are therefore missing in the terminology (for example, nouns with no article are very frequent). If the terminology is small enough, tagging can be corrected by hand. Alternatively, one can try to automatically improve it with some post-processing.

Another limitation appears when users work in specialized domains where particular linguistic entities (often called *named entities*) occur, such as molecular formulae or gene names. FASTR is meant to work on terms and term variants that comply to relatively standard linguistic patterns and such entities are outside its scope. If users care about such linguistic entities and their possible variants (for example, *EC 3-1-1-55* and *EC 3.1.1.55* for *acetylsalicylate deacetylase*), they have to develop their own specific process to handle these.

4. Implementation in the TermSciences Web site

4.1. Resources

The implementation at INIST had to take into account specific constraints, in particular the size of the corpus and its

composition. At INIST, around fifty information specialists in different scientific fields analyze several thousands of scientific publications each year and produce bibliographical records registered in two databases, PASCAL, for scientific domains and FRANCIS, for the humanities.
We used as corpus a selection from PASCAL: 30,744 records for French and 398,952 for English, i.e. all the records from year 2005.
Also, TermSciences, the terminological resource we used, contains more than 540,000 terms, we therefore needed to optimize the compilation process. Using the semantic tags attached to terms, we filtered out terms for which FASTR variants *a priori* seemed irrelevant. For example, geographical names, drug names, chemical compounds, etc. were excluded from the compilation process.
For these terms, indexing does not rely on FASTR but on a simpler processing called IRC3 (Royauté et al., 2003) that nevertheless allows to discover some simple graphical variants, which can prove important in certain specialized domains, for example a whitespace normalization allowing for *1,3,4-thiadiazole(2-amino)* to be recognized as well as *1 , 3 , 4 - thiadiazole (2 - amino)*. In TermSciences, the concordancer uses both FASTR and IRC3.
As for terms undergoing compilation by FASTR, i.e. terms that are neither drug names, nor chemical compounds, etc. and that usually conform to relatively standard syntactical patterns, an evalutation conducted on 1,000 of them showed that, even with some post-tagging rules, only 70% are correctly tagged. The low result is linked to using POS-tagging on a list of terms that are out of context and possibly to syntactic specificities (for example, missing articles and/or prepositions).

4.2. Performances

A variety of tests were run on the version of the system that was installed at INIST and a summary concerning the English corpus (the biggest one) is given in table 1. Note that this test was run on a 4 Intel processor 1.6 Ghz computer.

Nb English terms	117,248
Nb records (abstracts)	398,952
Corpus (abstracts) size	465.7 Mo
Compilation time	6 min.
Indexing time	66 h

Table 1: Performances for the English corpus processing

Indexing time is mainly due to FASTR, as shown in the results of an indexing experiment run on a 2 UltraSparcIII processor 750 MHz computer[8] given in table 2, concerning 20,197 records (titles and abstracts).

4.3. Result

The result of indexing is then put into a MySQL database which is accessed from the TermSciences Web interface[9].

[8]For technical reasons, we were unable to run this test on the same machine, but the proportions we give should be valid on most computers.

[9]`http://www.termsciences.fr/-/Index/Search/Concordancer/`

Total time	800 min.	100%
FASTR	775 min.	96.87%
TreeTagger time	5 min.	0.62%

Table 2: Performance details on a sample

An example is provided in the screenshot of the figure 4 for the French term *Développement affectif*.
The "Concept" area of the interface (see figure 5) shows the terms found in TermSciences in all the available languages (French, English and in some cases Spanish and German), as well as those that are related in the thesaurus (*related*, *broader*, *narrower*).
In this example, the French term *Développement affectif* is shown with:

- a synonym: *Développement psychoaffectif*,
- an English translation: *Affective Development*, *Affective development*
- a Spanish translation (that will not be used for the search): *Desarrollo afectivo*.

and with other associated terms:

- related terms: *Trouble du développement, Motricité*
- a broader term: *Psychologie*
- narrower terms: *Affect affectivité, Aptitude sociale, Sentiment, Vulnérabilité*

The concordancer area of the interface (see figure 6) shows the results of the search with the 4 selected terms (*Développement affectif*, *Développement psychoaffectif*, *Affective development*, *Affective Development*).
As can be seen, FastKwic also provides here:

- French variants with coordinations:
  - *développement social et affectif*,
  - *développement psychomoteur et affectif*
- an English variant with an insertion: *affective aspects of development*.

Each occurrence is shown with a limited context but is linked to the source document that provides a larger context. As we deal with large corpora, the concordancer interface provides two additionnal filters, one according to the scientific domain (*Display results for domain*) and the other according to the variant (*Display results for variant*). This feature can be used to analyze the different contexts or variants available for the term according to the domains.

5. Conclusion

FastKwic is being documented and made freely available from the CNRTL[10]. It will provide the community with an easy-to-integrate, easy-to-use concordancer integrating

[10]`http://www.cnrtl.fr/outils/`

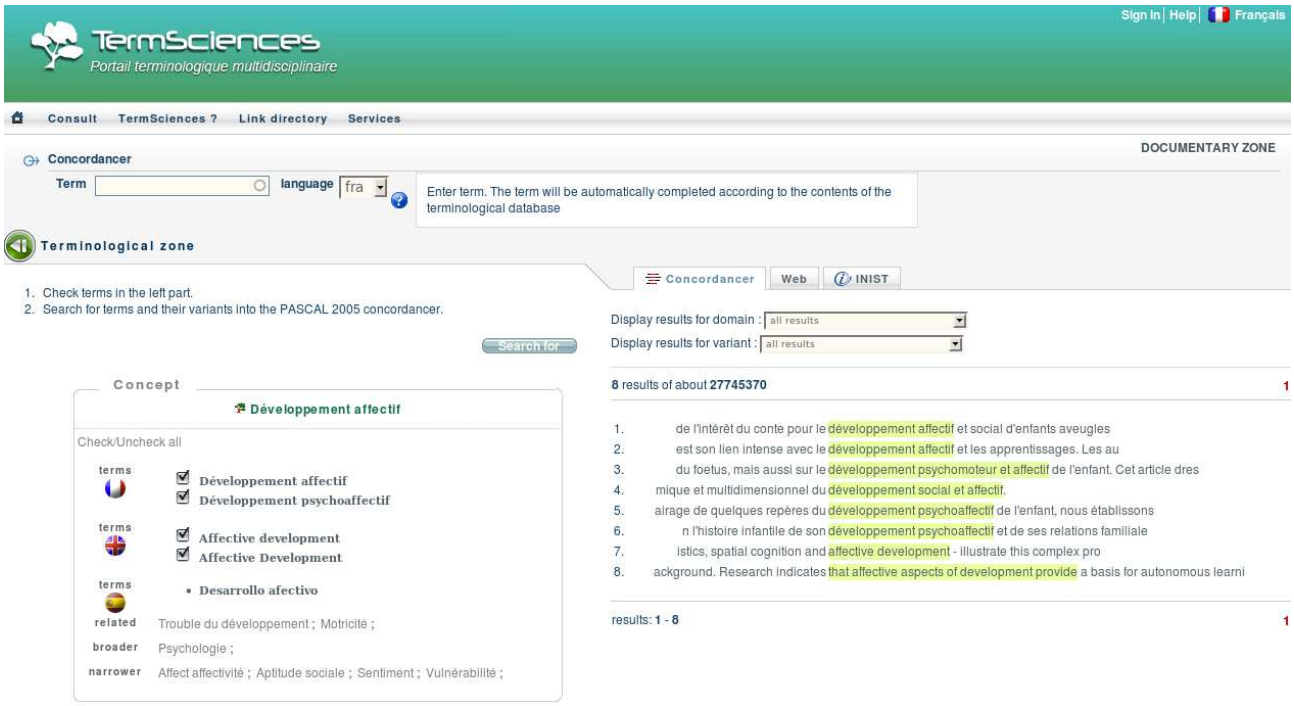

Figure 4: TermSciences Web interface

Figure 5: TermSciences Web interface (detail of the "Concept" area)

FASTR, that everybody will be able to install and feed with their own terminologies and corpora.

A short term perspective will be for us to improve the acronyms processing, as, at the moment, everything is put in lowercase before tagging and prevents their correct recognition. Another development concerns the integration of new languages, like Spanish, for which we need to model term variation.Finally, linking FastKwic to a terminology extraction tool, like ACABIT (Daille, 2003), would certainly be of great help, but the well-known output filtering problem is still a blocking issue.

6. Acknowledgments

For developments around TermSciences, we acknowledge support from the Région Lorraine via the CPER (Contrat Plan Etat Région) program. Also, a special thank goes to our colleague, Claire François, for her many suggestions on a previous version of this article, especially concerning ILC.

7. References

Ramzi Abbes. 2004. *La conception et la réalisation d'un concordancier pour l'arabe*. Ph.D. thesis, Institut National des Sciences Appliquées de Lyon, France.

Béatrice Daille. 2003. Conceptual structuring through term variations. In D. MacCarthy F. Bond, A. Korhonen

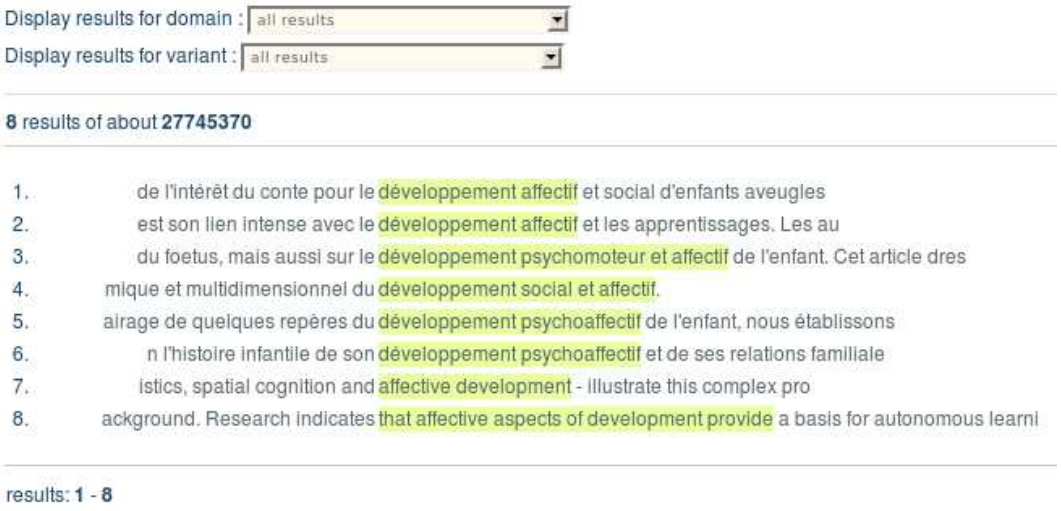

Figure 6: TermSciences Web interface (detail of the concordance results)

and A. Villacicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 9–16.

Christian Jacquemin and Jean Royauté. 1994. Retrieving terms and their variants in a lexicalized unification-based framework. In *Proceedings of the 17th Annual International ACP SIGIR Conference on Research in Information Retrieval (SIGIR'94)*, pages 132–141. Springer Verlag.

Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. 1997. Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL 1997)*, pages 24–31, Madrid, Spain. ACL.

Christian Jacquemin. 1994. Fastr: A unification grammar and a parser for terminology extraction from large corpora. In *Proceedings of IA-94*, Paris, France.

Christian Jacquemin. 1997. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches - Université de Nantes, France.

Majid Khayari, Stephane Schneider, Isabelle Kramer, and Laurent Romary. 2006. Unification of multi-lingual scientific terminological resources using the ISO 16642 standard, the TermSciences initiative. In Stefan Schulz Pierre Zweigenbaum and Patrick Ruch, editors, *Proceedings of the LREC 2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons: the Case of Biomedicine*, Genoa, Italy.

Olivier Kraif. 2008. Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest. In *Proceedings of JADT 2008*, Lyon, France.

Fiammetta Namer. 2000. Flemm : Un analyseur flexionnel du français à base de règles. *Traitement automatique des langues*, 41:2:523–547.

Jean Royauté, Claire François, Alain Zasadzinski, Dominique Besagni, Philippe Dessen, Sylvaine Le Minor, and Marie-Thérèse Maunoury. 2003. Mining corpora of texts on genes involved in thyroid cancers: a bioinformatic text mining and clustering process. In *Proceedings of the European Conference on Computational Biology, ECCB2003*, pages 77–78, Paris, France.

Jean Royauté. 1999. Les groupes nominaux complexes et leurs propriétés : application à l'analyse de l'information. Thèse de doctorat en informatique, LORIA, Université Henri Poincaré-Nancy I.

Helmut Schmid, 1997. *New Methods in Language Processing, Studies in Computational Linguistics*, chapter Probabilistic part-of-speech tagging using decision trees, pages 154–164. UCL Press, London.

Stuart M. Shieber. 1986. *An Introduction to Unification-Based Approaches to Grammar*, volume 4 of *CSLI Lecture Notes Series*. Center for the Study of Language and Information, Stanford, CA.

Nicolas Soccol. 2001. Indexation automatique: une expérience. Master's thesis, Ingénierie multilingue, INALCO, under the supervision of Jean Royauté (INIST) and Monique Slodzian (CRIM- INALCO).

Fabienne Ville-Ometz, Jean Royauté, and Alain Zasadzinski. 2007. Enhancing precision in automatic recognition and extraction of term variants with linguistic features. *Terminology*, 13:1:35–59.