# A Methodology for Developing Multilingual Resources for Terminology

**Marie-Claude L'Homme, Hee Sook Bae**

Observatoire de linguistique Sens-Texte (OLST)
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec)
H3C 3J7
{mc.lhomme, hee.sook.bae}@umontreal.ca

## Abstract

This paper presents a project that aims at building lexical resources for terminology. By *lexical resources*, we mean dictionaries that provide detailed lexico-semantic information on terms, i.e. lexical units the sense of which can be related to a special subject field. In terminology, there is a lack of such resources. The specific dictionaries we are currently developing describe basic French and Korean terms that belong to the fields of computer science and the Internet (e.g. *computer, configure, user-friendly, Web, browse, spam*). This paper presents the structure of the French and Korean articles: each component is examined and illustrated with examples. We then describe the corpus-based methodology and the different computer applications used for developing the articles. Our methodology comprises five steps: design of the corpora, selection of terms; sense distinction; definition of actantial structures and listing of semantic relations. Details on the current state of each database are also given.

## 1. Background and Motivations

This paper presents a project that aims at building lexical resources for terminology. By *lexical resources*, we mean dictionaries that provide detailed lexico-semantic information on terms, i.e. lexical units the sense of which can be related to a special subject field. In terminology, there is a lack of such resources since, typically, terminological dictionaries (and even more recent resources such as ontologies) focus on the knowledge structure of specialized subject fields, thereby ignoring important linguistic properties of terms. Furthermore, we cannot entirely rely on general resources (such as WordNet or other general dictionaries in electronic form), even if they have a good coverage of terms, since they do not always capture subtle semantic distinctions that appear in specific fields of knowledge.

The specific dictionaries we are currently developing describe basic French and Korean terms that belong to the fields of computer science and the Internet (e.g. *computer, configure, user-friendly, Web, browse, spam*). Part of the French dictionary can be accessed on the Internet (DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet: http://olst.ling.umontreal.ca/dicoinfo/). Both dictionaries take into account: a) four different parts of speech (nouns, verbs, adjectives and adverbs); b) the polysemy of terms; c) describe their actantial structure in terms of actantial roles; d) list the terms that can fill an actantial position, and, finally, list all the terms that are semantically related to a term being described. Our descriptions are based on Explanatory Combinatorial Lexicology (ECL) (Mel'čuk et al. 1984-1999, 1995).[1]

Section 2 of the paper presents the structure of the articles contained in the dictionary. Each component is examined and illustrated with French and Korean examples. In section 3, we describe the corpus-based methodology used in both languages. Section 4 gives details on the current state of each database. Finally, we will conclude with a short list of forthcoming projects.

## 2. Structure of the Dictionary

As was said above, the dictionary provides a description of various lexico-semantic properties of terms. More specifically, the articles take into account:

A) The polysemy of terms: For example, three different meanings for *adresse* (*address*) have been identified. In the dictionary, separate meanings are distinguished with a numbering system. Similarly, the meanings of the Korean form 주소 (*address*) are disambiguated.
*adresse 1*: 'address in a storage device'
*adresse 2*: 'address of a computer in a network'
*adresse 3*: 'address of a user' (e.g. an email address)'
주소 1: 기억장치 내의 주소 (adresse 1)
주소 2: 통신망에서 단말기의 주소 (adresse 2)
주소 3: 사용자의 전자메일 주소 (adresse 3)

B) The actantial structures of terms: Each separate meaning is accompanied by its actantial structure, which gives the position of actants and explains them in terms of actantial roles. In addition, linguistic realizations of actants are provided. We reproduced below the actantial structure and the the actants of the term *naviguer* (*browse*).

[1] Other details on why this framework is useful for terminology can be found in L'Homme (2002, 2003).

naviguer 1, v. intr.
AGENT navigue dans LIEU avec INSTRUMENT[2]

| AGENT | LIEU | INSTRUMENT |
|---|---|---|
| internaute 1<br>utilisateur 1 | Internet 1<br>Web 1<br>Toile 1<br>réseau 2 | navigateur 1<br>fureteur 1 |

Table 1: Linguistic realizations of actants in French

The Korean equivalent of the French term *naviguer* is *브라우징하다*. Here again (Table 2), the actantial structure and the realizations of actants are listed:

AGENT-가 LIEU-에서 INSTRUMENT-을 통해서 *브라우징하다*

| AGENT | LIEU | INSTRUMENT |
|---|---|---|
| 네티즌 1<br>사용자 1<br>이용자 1 | 인터넷 1<br>웹 1<br>인터넷망 1 | 브라우저 1<br>웹브라우저 1 |

Table 2: Linguistic realizations of actants in Korean

C) Semantically-related terms along with a formal explanation of the relation: all paradigmatic (relations within the lexicon) and syntagmatic (collocations) relationships are listed under each term being described. The listing comprises the following: the related term, a formal explanation of the relationship with a lexical function (LF) (Mel'čuk et al. 1984-1999, 1995), and a natural language (NL) explanation.[3] Table 3 gives a selection of the 78 semantic relations for the French term *navigateur* (*browser*) and some relations for the Korean term *브라우저*.

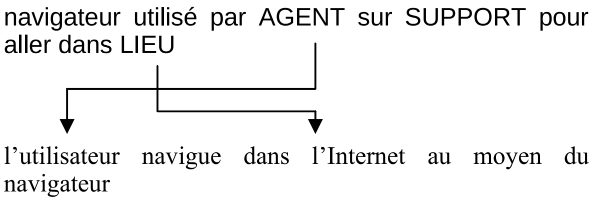
navigateur 1, n. f. : navigateur utilisé par AGENT sur SUPPORT pour aller dans LIEU

브라우저 1 : LIEU-에 연결하기 위해 SUPPORT-에서 AGENT-에 의해 사용되는 브라우저 1

| **NL explanation** | **LF** | **Related term** |
|---|---|---|
| ***Paradigmatic relationships*** | | |
| Synonyme (synonym) | Syn | fureteur 1, ~ Web, de navigation, ~ Internet<br>웹브라우저 1 |
| Générique (generic term) | Gener | logiciel 1, application 1<br>소프트웨어 1, 응용프로그램 1 |
| A le même générique (co-hyponym) | $Syn_{\cap}$ | traitement de texte 1, chiffrier 1, tableur 1<br>워드프로세서 1 |
| Partie (part) | Part | menu 1, barre 1<br>메뉴 1, |
| ***Syntagmatic relationships*** | | |
| L'agent prépare le mot clé (the agent prepares the keyword) | $Prepar_1$ | configurer 1 le ~<br>~을 구성하다 |
| L'agent fait fonctionne le mot clé (the agent causes the keyword to function) | $Caus_1Fact_0$ | lancer 1, exécuter 1 le ~<br>~을 실행하다 |
| L'agent se prépare à utiliser le mot clé (the agent gets prepared to use the keyword) | $Prepar_1Real_1$ | appeler 1 le ~<br>~을 호출하다 |
| L'agent cesse d'utiliser le mot clé (the agent stops using the keyword) | $FinReal_1$ | quitter 1 le ~, sortir du ~<br>~을 멈추다 |
| L'agent utilise le mot clé pour intervenir sur le lieu (the agent uses the keyword to do something in the location) | $Labreal_{12}$ | naviguer 1 dans / sur le lieu<br>lieu-에서<br>브라우징하다 |

Table 3: Semantic relations for the terms *navigateur* and *브라우저*.

The explanation of most semantic relations points to the actantial structure. As can be seen in Table 3, NL explanations include the actants which are involved in a collocation. Figure 1 shows the relations between the actantial structure and the collocation *naviguer dans l'Internet (au moyen du navigateur) 'browse the Internet (with a browser)'*. In this case, the agent and the location are involved.


navigateur utilisé par AGENT sur SUPPORT pour aller dans LIEU

l'utilisateur navigue dans l'Internet au moyen du navigateur

Figure 1: Actants in a collocation

In LF notations, we comply to the notation used in ECL and use numbers to indicate actantial positions ($Labreal_{12}$).

D) Finally, contexts are also provided to illustrate the functioning of terms in sentences.

## 3. Methodology

We devised a methodology which combines computer applications and human intervention based on Korean and French corpora. It is important to note that resources developed for French and Korean have not reached the

[2] It is worth pointing out that the notation of actantial structures in the DiCoInfo differs from the one that is proposed in ECL. In the latter, actants are identified with variables (e.g. X browsed Y with Z) to which no semantic content is given.

[3] In the Web version of the dictionary, LF notations are not displayed. In addition, if related terms have their own article, a link enables users to access it directly.

same stage, since the French project started in 2002, whereas the Korean project started in 2004.

Our methodology is divided into five steps, namely 1) compilation of the corpora; 2) selection of terms; 3) sense distinction; 4) definition of the actantial structure; 5) listing of semantic relationships. Each step is further described in the following subsections.

## 3.1 Corpora

The development of our resources relies heavily and at all stages on specialized corpora. The French corpus amounts to approximately one million words and contains texts published between 1996 and 2004. Texts deal with micro-computing, networks, the Internet, operating systems, hardware, software and programming. The corpus was tagged morpho-syntactically and lemmatized with TreeTagger (Schmid 1994).

The Korean domain-specific corpus comprises 4,230,000 eojeols and is subdivided into various subfields (such as hardware, software, Internet, program, operating systems). The corpora are partly comparable: the Korean corpus contains more texts on telecommunications than the French corpus; in contrast, the French corpus contains more texts dealing with the Internet. This will have an impact on the list of candidate-terms (refer to the table 4). The Korean corpus was tagged using a Korean tagger supplied by the research center KAIST.

Corpora must be updated regularly in order to provide occurrences of new terms. A first update has been performed in 2004 for the French corpus and another one is currently taking place.

## 3.2 Selection of Terms

Terms are selected using an automatic technique combined with an analysis performed by the terminologist. First, a list of candidate single-word terms is produced using an extractor called *TermoStat* developed by Drouin (2003). TermoStat compares a specialized corpus (called an *analysis corpus*) to a general corpus (referred to as the *reference corpus*) and ranks candidates according to their specificity. The French domain-specific corpus has been compared to the corpus *Le Monde* (30 million words); the Korean corpus was compared to a general corpus of 40,000,000 eojeols composed of different types of texts on literature, society, health, culture, history, etc. Table 4 presents the ten most specific terms identified by TermoStat in French in Korean.

TermoStat can identify both single-word and multi-word terms, but in this project, only single-words terms were identified. This decision is justified by the approach taken to compile the dictionary. Since the focus is on semantic properties of terms and not the knowledge organization of specialized fields, the only multi-word units analyzed do not have a compositional meaning (e.g. *système d'exploitation 'operating system', traitement de texte 'word processor*').

| **French** | **Korean** |
|---|---|
| *fichier (file)* | *사용자 (user)* |
| *commande (command)* | *시스템 (system)* |
| *Internet (Internet)* | *프로세스 (process)* |
| *serveur (server)* | *메시지 (message)* |
| *utiliser (use)* | *파일 (file)* |
| *utilisateur (user)* | *검색 (search)* |
| *logiciel (application or software)* | *경로 (directory)* |
| *option (option)* | *호스트 (host)* |
| *ordinateur (computer)* | *웹 (Web)* |
| *système (system)* | *소프트웨어 (software)* |

Table 4: Ten most specific terms in French and Korean

Once terms have been extracted automatically, terminologists analyze the lists produced to retain those terms that will appear in the dictionaries. To assist them in this task, they use four different lexico-semantic criteria:

a) Lexical units selected should refer to entities related to computing: hardware (e.g. *carte 'board', ordinateur 'computer'; 시스템 'system'*), software (e.g. *compilateur 'compiler', programme 'program'*), representational entities (e.g. *fichier 'file', données 'data'; 데이터 'data'*), units of measure (e.g. *octet 'byte', mégahertz 'megahertz'*), animates (e.g. *pirate 'hacker', programmeur 'programmer'; 사용자 'user'*).

b) If lexical units are predicative — verbs, nominalizations, adjectives, etc. — they are selected if their actants are entities accepted according to criterion a. (e.g. *l'utilisateur charge un programme en mémoire 'the user loads the program into the memory'*). However, the lexical unit must convey a specific meaning when combined with specialized actants (e.g. *specify* will not be considered a term since it conveys the same meaning with a wide variety of actants (specialized and non-specialized)).
This criterion can also be used to identify Korean terms. For example, *접근 'access'* is considered to be a valid term since it is used with its arguments *사용자 'user'* and *네트워크* '*network*' that convey the meanings specific to this domain (e.g. *사용자의 네트워크 접근이 용이하지 않은 상태이다. 'access to the network by the user is difficult*').

c) If the lexical unit is a derivative, it is selected if it is semantically related to a term selected according to criteria a. or b. (e.g. *bogue: déboguer, dégogage 'bug: debug, debigging*'). In the Korean corpus, we found series such as *입력, 접근, 사용자: 입력 'input') / 입력하다 'to input' / 입력키 'input key*', *접근하다 'access'/ 접근성 'accessibility') / 접근불가 'is* *not accessible', 사용자 'user', 사용 'use', 사용하다 'to use', 사용성 'usability'.*

d) Any lexical unit sharing a paradigmatic relationship with a term selected according to criteria a., b. or c. is selected. For example, if *serveur 'server'* and *ordinateur 'computer'* are selected according to a., then *client 'client'* and *portable 'portable computer'* must be selected. Similarly, if *coller 'paste'* is selected according to b., then *couper 'cut'* and *copier 'copy'* must be considered. In Korean, this criterion helped select terms such as *입력 'input'* and its antonym *출력 'output'. Similarly, 접속 'connection*' and its near synonym *접근 'access'* were considered according to this criterion.

Once applied to the list of candidate-terms extracted by TermoStat, evaluations of the French output (Lemay et al. 2005; L'Homme 2005) have shown that precision is fair (approx. 50 %). In Korean, precision was evaluated at approximately 50% as well.

### 3.3 Sense Distinction

Once, terms are selected, each lexical form is treated separately in order to distinguish different senses some of them convey. This step is carried out manually, again by applying lexico-semantic criteria:

a) Compatible and differential cooccurrence (Mel'čuk et al. 1995: 64-65): When cooccurrents can be combined and produce an acceptable sentence, a single meaning is identified. In contrast, when cooccurrents are combined and produce an unacceptable sentence, different meanings are identified. For exemple, the verb *exécuter 'execute'* can be found with *installation 'installation'*. It can also be found with *logiciel* '*application'*. However, both cooccurrents cannot be combined (e.g. **exécuter une installation et un logiciel* '**execute an application and an installation'*). In Korean, this criterion allows us to disambiguate the senses of *저장하다 'save'*. The verb can be combined with *파일 'file'* and *디스크 'disk'*. It can also be found with *값 'value'* and *변수 'variable'*. However, when these cooccurrents are combined, the sentence is not acceptable in Korean (e.g. **이 파일과 이 값을 변수 X 에 저장하다 '*Save this file and this value in the variable X*').

b) Synonymy: If a synonym can be substituted in a first set of occurrences, but not in a second set, then two different meanings are identified. This criterion helped confirm that *exécuter 'execute'* is polysemic. *Exécuter une tâche 'execute a task'* can be replaced by *accomplir une tâche 'accomplish a task'*. However, in *exécuter un logiciel*, the replacement is no longer possible. Similarly, in Korean, *파일을 삭제하다 'delete the file'* can be replaced by *파일을 지우다 'erase the file'*. However, in *삭제 프로그램을 이용해서 소프트웨어를 삭제하다 'uninstall the application using an uninstall program'*, *삭제하다 'delete'* cannot be replaced by *지우다 'erase'*. So, two different meanings can be identified for *삭제하다*.

c) Differential derivation: When lexical units can be linked to different series of derivatives, then separate meanings are identified. This criterion will help validate two different meanings for the verb *programmer 'to program'*. The first meaning can be linked to derivatives such as *programmable 'programmable'* and *rreprogrammer 'reprogram* (e.g. *mémoire programmable 'programmable memory', reprogrammer la mémoire 'reprogram the memory'*). A second meaning cannot be linked to these derivatives (e.g. *programmer une application 'program an application', application programmable* **'***programmable application'*).

   In Korean, this criterion is of limited usefulness since few derivatives can be observed. However, by grouping series of compound terms, we can obtain valuable information. For example, two groups of compound terms, in which this term is comprised as a constituent, are found:

   *삭제 가능 광디스크 'erasable optical storage' / 널 삭제 'null suppression' 자외선 삭제 가능 프롬 'ultraviolet erasable PROM', etc.*

   *프로그램삭제 'unistall'/ 프로그램추가삭제 ' install-uninstall'*

d) Other paradigmatic relationships: This criterion is similar to the previous one. However, it applies to lexical units that are not morphologically related. When lexical units can be linked to different series of semantically related terms, then different meanings are identified. For example, *page* can be linked to (at least) two different series of lexical units: 1. *page: Web, link, portal, address*; 2. *page: document, page down, page up*. The same can be said about the noun *address*: 1. *address: memory*; 2. *address: URL, Web site*. In Korean, *사용자 'user'* has two different series of semantically related terms: *사용자 1 'user': 그룹 'group', 인터넷 'internet', 컴퓨터 'computer', 인증 'authentification'; 사용자 2 'end-user': 개발자 'developer', 제공자 'provider', 응용프로그램 'application'*.

### 3.4 Definition of the Actantial Structure

The actantial structure is described for each sense. Actants that participate in the meaning of each term are generalized from the observation of concordances in corpora, then listed and described in terms of semantic roles (currently, 14 semantic roles are used in the descriptions). For example, the two senses of *installer* are described thus:

> *Installer 1: AGENT installe PATIENT sur SUPPORT (e.g. a user installs an application on a computer)*
> *Installer 2: AGENT installe PATIENT (e.g. a user installs a printer)*

Then, the linguistic realizations of actants are listed according to the observations made in the corpus. For example, patients for *installer 1* are *application ‘application’, navigateur ‘browser’, ‘traitement de texte ‘word processor’, pilote ‘driver’* (also, refer to Tables 1 and 2).

In principle, the description of actantial structures is the same in both languages. However, in Korean, case markers are added to each actantial role. Since the order of words is more or less free in a sentence, case markers help clarify the actantial structure. Also, in Korean, the order in which actants are given reflects what can be observed in concordances. For example, we could describe the actantial structure of *설치/하다 1 ‘install’* thus: <Destination, Agent and Patient>, <Agent, Destination and Patient> or <Agent, Patient and Destination>. However, in the dictionary, the latter was chosen because this order was found more frequently in the corpus. We reproduced below some actantial structures listed in the Korean dictionary.

> *설치/하다 1 ‘install’*: AGENT-이(subjective case) PATIENT-을(objective case) DESTINATION-에(locative case) *설치/하다 1*
> *설치/하다 2 ‘install’*: AGENT-이(subjective case) PATIENT-을(objective case) *설치/하다 2*

Paradoxically, this part of the work is still carried out manually. However, a syntactic parser could certainly be used to identify linguistic realizations of actants.

### 3.5 Listing of Semantic Relationships

The last step consists in listing, in each article, all the other terms with which the head word has a semantic relationship, and providing a systematic explanation for the relationship. As was said above, this listing takes into account paradigmatic as well as syntagmatic relationships (i.e. collocations) and the framework used to guide terminologists are LFs. Hence, in this project, LFs are not viewed simply as a means to encode semantic relationships, they represent a coherent system upon which terminologists rely to find potential relevant semantic relationships in corpora. There are approximately 60 LFs which can be combined to capture complex meanings.

Most of the work is carried out by terminologists themselves. Relationships are discovered by looking at the behaviour of terms in the corpus and at descriptions of other terms. However, this work can be assisted by automatic procedures. Two specific tasks have been automated partly for finding relations between French terms.

A first method has been developed to identify specific pairs of collocations, i.e. verb-noun pairs in which verbs convey a meaning of realization (e.g. *naviguer dans l’Internet ‘browse the Internet’; appuyer sur une touche ‘hit a key’; traiter des données ‘process data’*) (Claveau and L’Homme 2006), taking into account the syntactic position of nouns (subject, object or other complement). A second method was devised to identify morphologically related pairs of terms and the semantic relations shared by the terms in the pair (Claveau and L’Homme 2005). In both cases, the applications used the information in the dictionary (i.e. the description of the relationship shared by pairs that had already been encoded) to identify new valid pairs in the corpus.

Other methods are currently being investigated in order to find other semantically related pairs, i.e. causal senses and antonyms, with the use of linguistic markers.

## 4. Current State of the Dictionaries

The different parts of the descriptions are stored in a relational Access database. Once finished and revised, articles are exported in a MySQL database and displayed in the Internet. As was said above, for the time being, only French articles can be accessed through the Web.

The French dictionary currently contains 1810 articles: 606 articles are completed; and the remainder are being written. Our sense distinction method has led to the identification of 1810 senses for 1457 lexical forms.

| | French | Korean |
|---|---|---|
| Lexical forms selected | 1457 | 100 analyzed (out of 946 previously selected) |
| Senses identified | 1810 | 168 |
| Ratio sense/form | 1.24228 | 1.68 |
| Parts of speech | | |
| • Nouns | 1,130 | 145 |
| • Verbs | 385 | 21 |
| • Adjectives | 267 | 2 |
| • Adverbs | 10 | 0 |
| • Phrases | 18 | 0 |
| Articles completed | 606 | 168 (nearly completed) |
| Articles to be completed | 1204 | --- |
| Total number of semantic relationships described | 20,100 | 1352 |
| Number of relationships listed under articles that are completed | 13,500 | --- |
| Ratio semantic relationships / completed articles | 22.27 | --- |

Table 5: Current state of the dictionaries

In the 606 articles which are completed, approx. 13,500 semantic relationships are listed and fully described with lexical functions, on the one hand, and a natural language explanation, on the other. Hence, on average, 22.27 semantic relationships appear in each article. The total of semantic relationships described amounts to 20,100.

In Korean, the first 100 candidates among 1022 terms correspond to 168 different senses (1.68). In Korean, 1352 related terms are listed under the first 100 terms (168 meanings) that have been described.

Table 5 summarizes the figures given in this section and provides other details on the descriptions. In Korean module, we have been working on the first 100 lexical forms (168 articles) and are just beginning to analyze the following 514 lexical forms, so that we cannot provide all the information given for French.

## 5. Future Work

Presently, most of the work carried out aims at completing the articles which are still under construction. In French, this represents approximately 1,800 articles. Based on the ratio obtained for the first 606 articles we have completed, we expect to find an additional 30,000 list of relationships. We would also like to extend the coverage of the dictionary to other terms, perhaps add more technical terms than those that have been included up to now.

Regarding the Korean dictionary, we would like to complete the descriptions of the selected terms. Also, since the listing of semantic relationships is not as developed as in the French dictionary, we would like to enrich this component. Secondly, the phenomena and the difficulties found in the process should be systemized in order to display rules regarding the similarities and differences between the two languages. For example, we are interested in the relation between French derivational morphemes and Korean corresponding forms. This kind of systematization will us allow to implement efficient bilingual systems.

Also, two extensions of this work are currently being developed (for English and Spanish) using the methodology described in this paper.

## Acknowledgements

This research work was financially supported by the Quebec organization Fonds québécois de la recherche sur la Société et la culture (FQRSC).

## References

Claveau, V. & M.C. L'Homme (2005). Terminology by Analogy-Based Machine Learning, In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering (TKE 2005)*, Copenhagen (Denmark), 301-312.

Claveau, V. & M.C., L'Homme (2006, forthcoming). Discovering and Organizing Noun-Verb Collocations. In *Specialized Corpora Using Inductive Logic Programming. International Journal of Corpus Linguistics*, 11(2).

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology* 9(1), 99-117.

Lemay, C., M.C. L'Homme & P. Drouin (2005). Two Methods for Extracting "Specific" Single-word Terms from Specialized Corpora: Experimentation and Evaluation, *International Journal of Corpus Linguistics* 10(2), 227-255.

L'Homme, M.C. (2002). Fonctions lexicales pour représenter les relations sémantiques entre termes. *Traitement automatique des langues (TAL)*, 43(1), 19-41.

L'Homme, M.C. (2003). Capturing the Lexical Structure in Special Subject Fields with Verbs and Verbal Derivatives: A model for specialized lexicography. *International Journal of Lexicography*, 16(4), 403-422.

L'Homme, M.C. (2004). *La terminologie : principes et techniques*. Montréal : Les Presses de l'Université de Montréal.

Mel'čuk, I. *et al.* (1984-1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, 1-IV, Montréal: Les Presses de l'Université de Montréal.

Mel'čuk, I., A. Clas & A. Polguère (1995). *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve (Belgique): Duculot / Aupelf - UREF.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, *Proceedings of the Conference on New Methods in Language Processing*, 44-49.