Using a specialized resource to enrich a general language dictionary

Marie-Claude Demers

Marie-Claude Demers has a Bachelor's degree in German studies and a Master degree in Translation from Université de Montréal. During the course of her studies, she contributed to enriching the English version of the DiCoInfo at the Observatoire de linguistique Sens-Texte (OLST). She now works as a translator for a Canadian retailer and pursues translation as a freelancer.
marie-claude.demers.2@umontreal.ca

Abstract
This paper describes a method for comparing a specialized lexicographic resource and a general one, thus evaluating the extent to which the former can contribute to increase the coverage of the latter. Concretely, it compares the contents of the English wordlist of the *Dictionnaire fondamental de l'informatique et de l'Internet* (DiCoInfo), developed at the Observatoire de linguistique Sens-Texte (OLST), and the wordlist of the *Random House Kernerman Webster's College Dictionary* (RHKWCD), of K Dictionaries. Firstly, the entries from both resources were automatically extracted and compared. Then, we carried a manual analysis of every lexical item that we classified in different categories according to their presence in RHKWCD and the way they are described in it. Based on this research, recommendations regarding ways of improving the integration of specialized units in a general language dictionary were made. Overall, this paper concludes that both lexical resources are compatible and that it is possible to incorporate information recorded in a specialized resource into a general one. (Parrallel research to extract lexical units from RHKWCD and record them in DiCoInfo has demonstrated that the reverse is possible asd well.)

Keywords: terminology, lexicography, general language, computing language

1. Introduction

Over the last three decades, computational linguistics has evolved constantly, providing an increasing number of tools—such as term extractors, database management software, concordancers, and translation memories—that accelerate, automate, and ease the work of linguists, terminographers, lexicographers, and translators. There is undoubtedly an infinite amount of data compiled by organizations, companies, institutions, and individuals, creating the possibility of sharing research findings and information. This issue is at the core of this project on bilingual lexicography that studies the compatibility between lexicography and terminology, examined by different researchers such as Cabré (2007) and Béjoint (2007). More precisely, by focusing on the integration of terms in a general language dictionary, we compare the wordlist of the *Dictionnaire fondamental de l'informatique et de l'Internet* (DiCoInfo), developed at the Observatoire de linguistique Sens-Texte (OLST), with that of the *Random House Kernerman Webster's College Dictionary* (RHKWCD)[1], of K Dictionaries. We begin by presenting each resource. Then, we explain the different steps taken to extract, analyze and select relevant data from these dictionaries. Finally, we conclude with some observations and recommendations as to how information found in a specialized resource can be incorporated correctly in a general language dictionary.

2. Presentation of DiCoInfo and RHKWCD

RHKWCD was originally published in 1947 under the name of *American College Dictionary* (Demers et al. 2012). The dictionary was revised and updated annually, was eventually retitled *Random House Webster's College Dictionary*, and K Dictionaries acquired the last version published in 2005. RHKWCD is intended for college students and the general public, native English speakers and advanced non-native users. Today, it comprises approximately 130,000 words and expressions from all language ranges. Common meanings are ordered before specialized ones and frequent units appear before older ones. The entries include pronunciation, definitions, and examples of usage, as well as information on etymology and usage. Although RHKWCD and the DiCoInfo compile different types of information, they share a common encoding system both using a markup language to record data.

The DiCoInfo is a specialized dictionary created by the Observatoire de Linguistique Sens-Texte at the Université de Montréal. It is a free online resource, focusing on terms related to the fields of computing and the Internet. Its objectives are to describe fundamental terms, such as *email*, *bug* and *network*, as well as to list and explain the relations between the terms of the field. When compiling the entries, terminographers refer

to a corpus that has more than a million words containing mainly pedagogical texts dealing with topics such as the Internet, networks, programming, micro-computing, and operating systems.

The records in the DiCoInfo are divided into sections. The sections *headword*, *part of speech*, *status*, *actantial structure*, *written by,* and *last update* appear in every entry. The sections *synonym(s)*, *linguistic realization of actants*, *contexts*, *variant(s),* and *French* are shown by default, but only in records for which the information is available. The section *definition* is only provided for records of status 0. Figure 1 illustrates how records are written in the DiCoInfo.

Based on the Explanatory Combinatorial Lexicology method (Mel'čuk 1999), this resource is still under construction and enriched on an ongoing basis. Some records are complete and available online, while others appear in the wordlist but still need to be compiled (presenting only a few contexts and including no actantial structure). The achievement level of records is indicated by a status number that ranges from 0 to 3. Completed records are attributed the number 0.

3. Extraction, analysis and integration of data

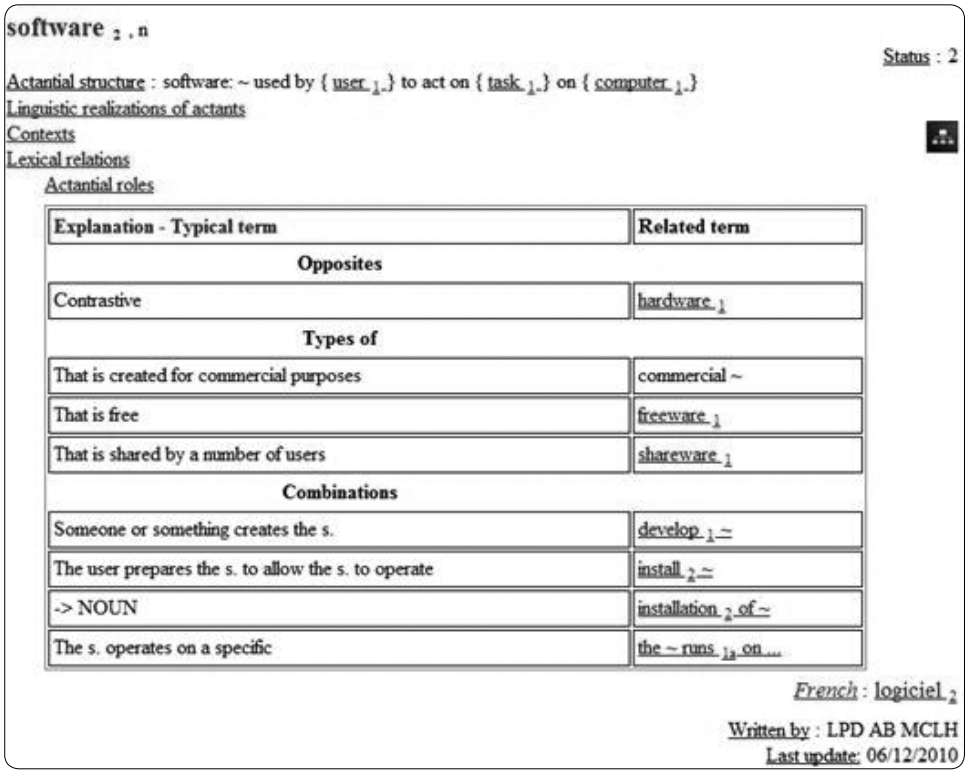
We started the project by automatically extracting all entries from both dictionaries and comparing them. This was facilitated by the fact that both resources are encoded in XML. The first step consisted of identifying lexical units that were not recorded in RHKWCD. Since many items are polysemous, we also had to carry out a manual analysis of every meaning defined under each dictionary entry of RHKWCD. The entries of the DiCoInfo were classified in one of the following categories depending on how they were taken into account in RHKWCD. We give below (Tables 1 to 12) the description and an example of a lexical unit for each of the six categories.

Table 1. Category A1

Category	A1
Criterion	A lexical item that is listed in RHKWCD with a clear indication that it belongs to the field of computing, such as a usage label or the presence of the word *computer* in its definition.
Quantity	302

Table 2. Example of a lexical unit from category A1: *batch1*

Lex. Unit	batch1
Part of sp.	noun
Definitions in RHKWCD	1. a quantity or number coming at one time or taken together; group; lot: *a batch of prisoners*. 2. the quantity of bread, dough, etc., made at one baking: *a batch of cookies*. 3. the quantity of material prepared or required for one operation: *to mix a batch of concrete*. 4. a group of jobs, data, programs, or commands treated as a unit for computer processing. 5. a. a quantity of raw materials mixed in proper proportions and prepared for fusion into glass. b. the material so mixed.



software 2 , n

Status : 2

Actantial structure : software: ~ used by { user 1 } to act on { task 1 } on { computer 1 }
Linguistic realizations of actants
Contexts
Lexical relations
Actantial roles

Explanation - Typical term	Related term
Opposites	
Contrastive	hardware 1
Types of	
That is created for commercial purposes	commercial ~
That is free	freeware 1
That is shared by a number of users	shareware 1
Combinations	
Someone or something creates the s.	develop 1 ~
The user prepares the s. to allow the s. to operate	install 2 ~
-> NOUN	installation 2 of ~
The s. operates on a specific	the ~ runs 1,3 on ...

French : logiciel 2

Written by : LPD AB MCLH
Last update: 06/12/2010

Figure 1: Record of SOFTWARE$_2$

Table 3. Category A2

Category	A2
Criterion	A lexical item that appears in RHKWCD with a meaning that clearly belongs to computing but without any clear indication, such as a usage label or the presence of the word *computer* in its definition.
Quantity	43

Table 4. Example of a lexical unit from category A2: *bot1*

Lex. unit	bot1
Part of sp.	noun
Definitions in RHKWCD	1. the larva of a botfly. 2. a device or piece of software that can execute commands or perform routine tasks, as electronic searches, usually without user intervention (often used in combination): *intelligent infobots*; *shopping bots*.

Table 5. Category B

Category	B
Criterion	A lexical item that is listed in RHKWCD, but the computing meaning is not recorded.
Quantity	273

Table 6. Example of a lexical unit from category B: *script1*

Lex. unit	script1
Part of sp.	noun
Definitions in RHKWCD	1. the letters or characters used in writing by hand; handwriting. 2. a manuscript or document. 3. the written text of a play, motion picture, television program, or the like. 4. any system of writing. 5. Print. a type imitating handwriting. 6. a plan.

Table 7. Category B-C

Category	B-C
Criterion	A lexical item that belongs to both categories B and C.
Quantity	193

Table 8. Example of a lexical unit from category B-C: *partition1*

Lex. unit	partition1
Part of sp.	noun
Definitions in RHKWCD	1. a division into or distribution in portions or shares. 2. a separation, as of two or more things. 3. something that separates or divides. 4. a part, division, or section. 5. an interior wall or barrier dividing space into separate areas. 6. Logic. the separation of a whole into its integral parts. 7. Math. a mode of separating a positive whole number into a sum of positive whole numbers.

Table 9. Category C

Category	C
Criterion	A lexical item that is recorded in RHKWCD whose general language meaning is applicable to its meaning in the field of computing.
Quantity	111

Following this analysis, two types of lexical units were considered for inclusion into RHKWCD: units absent from RHKWCD and units that are present in RHKWCD but that do not convey a meaning related to the field of computing. Therefore, we selected units labeled B and D. After having selected which units could be added, we had to determine how these could be integrated based on their presence in RHKWCD and how they are described in this resource.[2]

4. Observations and recommendations

A few scenarios were identified based on the analysis in section 2. We made recommendations for the inclusion of lexical units depending on the category in which they were classified.

When a term is used exclusively in a specialized context, L'Homme and Polguère (2008) recommend adding a label indicating the field to which it belongs. Lexical units from category D, which belong exclusively to the field of computing, should be accompanied by such a label. However, if the field is clearly indicated in the definition, for example when the word *computer* is mentioned, as in cases A1, the label becomes superfluous (Josselin-Leray and Roberts 2004).

Although lexical units B appear in RHKWCD, no meaning from the field of computing is recorded. To integrate these lexical units into the dictionary, lexicographers could simply add a new meaning.

Lexical units in groups B-C and C are listed in RHKWCD and their definition could also apply to computing. In many cases, in addition to conveying a general meaning, those units also cover a terminological usage. To illustrate how these lexical units are used in the field of computing, sentences from that domain may be added in the form of examples after the general language definition.

Figure 2 presents an example of how this latter scenario applies. In the field of computing as well as in general language, the verb *decipher* means "to decode a message". In computing, that message is in an electronic format and is decoded with a key. This figure shows the three meanings of *decipher* as listed in RHKWCD. We added an example (in boldface) to illustrate the usage of the verb in computing.

Three criteria motivate the inclusion of specialized units in general language dictionaries: the level of specialization of the term, the nature of the term (single-word or multi-word unit) and morphological relations between lexical units (Josselin-Leray and Roberts 2004). First, lexicographers prefer less specialized terms than specialized ones since the former

de•ci•pher/dɪˈsaɪ fər/ *v.t.*
1. to make out the meaning of (something obscure or difficult to read or understand): *I couldn't decipher his handwriting.*
2. to interpret by the use of a key, as something written in cipher: *to decipher a secret message, to decipher data with an algorithm.*
3. Obs. to depict; portray.
[1520–30; MF *déchiffrer*]

Figure 2: *decipher* recorded in RHKWCD

are more likely to be relevant for a vast audience. Furthermore, they prefer single terms over multi-word units. Lastly, they not only consider the meaning of the unit but also its formal resemblance to other units and would rather work on a group of units than on individual units.

5. Conclusion

In this study, we evaluated whether it is possible to use a specialized lexical resource, the DiCoInfo, to enrich a general language dictionary, namely RHKWCD. We compared the wordlist of the DiCoInfo with that of RHKWCD. We then proceeded by classifying each lexical unit found in the DiCoInfo into six different categories according to the way they were taken into account in RHWKCD. Lexical units in categories A1 and A2 were described in RHKWCD and it was obvious that they belonged to the field of computing. Only lexical units B and D were considered for inclusion into RHWKCD. The former were present in the dictionary but a computing meaning had to be added, while the latter were completely absent from it. For lexical units in B-C and C, examples could be added after the definitions to show that the lexical units are used in the field of computing, as demonstrated by the term *decipher*. We thus showed that it is possible to use an existing specialized resource to increase the coverage of a general language dictionary. We also provided a few guidelines on how to proceed based on the presence and the type of definition of units in the general language dictionary.

According to L'Homme and Polguère (2008), lexical units should be selected based on the target audience of the resource. However "different users require different things from their dictionaries, but even where dictionaries set out to address similar userships, there are discrepancies between the levels of information and kinds of detail for scientific and technical words or meanings." (Moon 2008) How can we determine which units are relevant to the target audience and which information should be added in the definition? To answer this question, we suggest selecting the units that are to be added to the general language resource based on their occurrences in a general language corpus. This allows us to objectively determine which specialized unit is now part of the general language vocabulary. For instance, in this study, we could have selected which computing lexical units labeled as B and D are part of the general language and should be added to RHKWCD based on their level of occurrence in a general language corpus. The following step would be to decide which information the definitions of those computing lexical units should provide.

Table 10. Example of a lexical unit from category C: *decipher1*

Lex. unit	decipher1
Part of sp.	transitive verb
Definitions in RHKWCD	1. to make out the meaning of (something obscure or difficult to read or understand): *I couldn't decipher his handwriting*. 2. to interpret by the use of a key, as something written in cipher: to decipher a secret message. 3. Obs. to depict; portray.

Table 11. Category D

Category	D
Criterion	A lexical item that is not listed in RHKW*CD*.
Quantity	421

Table 12. Example of a lexical unit from category D: *server machine1*

Lex. unit	server machine1
Part of sp.	noun
Definitions in RHKWCD	Not applicable.

Notes

[1] RHKWCD consists of the core of *Random House Webster's College Dictionary* (Random House, New York, 2005) updated by K Dictionaries.

[2] A total of 1,353 lexical units were analyzed, although DiCoInfo does not contain as many records. This occurred because many lexical units are synonyms or variants, and many terms have a compositional meaning. Multi-word terms containing the word *internet*, such as *internet site*, *internet access*, *internet browser* and *internet network*, can be quoted as examples. Thus, although 273 lexical units pertained to category B and 421 units were labeled as D, not as many units could be added to RHKWCD. By excluding variants, the number of potential entries that could be added to RHKWCD decreased.

References

Béjoint, H. 2007. Nouvelle lexicographie et nouvelle terminologie : convergences et divergences. In L'Homme, M.-C. and Vandaele, S. (eds.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes*. Ottawa: Les Presses de l'Université d'Ottawa, 29-78.

Cabré, T. 2007. La terminologie : une discipline en évolution : le passé, le présent et quelques éléments prospectifs. In L'Homme, M.-C. and Vandaele, S. (eds.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes*. Ottawa: Les Presses de l'Université d'Ottawa, 79-109.

Demers, M.-C., Kernerman, I. and L'Homme, M.-C. 2012. Lexicographic interchange between a specialized and a general language dictionary. In *Euralex 2012 Proceedings*.

DiCoInfo: Le dictionnaire fondamental d'informatique et de l'Internet. 2011. http://olst.ling.umontreal.ca/cgi-bin/*DiCoInfo*/search.cgi/.

Josselin-Leray, A. and Roberts, R.P. 2004. Le traitement des termes dans les dictionnaires généraux. In Béjoint, H. and Maniez, F. (eds.), *De la mesure dans les termes*. Lyon: Presses de l'Université de Lyon, 322-348.

L'Homme, M.-C. and Polguère, A. 2008. Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. In Maniez, F. and Dury, P. (eds.), *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*. Lyon: Presses de l'Université de Lyon, 191-206.

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

Moon, R. 2008. Technicalities of definition. In Maniez, F. and Dury, P. (eds.), *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*. Lyon: Presses de l'Université de Lyon, 83-98.

Random House Kernerman Webster's College Dictionary. 2011. http://kdictionariesonline.com/DictionaryPage.aspx?ApplicationCode=18/.

***Random House Webster's College Dictionary*. 2005.** New York: Random House.

Practical aspects of the description of terms: contexts, actantial structure and lexical relationships

Geneviève Camirand

As a translation student, my contribution to Marie-Claude Demers's directed study on the enrichment of a general dictionary's wordlist with the relevant contents of a specialized dictionary gave me the opportunity to investigate hidden aspects of some of the resources I will likely use extensively in a professional setting. My role has been to participate, as a research assistant, in developing terminological dictionary entries related to the computer field and contained in a terminological dictionary, the English version of the *DiCoInfo* developed at the Observatoire de linguistique Sens-Texte (OLST), that had been selected with the aim of supplying a general dictionary, the *Random House Kernerman Webster's College Dictionary* (RHKWCD), with new entries and meanings. And indeed, specialized and general resources being some of the main tools for translators, I took a particular interest in the various aspects of the project. The specific challenge of this project was to write entries that could be added to RHKWCD while respecting the guidelines usually applied in *DiCoInfo*. My role was to add data categories compatible with *DiCoInfo* (contexts, actantial structure, lexical relationships). Once added, these data categories could be used to write a definition and select examples that could be incorporated into RHKWCD.

The criteria for the selection of the terms to be included in the project were basically the following: among the terms whose meaning relative to the computer field was not already described in RHKWCD, only those that were not too specialized to be part of the general language were accepted. It is worth mentioning here that, since *DiCoInfo* is in constant evolution, as is the computer field, the list of terms established the first time is open to new additions.

Figure 1 is a screenshot of part of the list