Lexicographic interchange between a specialized and a general language dictionary[1]

Marie-Claude Demers, Ilan Kernerman & Marie-Claude L'Homme

Keywords: *general language dictionary*, *terminological database*, *specialized meaning*, *term*, *wordlist*.

Abstract

One of the important issues lexicographers need to address concerns the desired coverage of a dictionary's wordlist. This paper addresses the issue from a practical angle. We propose a method for comparing the contents of two resources and evaluating to what extent each can contribute to increase and improve the coverage of the other. Concretely, the project consists of comparing the contents of the English version of DiCoInfo (a dictionary of computing and Internet terms) with the appropriate entries of the *Random House Webster's College Dictionary* (RHWCD). The entries missing in one resource are considered for inclusion in the other, and vice versa. The approach proves beneficial for both resources. Approximately 100 entries were added to DiCoInfo and over 500 lexical items or meanings are being included in the RHWCD.

1. Introduction

An important issue lexicographers need to address concerns the coverage of a dictionary's wordlist. The question is relevant from the points of view of general as well as specialized lexicography, although it leads to different answers in each area. Specialized dictionaries should include all items that are related to the field they aim to cover. General language dictionaries include many specialized lexical items but attempt to cover fields and terms of more general interest (Alonso Campo 2008; Béjoint 1988; Boulanger 1996; Josselin-Leray 2005; Svensén 2009; Wiegand 1999).

This paper addresses the issue of coverage from a practical angle. We propose a method for comparing two specific lexicographic resources and evaluating to what extent each can contribute to increase and improve the coverage of the other. Concretely, we use a semi-automatic method to compare the content of the English version of DiCoInfo (a dictionary of computing and Internet terms) with the appropriate entries of the *Random House Webster's College Dictionary* (RHWCD).[2] The entries missing in one resource are considered for inclusion in the other, and vice versa. To our knowledge, while such comparisons using existing resources compiled separately may have been carried out in the past, none have consisted of such a systematic computerized process. From the point of view of DiCoInfo, the method should contribute to fill some gaps in its wordlist. Since the dictionary is still under construction, some entries are missing. From the point of view of RHWCD, this should also contribute to enhancing the vocabulary on these specific domains that might be overlooked in a general-purpose dictionary. With respect to the infiltration of technology, computers and the Internet in everyday life, it is assumed that such a dictionary should include more extensive vocabulary pertaining to these domains for the educated public.

2. DiCoInfo

DiCoInfo is an online specialized lexical database that contains computing and Internet terms. The wordlist consists of terms that can be found in many different texts on computing and the Internet (e.g. *boot*, *connect*, *dynamic*, *spam*, *upload*, *virtual*) but excludes very specialized items (terms that would be specific to a given operating system, e.g. *X server* or *run-level* that are more associated with Linux). The compilation process started nearly a decade ago and preliminary work focused on French only. The dictionary currently consists of three language versions: French (approximately 1,000 entries, including 15,000 lexical relationships), English (approximately 800 entries, including more than 4,500 relationships)[3] and Spanish (approximately 100 entries are currently online).[5] Entries contain the following data categories: part of speech, actantial structure, linguistic realizations of actants (i.e. arguments), lexical relations, contexts, and equivalents in other languages. Figure 1 shows a simplified reproduction of the entry *browse*.

browse, vt Status: 2

Actantial structure: browse: {internaut} ~ {Internet } with {browser 1}

Linguistic realizations of actants

Contexts

- … and especially if you work at a larger company and browse the Web while you are at work …
- You can use it to type documents, send e-mail, browse the Internet and play games.

Lexical relations

Related meaning	navigate
Noun	browsing

French. naviguer

Spanish: navegar

Figure 1. Entry for the term *browse* in DiCoInfo.

DiCoInfo's theoretical and methodological principles are based mostly on Explanatory Combinatorial Lexicology (ECL, cf. Mel'čuk et al. 1984-1999). In addition, it adheres to some principles that are closer to lexicography than to terminology: most terms considered are single-word units (multi-word units are selected if their meaning is non-compositional); and, polysemy is analyzed from a structural perspective (series of interactions a unit has with others), etc. These characteristics made the specialized resource particularly interesting from the point of view of a comparison with a general language dictionary.

3. RHWCD

K Dictionaries (KD) acquired the rights for the acclaimed *Random House Webster's College Dictionary* (RHWCD) in 2009 (cf. Levine and Pearsons, 2010). Since its launch in 1947 (as the *American College Dictionary*), RHWCD underwent regular revision and updating until the last edition appeared in 2005, in conjunction with dismantling Random House's dictionary department. This is a major English general-language native-speaker dictionary, including well over 200,000 definitions, targeted primarily at American university students. Following its purchase, KD started to revive RHWCD and prepare to use it as a base for digital and bilingual versions. To begin with, the data was converted from SGML to XML format and the proprietary

phonetic transcription converted into standard IPA. Then, the first new entries began to be compiled, and a meticulous manual was drafted.

The headwords were selected by specialist editors and consultants in many fields, and the entries generally consist of the headword, pronunciation, part of speech, inflections, definitions, examples, phrasal verbs and idioms, usage notes and grammatical information, subject labels and labels of time and of place, synonym studies and etymology. Some of these elements appear in the sample entry shown in Figure 2.

browse /braʊz/ *v.* **browsed, brows•ing,** *n.* --- *v.t.* **1.** to eat, nibble at, or feed on (foliage, berries, etc.). **2.** to graze; pasture on. **3.** to look through or glance at casually. --- *v.i.* **4.** to feed on or nibble at foliage, lichen, berries, etc. **5.** to graze. **6.** to glance at random through a book, magazine, etc. **7.** to look leisurely at goods displayed for sale, as in a store. --- *n.* **8.** tender shoots or twigs of shrubs and trees as food for cattle, deer, etc. **9.** an act or instance of browsing. [1400–50; late ME]

Figure 2. Entry for *browse* in RHWCD.

4. Interchange

The methodology of the cooperation between OLST and KD concerning DiCoInfo and RHWCD basically consists of identifying lexical items or meanings in one of the resources by comparing its wordlist to the other (the process was partially automated; this was facilitated by the fact that the data in both resources is encoded in XML), and including relevant items or meanings in each resource.[4] The following sections describe the specific steps of each task.

4.1 *Data from RHWCD to the English version of DiCoInfo*

One part of the comparison between the two resources consisted of identifying missing lexical items or meanings in DiCoInfo. This was done as follows:

- Extract from RHWCD all the entries that contain the words *computer*, *Internet* or *technology* (as part of the headword, subject field label, definition or example). The entries were extracted with their full information and matched against the DiCoInfo headword list. The extraction produced a list from which some lexical items missing from DiCoInfo could be identified (e.g. *artificial intelligence, avatar, capture, clear*).
- Create entries for the terms found in RHWCD and not in DiCoInfo by adapting the RHWCD entry to the DiCoInfo editorial style, link them to their French equivalents, and post them on the online version of DiCoInfo. About 100 entries were compiled so far. Figure 3 shows how the term *avatar* from RHWCD was added to DiCoInfo.

av•a•tar /ˈæv əˌtɑr, ˌæv əˈtɑr/ *n.* **1.** an incarnation of a Hindu god. **2.** an embodiment or personification, as of a principle, attitude, or view of life. **3.** Computers. a graphical image that represents a person, as on the Internet. [1775–85; < Skt *avatāra* a passing down]

Avatar, n
Actantial structure: an avatar: ~ created by {user 1} to represent {user 1}[6]
Linguistic realizations of actants
Contexts:
• … three dimensional avatars which represent you in a computer based world.
• Users create a math avatar, then take part in the activities, where the avatar appears as a participant in different game shows.
Lexical relations

Related meaning	character
The user creates an a.	create an ~
The a. acts on the user	the ~ represents …

French. avatar

Figure 3. Addition of RHWCD data to DiCoInfo.

4.2 *Data from the English version of the DiCoInfo to RHWCD*

The second part of the comparison consisted of identifying missing lexical items or meanings in RHWCD. This was done as follows:

- Extract the terms listed in DiCoInfo and compare them with the headword list of RHWCD. The lexical items were extracted along with the following data categories: part of speech; number of meanings for polysemous items; and any additional information (e.g. actantial structure) currently available in order to further compare with these entries in RHWCD.
- Automatically compare the terms stored in DiCoInfo with those listed in RHWCD. The comparison took into account the part of speech of the headwords in both sources. The articles were extracted with all the information contained in the entries as shown in Table 1.

Table 1. Examples of results of the automatic comparison between DiCoInfo and RHWCD.

Lexical unit in DiCoInfo along with its actantial structure	**POS**	**Information extracted from RHWCD**
Address$_1$: an address: ~ used by Agent{processor 1} to act on Destination{data 1} on Location{memory 1}[8]	n.	• the place or the name of the place where a person, organization, or the like is located or may be reached. • a direction as to the intended recipient, written on or attached to a piece of mail. • a usu. formal speech or written statement directed to a particular group. • skillful and expeditious management; ready skill. • manner of speaking to others; personal bearing in conversation. • the use of a name or title in speaking or writing to a person: *forms of address.* • a label, as an integer or symbol, that designates the location of information stored in computer memory.

		• Usu., addresses. attention paid by a suitor; courtship. • Obs. preparation.
addressable 1: addressable: ~ Location{memory}	adj.	• capable of being addressed. • (of a cable-TV system) capable of calling up any available channel. • (of computer data) capable of being accessed.

Some of the specificities of each resource had to be considered (for instance, the fact that entries in RHWCD are devoted to lemmas whereas in DiCoInfo each lexical unit – distinct meaning – is described in a separate entry). This led to the identification of two types of items: (a) lexical items missing from the RHWCD wordlist (e.g. *addressing, bootable, read/write head*);[7] and, (b) lexical items present in RHWCD.

- Carry out a manual analysis of the output of the automatic comparison. Five different cases were identified, as shown in Table 2.

Table 2. Manual analysis of the results of the automatic comparison.

Label of case	**Definition**	**Examples**
A1	Lexical item in RHWCD with a clear indication that it belongs to the field of computing (a usage label is provided or the word *computer* or something similar appears in the entry)	*boot*, vt *browser*, n
A2	Lexical item recorded in RHWCD with a meaning associated with the field of computing but without a clear indication	*calculate*, vt *computation*, n
B	Lexical item recorded in RHWCD but the meaning associated with the field of computing is missing	*bookmark*, n *run*, vi
C	Lexical item recorded in RHWCD with a "general meaning" that can apply to the field of computing	*cable*, n. *case*, n.
D	Lexical item not recorded in the RHWCD	*bookmark*, vt *computable*, adj *read/write head*, n.

Lexical items classified under B or D are those that could be added to RHWCD. More than 1,000 such items were identified as potential candidates for inclusion. However, this figure includes synonyms and graphical variants. Altogether, over 500 potential additions to RHWCD were identified (new entry or new meaning).

4.3 *Addition of data to the RHWCD wordlist*

For some of the terms that are not listed in RHWCD and that seem interesting to add to a general language dictionary, certain issues concerning the compilation of new entries still need to be worked out.

For the time being, part of the data recorded in the DiCoInfo entries (headwords, part of speech, synonyms, variants, antonyms, and French equivalents) is extracted as such with some adaptations (definitions written in the RHWCD style and examples that are simplified versions of the contexts provided in DiCoInfo). Figure 4 gives examples of the form in which entries are extracted from DiCoInfo.

browse, vt
Def.: to examine (pages on a network) in search of specified information.
Ex.: *to browse the Web*
Syn.: surf
Fr.: naviguer 1

executable 1, n
Def.: a file or program that can be run by a computer.
Ex.: *Scripts are compiled into executables.*
Syn.: executable file
Fr.: exécutable 1

unzip, vt
Def.: to restore the size of a compressed file with a decompression program.
Syn.: decompress 1, decrunch, uncompress
Ant.: zip 1
Fr.: dézipper 1
Figure 4. The form in which entries are extracted from the DiCoInfo data.

However, before being added to RHWCD, further adaptations are necessary. New lexical items (such as *executable*) can simply be added to the wordlist with some additional data categories (e.g. pronunciation, etymology, subject field label). New meanings (such as for *browse* and *unzip*) are somewhat more difficult to deal with, as lexicographers have to decide where to insert the added meaning to an existing entry.

5. Conclusion

Our approach, which consists of comparing the contents of a general language dictionary and a specialized one, proves beneficial to both sides, since it allows us to identify missing lexical items or meanings in each resource. Many items have already been added to the DiCoInfo wordlist, and new lexical items and meanings are currently being added to RHWCD. However, subtle adjustments remain necessary to suit the style of definitions and examples from one dictionary to the other.

Interestingly, this method has a beneficial side-effect: the compilation of new entries sometimes leads us to notice other terms (which were not identified during the initial comparison) that are missing in DiCoInfo or RHWCD (e.g., *spam -> spammer*; *case-sensitive -> case-insensitive*). These terms can later be added to each wordlist.

Since it is bidirectional, our method offers an excellent starting point for identifying missing items in different kinds of lexicographic resources. It allows us to work with data that is already encoded either in a general language dictionary or a specialized dictionary and simply adapt it to the specific requirements of each resource.

Notes

[1] The authors would like to thank Caroline Gagné, Charles Levine, Enid Pearsons and Benoît Robichaud for their help with the work described in this paper, as well as two anonymous reviewers for their comments on a previous version.
[2] In 2010 OLST and KD began to collaborate on introducing entries from RHWCD into DiCoInfo and vice versa, as part of developing a method for comparing a specialized lexicographic resource and a general one.
[3] The coverage in English was boosted thanks to the project described in this paper.
[4] The Spanish version is developed in collaboration with the team TecnoLeTTra of the Universitat Jaume I, in Spain, and its coverage should increase in the coming years.
[5] Part of the work described in this section was carried out by Marie-Claude Demers and is described in her M.A. dissertation (Demers 2011, 2012).
[6] The definition is not available yet on the online version, but the definition given in the RHWCD has been stored in the term *record* and will be edited according to the DiCoInfo rules at a later stage.
[7] Some lexical items identified at this stage were actually present in RHWCD, not as main headwords but as sub-entries (e.g. *algorithmic*, *hard-wired*). However, since most of these items were not defined, we still considered them as relevant candidates for inclusion.
[8] At the time, DiCoInfo did not contain definitions. The actantial structure gives a basic idea of the meaning of the unit being analyzed.

References

A. Dictionaries

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

RHWCD. *Random House Webster's College Dictionary*. 2005. New York: Random House; 2010. Tel Aviv: K Dictionaries.

B. Other literature

Alonso Campo, A. 2008. 'Environmental terminology in general dictionaries.' In Bernal, E. and J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress*. Barcelona: Institut universitari de lingüística aplicada (IULA).

Béjoint, H. 1988. 'Scientific and technical words in general dictionaries.' *International Journal of Lexicography* 1.4: 354–368.

Boulanger J.-C. 1996. 'Les dictionnaires généraux monolingues, une voie royale pour les technolectes.' *TradTerm* 3: 137–151.

Demers, M. C. (2011). *Travail de recherche en lexicographie bilingue: enrichissement de la nomenclature d'un dictionnaire général à partir de celle d'un dictionnaire spécialisé.* Travail dirigé réalisé à l'Université de Montréal. Université de Montréal, Montréal.

Demers, M. C. (2012, forthcoming). 'Using a specialized dictionary to enrich a general language lexical resource.' *Kernerman Dictionary News 20.*

Josselin-Leray, A. 2005. *Place et rôle des terminologies dans les dictionnaires unilingues et bilingues. Étude d'un domaine de spécialité: vulcanologie*. Thèse de doctorat, Université Lumière Lyon 2, Lyon.

Levine, C. M. and E. Pearsons. 'The Random House dictionary tradition.' *Kernerman Dictionary News* 18, 2–5.

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

Svensén, B. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-making*. Cambridge: Cambridge University Press.

Wiegand, H. 1999. 'Languages for Special Purposes in the Monolingual Dictionary: Criticism, Provocations, and Practical and Pragmatic-oriented Suggestions.' In Inmmken, A. and W. Wolski (eds.), *Semantics and Lexicography. Selected Studies (1976-1996).* Tübingen: Max Niemeyer Verlag.