Improving the terminological dimension in a computational lexicon

Thierry FONTENELLE

Abstract

Cet article décrit une expérience effectuée sur la base de données du Longman Dictionary of Contemporary English (L DOCE). Nous proposons une méthode permettant d'améliorer le dictionnaire en localisant des omissions dans l'attribution des codes matière ou champs sémantiques. Nous nous concentrons plus particulièrement sur la terminologie juridique et évoquons les utilisateurs possibles des informations terminologiques contenues dans une base de données lexicale.

Introduction

The past decade has witnessed the emergence of a new discipline within the field of computational linguistics, viz computational lexicography. The latter term refers to the use that is being made of various machine-readable dictionaries and terminological data banks to feed the lexical component of Natural Language Processing (NLP) systems for applications such as machine translation (MT), man-machine interface, automatic indexing, speech recognition, etc.

The development of computational lexicography is largely due to the attention paid by researchers to several commercial dictionaries whose electronic versions are believed to house a lot of the information that is necessary for the above-mentioned NLP applications. Part of the research community is indeed convinced that it is neither desirable nor feasible to start from scratch if we want to construct large lexicons with hundreds of thousands of entries. Another term has therefore been coined : *reusability of lexical resources*. It has now become a buzzword and is widely used to refer to various types of effort to re-use the more or less formalized pieces of information that are to be found in published dictionaries. The present paper should

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

be seen as an attempt to show that some computerized dictionaries, although far from perfect, can be used and improved once they have been arranged in a format that allows efficient processing.

A lexical data base

The experiment described here is based on the Longman Dictionary of Contemporary English (Procter 1978, henceforth LDOCE). In 1979, Professor Noël's English Department at Liège University was granted access to the magnetic version of this dictionary under contract with the publishers. The magnetic tape had originally been used for the typesetting process and it took a few years to transform and implement the dictionary in data base format. The relational model (dBase III+)[1] was chosen for the implementation of LDOCE. The data base design and the transformation of the tape into the dBase relational tables were carried out by Jacques Jansen. The retrieval software used here was written in Clipper[2] by Archibal Michiels.

More information on the structure of this data base can be found in Michiels 1990. Suffice it to say that the Liège lexical data base is based on a series of distinct tables related by one or several common fields. This approach enables us to take into account the intricate nature of lexical data. Lexical items may have different numbers of definitions and definitions may have different numbers of associated examples: the relational model is found to be very efficient in this respect (broadly speaking, the Liège data base includes the following tables: COWORD with a.o. the lemmas and part-of-speech information; CODCOD with definition numbers and grammatical codes; CODEFI with definition lines, semantic and subject field information; COEXAM with example lines, etc. All these indexed files are linked by one or several common fields).

It is important to note that the tape provides some types of information which are not found in the printed version of this learner's dictionary. For example, it includes semantic codes à la Katz and Fodor which enable the lexicographer to encode lexical co-occurrence restrictions. These restrictions are based on a system of binary features assigned at definition level. For instance, T will be used to mark [+ABSTRACT] nouns whereas H will be used for [+HUMAN] nouns (note that it can also be used to impose a constraint on the choice of co-occurring items: H can also be used to refer to the head of the noun phrase that can be the subject of a verb or

[1] dBase is a trademark of Ashton-Tate, Inc.

[2] Clipper is a trademark of Nantucket Corporation.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

that can be modified by an adjective – in this case it means that the noun in question should be [+HUMAN]).

Beside semantic information, LDOCE also includes subject field information assigned at definition level. A 4-byte field (deflmat in the Liège data base) can be filled by one or two Merriam-Webster codes : the first two letters refer to a subject field (terminological sublanguage) code for a given definition (e.g./SP/=Sports; /MD/=Medicine; /BZ/=Business...). The last two letters, if any, refer either to another subject field or to a subdivision within the first field (Z being used as a separator). The following examples illustrate the two possibilities :

abortion	MDLW	(Medicine/biology + Law)
customs duty	ECZT	(Economics + subfield = Tax)

The LDOCE data base makes it possible to retrieve all the items that belong to a particular discipline. At least three applications are possible :

1. Translation

Translators always need information about the topic they are dealing with. Having access to a domain-related sublanguage may prove interesting before tackling the translation of a specialized text. The automatic generation of a glossary on the basis of subject field information may therefore be considered as a crucial by-product of dictionary making.

2. Computer-Assisted Language Learning (CALL)

CALL exercises can be automatically created on the basis of a lexical data base. For example, the words which are associated with a given subject field can be erased from a text, the learner having to fill the gaps by choosing among the words listed in alphabetical order (cloze exercises). This type of exercise is extremely interesting in the field of ESP (English for Special Purposes) : law students, for instance, will only be tested on the vocabulary of the sublanguage they are expected to master (see Noël & Jansen 1986).

3. Automatic indexing

Jansen 1989 shows that LDOCE can be used to automatically produce thematic indexes on the basis of the subject field codes contained in the computerized version.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

Indeed, all the words of any text can be scanned by the computer to establish the statistical frequency of the subject fields. This makes it possible to discover the topic of a given text, which proves crucial in the domain of machine translation where this type of information can be used to select a given reading and solve possible ambiguities (preference mechanism).

Subject field information and consistency

The above considerations point to the necessity of including such terminological information while compiling a dictionary. However, it should be borne in mind that LDOCE has been compiled by human beings and for human beings (basically for learners of English). As has already been pointed out by Michiels 1982, no attempt has ever been made to develop automatic checking procedures to ensure consistency. This explains why the dictionary is often found to be marred by errors of various types (errors in the assignment of grammatical codes, missing parentheses, erroneous semantic coding, etc). The purpose of this paper is to show that it is possible to enrich the dictionary and supplement it with subject field codes that the lexicographer has failed to assign.

I chose to focus on the subject field law (LW in the field defimat). This discipline has indeed developed a sublanguage of its own and is particularly well represented in LDOCE (1544 occurrences of LW in the data base). It should be stressed that, while the Merriam-Webster codes are found only on the tape, the LDOCE lexicographers sometimes mentioned the specialized subject field in which a given word sense can be used (field label). *Codicil* is a case in point :

> **codicil** *n law* an addition to a WILL (a paper stating who is to have one's possessions after one's death).
> **(defimat** : LW..)

In the Liège data base, the field label (*law*) appears within the definition field. In the present case, it means that there is some redundancy in the assignment of terminological information. The following example shows that this type of practice is, however, far from being general and that a human reader is quite often not given any clue apart from definitions themselves :

> **code** *n* 4 [C] a collection of laws : *the Napoleonic code*
> **(defimat** : LW..)

In passing, it may be worth stressing that only the fourth definition of *code* is assigned the /LW/ code. This reading would be chosen as the most probable one if the statistically most significant code in a text was found to be LW, i.e. if the automatic content-analysis procedure discovered that the text in question deals with

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

a legal subject. Such a preference mechanism is of course not one hundred percent reliable but can be applied to solve an ambiguity and rule out less probable readings.

The two examples above show that the reader cannot always rely on the field labels that are provided within the definition field. They are frequently absent and not sufficiently formalized to be put to good use in statistical analyses. The best way to formalize that information is of course to encode it with the Merriam codes but, again, consistency has not been reached. The following example illustrates a case where the LW code should have been assigned :

> **breach** *n* 1 [C;U] an act of breaking, not obeying, or not fulfilling a law, promise, custom, etc. : *Your action is a breach of our agreement.* | *You are in a breach of your contract.*

The **defimat** field is empty and the user is left with the definition and the examples to discover that this term has a particular legal meaning.

Keywords and controlled vocabulary

Once it has become evident that a particular code has not been assigned where it should have been, a method has to be developed to locate possible candidates for the assignment of this code. It can be done readily thanks to one essential feature of the Longman dictionary : LDOCE definitions are written in a controlled vocabulary of around 2,000 items. The words that do not belong to the defining vocabulary are capitalized. This constraint imposed on the Longman lexicographers spurred them to use "ready-made" definition patterns that prove to be useful to extract semantic information from the dictionary (see inter alia Fontenelle & Vanandroye 1989, Fontenelle 1990, Michiels & Noël 1982).

In the present case, the power of the relational data base can be exploited to retrieve all the definitions that contain some legal keywords and that are not coded LW (two conditions). The analysis of the words that have been assigned the LW code shows that the following words have been used repeatedly in the definitions of these legal terms :

> law - laws - lawyer - lawyers - lawful - lawfully - lawfulness - lawlessness - unlawful - unlawfully - legal - court - judge - judges - judged - justice

As can be seen, a lot of these words are actually derived forms of "law". Searching on this particular string of characters (the three letters l-a-w) will then automatically yield a list of entries where the derived forms are also used.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

The problem with "judge" and "justice" is a tricky one since these two words are not always used in their prototypical senses in LDOCE. The General Introduction (p.ix) tells us that *"a rigorous set of principles was established to ensure that only the most "central" meanings of these 2,000 words [...] were used"*. A search on "judge" and "justice" in definitions that do not bear the /LW/ code reveals that these two words are always used in a non-legal sense (cf. appeal(4), appraise(1), armchair critic, contest(2), criticize(1), misjudge(1), redress(1), trust(1)...). We will therefore not consider them further here.

The polysemy of words such as "law" or "court" is easier to take into account. These two words are also used in fields that are related to law proper (politics, sociology) or in disciplines that are distantly related (science or religion for "law", sports or history for "court"). This means that many words that are defined in terms of "law" or "court" can actually be assigned other codes than LW. The Merriam-Webster classification provides lexicographers with codes as diverse as /SP/ (Sports : cf. appeal(4), net(2c), netball(1)), /SO/ (Sociology : cf. approved school, commission(4)), /PL/ (Politics : cf. anarchism(1), bill(1), civil disobedience), /PH/ (Philosophy : cf. moral(1), natural law(1)), /SI/ (Science : cf. formula(1), principle(2b)), /CT/ (Court games : cf. squash(3)), /HI/ (History : cf. durbar(1)), /RL/ (Religion : cf. canon(1), canon law(1)), etc. It is therefore preferable not to rule out any subject field code and to query all the words that contain the keywords we have examined and that are not coded for any specific subject field.

The queries on the data base will then be specified in the following way :

- Filters on definitions

"law" $ def .or. "legal" $ def .or. "court" $ def
(the definition field includes (operator $) law, legal or court)

- Filter on domain (subject matter at definition level)

defimat="...."
The field only contains dots, i.e. no code has been assigned.

Evaluation of the results

The query described above enables us to retrieve 176 lexical items which do not bear any subject field code : around one hundred should indisputably be assigned the /LW/ field code. A few words even include the field label *law* in italics in the definition. Consider the following examples :

- behaviour *n* 3 *law* to behave well

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

- **discharge** *n* 10 *law* to put an end to (a court order)
- **grievous** *adj* 3 *law* hurt done to a person's body in an attack, for which the attacker may be charged in a court of law
- **guardian** *n* 2 *law* a person who has the responsibility of looking after a child not his own, esp. after the parents' death
- **suicide** *n* 3 *law* a person who does this

It is also possible to retrieve collocations consisting of a verb and the noun *law* where the verb refers to the creation or activation of a law. Examples of verbs that have not been assigned the /LW/ code include **abide by(1)** and **enforce(1)**. Other collocations with *law* are found under verbs meaning eradication or nullification (e.g. **coach(7)**, **contravene(1)**, **revoke(1)**, **transgress(2)**).

Since subject matter is coded in a four-byte field, lexicographers can use two codes when a given item can be used in two different disciplines without change of meaning. The following example illustrates a case where /SI/ (Science) and /LW/ should have been assigned :

> **evidence** *n* [U (*of, for*), U5] 1 a (esp. in science or law) words which prove a statement, support a belief, or make a matter more clear

Similarly, the definition of **vacation** (1 a esp. BrE one of the periods of holiday when universities or law courts are closed) refers to two subject fields. **Defimat** should therefore have been coded /EDLW/ (ED stands for Education).

The preceding example illustrates the use of the word "law" as part of another term, namely *law court*. The term *court of law* is used more frequently by the LDOCE lexicographers. The co-occurrence of these two words in a definition is again an indicator that this word sense should have been assigned a /LW/ code. Twelve items display that particular pattern in their definition : **adjourn(1b)**, **appear(5)**, **charge(11)**, **examination(3)**, **exhibit(2)**, **grievous(3)**, **judgment(6)**, **judgment(7)**, **kiss(7)**, **suborn(1)**, **trial(5b)**, **vacation(1a)**.

The word "court" alone can be found in a few entries such as, for instance, **contumacious(1)**, **contumacy(1)**, **summing-up(1)**, **sum up(1)**, **swear(3)**, **unheard(1)** or **witness(4)**.

The word "lawyer" is also used in a few definitions. Consider for instance the idiom **be called to the Bar** (under **Bar** 3b) which is defined as "to become a lawyer".

The words "lawful", "lawfully" and "unlawful" are also used to define items which should have been coded /LW/. Consider **consent(2)**, **crackdown(1)**, **effect(5)**, **possession(8a)**, **vest in(1)**, **vest with(1)**, **wrongful(2)**.

Of course, caution should be exercised in assigning a subject field code to a given word sense. One should not be too hasty in concluding that a given item

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

should be considered as a legal term precisely because its definition includes the string "law". The following example illustrates a case where the entry definitely refers to religion (RL should thus be assigned to it) :

> **forbidden** *adj* [Wa5; A] 2 that may not be used, entered, or visited by ordinary people, as because of a religious law : *the forbidden city*

Observant(1) and **observe(2)** provide similar examples where RL ought to have been used.

Conclusions

The data obtained in the experiment described above show that the word senses that are retrieved should not be coded automatically. Indeed, we believe that the computer should be seen here as a tool used to interactively and semi-automatically ensure consistency in the lexical data base. We have tried to show that a relational data base provides an adequate and powerful tool which might be put to good use in a lexicographer's workbench. Checking procedures can be developed relatively easily and make it possible to avoid a number of errors and omissions.

The experiment carried out here on legal terminology would be equally valid for other sublanguages. Studying the language of definitions and identifying relevant keywords would enable lexicographers to enrich their dictionary and improve the terminological dimension in other fields as well.

Bibliography

FONTENELLE Th., VANANDROYE J. (1989) *Retrieving Ergative Verbs from a Lexical Data Base,* dans *Dictionaries : Journal of the Dictionary Society of North America*, Vol.11, pp. 11–39.

FONTENELLE Th. (1990) *Automatic extraction of lexical-semantic relations from dictionary definitions,* forthcoming in *Proceedings of the EURALEX Fourth International Congress*, Biblograf, Malaga.

JANSEN J. (1989) *Apport contrastif des dictionnaires généraux de la langue au problème de l'indexation automatique dans le discours techno-scientifique,* dans *Meta*, Vol. 34, n° 3, pp. 412–427.

MICHIELS A., NOËL J. (1982) *Approaches to Thesaurus Production,* dans *Proceedings of the International Conference on Computational Linguistics (COLING 82)*, ACL, Prague, pp. 227–232.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.

MICHIELS A. (1982) *Exploiting a Large Dictionary Data Base,* unpublished PhD thesis, University of Liège.

MICHIELS A. (1990) *Retrieving verb classes from LDOCE,* (forthcoming).

NOËL J., JANSEN J. (1986) *Auto-apprentissage de l'anglais langue étrangère sur micro-ordinateur ayant accès à des dictionnaires informatisés,* dans *Revue, Informatique et Statistique dans les Sciences humaines,* XXII, 1–4, pp. 105–134.

PROCTER P. (ed.) (1978) *Longman Dictionary of Contemporary English,* Longman, Harlow.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXVI, 1 à 4, 1990. C.I.P.L. - Université de Liège - Tous droits réservés.