DEFIDIC, a lexical database for computerized translation selection

Nicolas DUFOUR

Abstract. The present paper provides a description of the English-to-French lexical database developed in Liège in the framework of the DEFI word sense discrimination project. The elaboration process and structure of the database are set out in some detail, as well as its intended and potential uses.

Résumé. Le présent article décrit la base de données lexicale anglais-français développée à Liège dans le cadre du projet de désambiguïsation sémantique DEFI. L'élaboration et la structure de la base de données sont décrites en profondeur, ainsi que ses utilisations prévues et potentielles.

Keywords: Machine-tractable dictionaries, word sense discrimination, automatic translation selection, computational lexicography.

Mots-clés : Dictionnaires électroniques, désambiguïsation sémantique, sélection automatique de traductions, lexicographie computationnelle.

1. Introduction

DEFI is a five-year basic research project in the field of word sense discrimination and translation selection. The project's general objective is to create a prototype that would provide the reader of a text in a foreign language (in this case either French or English) with the best possible translation of any word he/she selects online, depending on

✉ DEFI project; Université de liège; Quai Roosevelt, 1 (Bât A4); B-4000 Liège (Belgique)
http://engdep1.philo.ulg.ac.be/michiels/defi.htm.
Fax: +32 4 366 57 84 E-mail: Nicolas.Dufour@ulg.ac.be

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

its environment in the source text. That look-up system, of which an early version is already up and running, works as a 'text-dictionary matcher' that tries to find the lexical database entry (*i.e.*, the translation) whose linguistic and metalinguistic information—part of speech, style and domain labels, collocational restrictions, etc.—best matches the elements found in the source text. The various possible translations of the selected word, or of the multi-word lexeme it is a part of, are given 'preference scores' depending on the number and quality of these matching elements, and provided to the user in order of decreasing preference. The DEFI prototype could thus be regarded as a 'comprehension assistant' similar in its goals to Rank Xerox's LOCOLEX (Bauer *et al.*: 1995), albeit with a very different approach.

The project makes use of a wide range of lexical resources to achieve its goals. Apart from our terminological database none were developed from scratch in Liège, our aim being to make the best possible use of the available data, to be obtained either from the public domain or via a research agreement with the copyright owners. DEFI's lexical resources are the following:

— the Collins-Robert and Oxford-Hachette English/French and French/English dictionaries (*cf.* Corréard & Grundy 1994, Duval & Sinclair 1993);
— WordNet (*cf.* Miller *et al.* 1990);
— Roget's Thesaurus of English Words and Phrases;
— the COBUILD, LDOCE and CIDE dictionaries of English (*cf.* Sinclair 1987, Procter 1978, Procter 1995);
— a home-made bilingual database of archaeological terminology (based on our testbed corpus of scholarly articles in the field of Aegean archaeology), which is still being compiled.

All three monolingual dictionaries are still 'on the shelves', they will be used in the later stages of the project to provide a bridge between the source text and the bilinguals. Our principal resources obviously consist in the two bilingual dictionaries, whose transformation into a single database has taken up most of the author's time in the first eighteen months of the project.

This paper focuses on the elaboration and structure of the English-to-French side of the database, which was completed earlier this year. Work on the French-to-English part is still in progress, but should lead to similar results.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

2. From 'raw tapes' to a common, machine-tractable format

The Oxford-Hachette and Collins-Robert English/French dictionaries (henceforth OH and CR) are popular general-use bilingual dictionaries. Both are medium-sized, and cover general language as well as some technical terms.

The electronic versions we obtained from the publishers (the 'raw tapes') are, in both cases, the files that are used to drive the typesetting process of the paper dictionaries. They contain exactly the same information as the print versions, and this information is organized in the same, human-oriented manner.

The CR tape is purely typographic, all tags signalling changes in typeface and typesize, or the inclusion of special characters such as subentry numbers/letters. The OH tape is SGML-tagged, each piece of information (except, strangely enough, the translation) being enclosed between specific markers (<hw> for headword, <la> for field labels, etc).

In order to illustrate the 'rawness' of the original tapes, I reproduce here the entries for *abstract* in the CR and OH tapes, together with their equivalents in print (only phonetic transcriptions are omitted).

Abstract in CR:

abstract **1** **adj** *idea, number, noun, art, artist* abstrait. ~ **expressionism** lyrisme abstrait. **2** **n** **a** (*Philos*) abstrait *m*; (*idea*) abstraction *f*. **in the** ~ dans l'abstrait. **b** (*summary*) résumé *m*, abrégé *m*. (*Fin*) ~ **of accounts** extrait de compte. **c** (*work of art*) œuvre abstraite **3** **vt** **a** (*also Chem: remove*) extraire (*from* de). **b** (*steal*) soustraire (*sth from sb* qch à qn), dérober. **c** (*summarize*) *book* résumer.

```
>H<abstract>X<
>[<>HST<>ae<bstr>ae<kt>]<
>g1< >POSB<adj
>MI<idea, number, noun, art, artist
>R<abstrait.
>B<>~< expressionism
>R<lyrisme abstrait.
>g2< >POSB<n
>sa< >MI<>(<Philos>)<
>R<abstrait >GENR<m>R<;
>MI<>(<idea>)<
>R<abstraction >GENR<f>R<.
>B<in the >~<
>R<dans l'abstrait.
```

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

```
>sb< >MI<>(<summary>)<
>R<re>ac<sume>ac< >GENR<m>R<, abre>ac<ge>ac< >GENR<m>R<.
>MI<>(<Fin>)<
>B<>~< of accounts
>R<extrait >GENR<m
>R<de compte.
>sc< >MI<>(<work of art>)<
>R<>oe<uvre abstraite.
>g3< >[<>ae<b>HST<str>ae<kt>]<
>POSB<vt
>sa< >MI<>(<also Chem: remove>)<
>R<extraire >(<>MI<from
>R<de>)<.
>sb< >MI<>(<steal>)<
>R<soustraire >(<>MI<sth from sb
>R<qch a>gr< qn>)<, de>ac<rober.
>sc< >MI<>(<summarize>)<
book
>R<re>ac<sumer>R<.>QL<
```

Abstract in OH:

abstract I *n* 1 (theoretical) **the** ~ l'abstrait *m*; **in the** ~ dans l'abstrait; **2** (summary) résumé *m*; **3** Fin, Jur extrait *m*; **4** Art œuvre *f* abstraite.
II *adj* **1** (theoretical) abstrait; **2** Art abstrait; **3** Ling [*noun, verb*] abstrait.
III *vtr* 1 (summarize) **to ~ sth from** tirer qch de [*documents, data*]; **2** (remove) sout dérober (**from sb** à qn; **from sth** dans qch); **3** (theorize) **to ~ sth from sth** extraire qch de qch.
IV *v refl* **to ~ oneself from sth** se soustraire à qch.

```
<se><hw>abstract</hw> <s1 num=I nl=n><pr><ph>"&bstr&kt</ph></pr>
<ps>n</ps> <s2 num=1>(<ic>theoretical</ic>) <lo>the &hw.</lo>
l'abstrait <gr>m</gr>; <le>in the &hw.</le> dans l'abstrait</s2>; <s2
num=2><ann><la>Univ</la></ann>(<ic>summary</ic>) r&ea.sum&ea.
<gr>m</gr></s2>; <s2 num=3><la>Fin</la>, <la>Jur</la> extrait
<gr>m</gr></s2>; <s2 num=4><la>Art</la> &oe.uvre <gr>f</gr>
abstraite</s2></s1>. <s1 num=II><pr><ph>"&bstr&kt</ph></pr>
<ps>adj</ps> <s2 num=1>(<ic>theoretical</ic>) abstrait</s2>; <s2
num=2><la>Art</la> abstrait</s2>; <s2 num=3><la>Ling</la> <co>noun,
verb</co> abstrait</s2></s1>. <s1 num=III><pr><ph>@b"str&kt</ph></pr>
<ps>vtr</ps> <s2 num=1>(<ic>summarize</ic>) <ls>to &hw. sth from</ls>
tirer qch de <co>documents, data</co></s2>; <s2
num=2>(<ic>remove</ic>) <la>sout</la> d&ea.rober (<pp><sp>from
sb</sp> &ag. qn; <sp>from sth</sp> dans qch</pp>)</s2>; <s2
```

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

```
num=3>(<ic>theorize</ic>) <ls>to &hw. sth from sth</ls> extraire qch
de qch</s2></s1>. <s1 num=IV><ps>v refl</ps> <lo>to &hw. oneself from
sth</lo> se soustraire &ag. qch</s1>.</se>
```

When considering these raw tapes, one should always bear in mind that they are solely intended to serve as the basis for printed dictionaries, which are themselves to be used by humans. Their formal structure, which a trained human user will mostly decode without giving it a second thought, is characterized by their frequent use of implicit inheritance rules and information elision. This is an obvious choice to make for paper dictionaries, which have to store information into the smallest possible space for reasons of marketing and user-friendliness. It is however a pity for the NLP people who want to turn these files into a machine-tractable format, and end up pondering for weeks and months over the meaning of semicolons, over what part of a complex headword should be substituted for the swung dash (~) within the entry body, and whether some pieces of information found at the beginning of (sub-)entries systematically apply to the remaining parts.

This is no less true of OH's SGML tagging than of CR typesetting codes: in both cases the basic information units are relatively easy to identify, and almost impossible to organize without human intervention. As a matter of fact, OH turned out to be the harder nut to crack, due partly to the inexperience of the author at the time (it was processed first) and partly to its more complex and less consistent entry structure.[1]

Turning the raw tapes into machine-tractable files meant disentangling the inheritance structures of the original entries in order to create a series of self-supporting, independent records containing all the linguistic and metalinguistic information needed for their interpretation. In the resulting two 'DEFI dictionaries', the record boundary is not the traditional *headword* any more, but a more specific unit which we call the *lemma*, *i.e.* the English word, phrase or example sentence to be translated. Each record is thus centered on a lemma/translation pair, to which such information is added as part of speech, field labels, collocational restrictions, prepositional environment and so forth. Criticism has been levelled at such an organization of dictionaries into independent records (Boguraev

[1] Lack of space does not allow a more thorough discussion of OH's structural haziness. The best example of it is the 'swung dash replacement' problem: many OH entries have sub-headwords, such as compounds or past participles, which the conventional swung dash replaces sometimes fully, and sometimes only partially, in an unpredictable fashion (see also below, description of the headword field).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

et al. 1990), on the ground that it is space-consuming and fails to reflect the logical structure of the original entries. We are convinced however that only this format—as opposed to logical/hierarchical databases relying on inheritance mechanisms—is flexible enough for many kinds of NLP applications, and especially for combining two distinct bilingual databases (*cf.* below).

Of course great stress was laid on bringing the formats of the two dictionaries as close to each other as possible, since our aim was to combine them into a single, richer and more exhaustive MTD: DEFIDIC. The main guidelines of that merging process are described later in this paper, but a more in-depth discussion is available in Dufour (1997).

This entailed among others harmonizing such types of information as part of speech and subject domain labels, and establishing a common classification of collocates. On the whole, however, the two dictionaries turned out to be strikingly similar both in the kind of information they provide and the way they provide it. This can probably be seen as a vindication of RC's collocation-oriented approach[2], which was first introduced in 1978 and is obviously the inspiration of the 1994 first edition of OH.

They are almost identical in size as well, the final DEFI versions of CR and OH numbering respectively 211,017 and 208,141 records. The following sections provide a description of the main types of information (*fields*) featured by DEFI records.

3. The DEFI format: NLP-oriented bilingual dictionaries

Dictionaries in DEFI format are intended for use by NLP systems. As such they are free from the space and structural constraints imposed on human-oriented paper dictionaries, and can be processed and re-formatted at will to suit the needs of any particular application. DEFI dictionaries are not full-blown 'databases' in the computational/relational sense of the word, since they remain nothing more than flat ASCII files with blank lines separating bunches of codified information-bearing lines. They can be better described as 'data collections', which developers of NLP systems can easily sift through in order to create the exact database

[2] For more details on the collocational resources of CR see Fontenelle (1997a, 1997b).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

they want. All our work has been geared towards automatic translation selection, and a few choices had to be made that might prove a handicap in other fields of endeavour. On the whole, however, we always tried to maintain as much information as possible even when it was of no immediate use to the project.

The following paragraphs give a detailed account of the various (22) information fields present in DEFI dictionary records. Not all 22 fields are always to be found together of course, since some mutually exclude each other.

3.1. Record identification: record number and origin

All DEFI records have a record identification number, which hardly needs commenting on, but more importantly an *origin* field. This field keeps track of the origin of the lexical information used by the NLP applications, which allows easier result checking and comparative studies of the two dictionaries. Origins are either `ohef`, `rcef` or, in the combined DEFIDIC, `efm` (*English-French merge*) for records that have been created using elements from both dictionaries.

3.2. Describing the source item: headword, lemma, lemmatype

The ***headword***, a remnant of the original entry structure, is the word by which entries are accessed in the print versions of the dictionaries. While it is almost always a single word in CR, OH often uses lexicalized compounds (such as *medical advice, cabin trunk* or *school age*) as headwords. OH also has 'sub-headwords', which govern only part of the entry attached to a main headword. Such sub-headwords are mainly lexicalized past participles (*accepted*), plurals (*levels*) and capitalizations (*Exile* in the Biblical sense). In accordance with widespread lexicographical usage, both CR and OH substitute a swung dash (~) for the headword in the body of multi-word lexemes (MWLs) and example sentences. This substitution was unfortunate from the point view of the DEFI team, who had to reverse it in order to get clear-cut and complete lexemes to work with. From this point of view OH's more complex headword distribution proved to be a drawback, all the more so because it is not always consistent: the dash may or may not stand for only one word in a compound headword, or for the canonical form of conjugated verbs. Suffice it to consider the following contrastive pairs:

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

accepted: in the ~ sense of the word
allotted: his ~ted task

medical advice: to seek ~ advice
medical opinion: ~ is divided

Headwords have been 'kept aboard' mostly for reasons of tractability, as we did not expect to use them in our translation selection programs. We later found out, however, that keeping track of the headword within a multi-word lexeme can be extremely useful. Identifying MWLs in running text is a prerequisite to their translation, and it requires comparing content words found in the text with those found within the lexeme in the dictionary. Unfortunately a complete overlap can never be required, since many MWLs are recorded as examples in the dictionary and thus contain words that are actually not lexically relevant. For such text/dictionary comparisons, clearly, identifying the headword as a 'vital' (as opposed to 'optional') part of the MWL is of paramount importance. Consider for example the MWL *to ring a bell*, which appears in the OH entry for *ring* in the examples *that name rings a bell* and *that number rings a bell*. In both cases, finding *ring* in the source text will go further towards identifying the MWL than finding *name* or *number*.

The **lemma** is the English item to be translated, and as such stands at the core of each record. Lemmas can be either identical to the headword (single-word lexemes), lexicalized phrases or example sentences. From the very first stages of the project we took the option of keeping absolutely all lemmas, even though thousands of example sentences will probably never be used by our look-up programs. The look-up programs are designed to translate single-word lexemes (SWLs) and to identify/ translate MWLs, and are therefore not in a position to exploit complex example lemmas, such as the following, that are meant mainly to illustrate usage (@ signals the location of the re-inserted headword):

he has a certain artistic @ability
the building stands at an @angle on the street
I like the @cut of his coat
this horse is @lame in one leg

These unwieldy lemmas were kept mainly for the reason that there is no safe way to distinguish them from slightly manipulated MWLs, such as *that name rings a bell*, which are of course invaluable. On the other hand this was perfectly in keeping with our maximalist approach, which dictated that all the information that could be kept, should be. These

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

lemmas are now used intensively by a 'language in use' interface[3] allowing human users to query DEFIDIC via keywords. That interface, which was developed as a by-product of DEFI, allows users to search the dictionary like an 'artificial' corpus, listing in seconds all lemmas sharing any number of keywords.

Lemmas are one area in the dictionary where information elision is omnipresent. Consider the following lemma/translation pairs:

to bear / take the @brunt of - être le plus touché par
to meet with / have an @accident - avoir un accident
@almost finished/cooked/cold - presque / à peu près terminé/cuit/froid
I'd like an @alarm call (for...) - je voudrais être réveillé (à...)
'@accommodation (to let)' - 'appartements *mpl or* chambres *fpl* à louer'
to be @like sb/sth - ressembler à qn/qch
judg(e)ment - jugement
(road) @accident figures/statistics - chiffres *mpl* / statistiques *fpl* des accidents de la route

Note that in the last example above the bracketing of the implicit information '(road)' is not reflected in the target language—implying even more strongly that accident figures are, by default, *road* acccident figures. Such 'split' lemmas are not fit for any NLP application—and especially not for the elaboration at a later stage of our Prolog database, which includes parsing all MWLs. All records whose lemmas featured forward slashes or brackets had to be split, creating two or more disambiguated records. So the last of the above examples produced the following:

road @accident figures - chiffres *mpl* des accidents de la route
road @accident statistics - statistiques *fpl* des accidents de la route
@accident figures - chiffres *mpl* des accidents de la route
@accident statistics - statistiques *fpl* des accidents de la route

Parentheses could mostly be handled by programs, but slashes (which are present in a not negligible 5 % of all lemmas) are another kettle of fish: it is impossible to determine automatically how many words on either side of the slash must be kept in each translation and lemma variant, and indeed the division is not always symmetrical (consider *to meet with / have an accident*, where *meet with* is the counterpart of *have*). Furthermore, divisions in the lemma and in the translation do not always match. Consider the following example, where the translation introduces a division that is not present in the lemma:

@almost finished/cooked/cold - presque / à peu près terminé/cuit/froid

[3] For more details please check (http://engdep1.philo.ulg.ac.be/michiels/lkpuser.htm).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

Note that this configuration does not produce six records: the lemma is split in three, and all three translations keep a *presque/à peu près* division. We decided not to 'clean' the translations since they, like other types of information in the target language, are not meant to be worked upon but simply provided to the user as results. Lemma splitting was performed semi-automatically, by means of a program that presented the user with the default 1-1 division (one word on either side of the slash). The user could then accept the proposed division, add words on either side via hot keys or edit the lemma and perform the division by hand. Translations were split in a similar fashion in a second pass. This decompacting process would be highly objectionable from a human—and a publisher's—point of view, as it adds about 12,000 entries to each dictionary with no net gain in lexical information. However, we think that it is absolutely necessary in order to get the cleanest possible database, and it is a prerequisite to the recognition of many MWLs, which often offer variants even in their canonical form (consider *to bear/take the brunt*, an item of information that would be lost if left in its original compacted form).

The classification of lemmas into ***lemmatypes*** was originally an OH feature, OH lemmas being tagged in different ways according to the kind of lexical information they were perceived to carry (example, compound, frozen structure, idiom...). Note that lemmatype distinctions are present only in the SGML tape, and do not translate into typographic differences in the printed dictionary. There are 8 different lemmatypes in OH:

— `standard`, when the lemma is a single-word lexeme identical with the headword.

— `compound`, for compound nouns, including lexicalized Adj + N constructs such as *abject poverty* and *open access*.

— `example`, according to the OH markup documentation, describes lemmas 'used for illustrations of a word or phrase'. These lemmas include example sentences, but also NPs and infinitive phrases centered on the headword. The distinction between `example` lemmas and other kinds of multi-word lemmas is fuzzy at best.

— `struc` and `obligstruc` lemmas are more or less frozen multi-word expressions, `obligstruc` signalling theoretically a greater resistance to lexeme manipulation. The classification of a lemma under any of these two labels, however, is not a safe indication that the lemma is really a 'canonical' multi-word lexeme. Many truncated example sentences (e.g., *it is @absurd that*) are recorded as `struc`, and even `obligstruc` lemmas are sometimes full-blown sentences (e.g., *I can't*

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

@abide sth/doing). While these lemmas do illustrate typical uses of the corresponding headword (such as the combination *can* + NEG + *abide*), they cannot be safely regarded as anything else than examples.

— `idiom` lemmas are idiomatic expressions, either in their canonical form or in use, such as *to try one's @level best to do sth* or *it's a @licence to print money*. As with `struc` and `obligstruc`, the presence among 'idioms' of full sentences makes it impossible to consider them as anything else than examples.

— `usage`, for all practical purposes, is indistinguishable from `example`.

— `phrasalverb` and `prepverb` indicate particle verbs, which are regarded as prepositional (*prepverb*) rather than phrasal (*phrasalverb*) if verb and particle cannot be separated.

The CR raw file, being only a typesetting tape, did not draw such distinctions. For the sake of homogeneity we re-created the most relevant ones, namely `standard` (as in OH), `example` (default value for all lemmas distinct from the headword), `compound` (based on the part of speech *comp* and including compound verbs, e.g., *backdate*), `phrasalverb` (part of speech *vtr sep*) and `prepverb` (part of speech *vtr fus* and multi-word *vi*).

3.3. Grammar and syntax: part of speech, phrasal verb pattern, prepositional and clausal environment, count/mass status

Almost all lemmas are attributed a ***part of speech*** (`pos`), which, in the case of multi-word lemmas, applies only to the corresponding headword. POS-less lemmas fall into the following categories:

— OH `idiom` lemmas, because idioms are grouped at the end of entries and the headword within them can have any of the various POS represented in the entry.

— In both dictionaries, all lemmas that do not appear in a fully-fledged 'entry', but whose headword listing is followed directly by a cross-reference to another form. This concerns elided forms such as *'em, ain't* or *'twould* (reference to *them, am/are/is/have/has not, it would*), or spelling variants (*Soudanese* → *Sudanese*).

— Both dictionaries, and CR much more frequently than OH, sometimes 'forget' to provide acronyms and abbreviations with a part of speech: *AD, AGR, Leics., MA*... There seems to be no rational explanation for that, especially since most such lemmas do receive a POS.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

Part of speech indications are vital in restricting the scope of translation possibilities of single-word lemmas: since we work on tagged and parsed[4] texts, the look-up programs start by giving a higher score to translations that match the POS found in the source text. On the other hand that preference score should not be too high either, since the success rate of parsers—notwithstanding their developers' claims—rarely exceeds 90 %. We have endeavoured to harmonize CR and OH parts of speech as far as possible, either by simply adapting the POS codes of the one on those of the other, or by giving up distinctions present in only one of the dictionaries (e.g., by giving up RC's *vt sep* and *vt fus* phrasal verb POS, which was made possible by the presence of equivalent information in the phrasal verb patterns, see below). The number of different POS codes has been reduced as well, eliminating 'marginal' POS such as *n abbr* (nominal abbreviation) or *n inv* (invariable noun). There are 56 different POS left in the combined OH-CR dictionary. Of these many could still be called 'marginal', since the 9 POS codes with more than 3,000 occurrences each account for over 96 % of all records (including those with empty POS fields). Nevertheless we decided not to push our simplification efforts too far, leaving the choice of further simplifications/customizations to NLP system developers.

Phrasal verb patterns are numerical codes, ranging from 1 to 7, that inform the user about the complementation capacities of phrasal verbs: what kind of object they can take, and whether that object may be inserted between verb and particle. Verb patterns were originally displayed in full text at the beginning of OH phrasal verb sub-entries, as in:

account for: ~ for [sb/sth]
add in: ~ [sth] in, ~ in [sth]
add to: ~ [sth]
take away: ~ [sb/sth] away, ~ away [sb/sth]
take back: ~ back [sth], ~ [sth] back

The pattern for intransitive phrasal verbs was the 'zero' pattern, where the rest of the sub-entry followed directly the first mention of the verb. The resulting seven pattern codes are as follows:

1. no object insertion, typical object is non-human
2. no object insertion, typical object is human
3. no object insertion, no object restriction

[4] The parser we use is ENGCG, a constraint-grammar parser developed at the University of Helsinki and marketed by LingSoft.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

4. intransitive
5. object insertion allowed, typical object is non-human
6. object insertion allowed, typical object is human
7. object insertion allowed, no object restriction.

CR did not feature such formalized object restrictions for phrasal verbs, so that its patterns codes are either 3 (POS = *vt fus*), 4 (*vi*) or 7 (*vt sep*). Human/non-human object restrictions, it must be admitted, will probably be of little use for DEFI. Indeed their full exploitation, which would entail understanding the nature of the object before disambiguating the verb, requires a degree of semantic comprehension and world knowledge which no NLP system has yet achieved on a sufficient scale.

Knowing the typical ***prepositional or clausal environment*** of a word (as associated with a given translation), on the contrary, can be invaluable. The so-called `envir` field lists the prepositions and the types of clauses (infinitive, gerunds, that-clauses) that are likely to follow the word when it is used with a particular meaning—and thus often a particular translation. A few examples of environmental information items as they would appear in the paper dictionaries:

ability: capacité (*to do* de faire)
bother: se donner la peine (*to do* de faire)
absolve: absoudre (*of, from* de)
balance: équilibre (*between* entre)
to be afraid: avoir peur (*of* de; *to do, of doing* de faire)

The presence of the required environment in the source text is a very safe translation selection factor. Consider the following short sentence, which was used to test the DEFI look-up programs:

But why bother to object?

The user-selected word *bother* was tagged as a noun by our parser, and that tagging caused our look-up system to give an undeserved 20-point headstart to the preference scores of all nominal translations of *bother*. *Bother* in the sense of 'find the will and courage (to do sth)', however, is recorded in our dictionary as having an infinitive clause as typical complement (*cf. supra*: *to do*, of course, is understood to stand for all *to + inf* clauses). Such a clause was identified in the source text, which gave the verbal translation a 50-point bonus and made it the final choice (all other verbal translations—referring to the meaning 'annoy, pester'—received no marks at all).

Finally, the ***count/mass status*** refers to the countable or uncountable quality of nouns. Its potential use is quite easy to see: given that

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

some nouns have different meanings and translations whether they are countable or not (e.g., *reason*, which means 'motive' as a count noun and 'common sense, understanding' as a mass noun), finding them in the plural in the source text automatically restricts the number of possible translations. Countability is of course the norm, so that count/mass fields are quite rare: in the combined CR-OH dictionary, they occur in about 2.5 % of all entries. Predictably, most of these are specifications of *un*-countability: only OH actually tags some nouns as explicitly 'countable', and this only by contrast with uncountable versions of the same nouns, e.g., *disorder* as 'lack of order' or as 'illness'. While the specification of countability may be rightly regarded as lexicographical overkill for most NLP applications (which would regard it as the default value anyway), it should be borne in mind that OH, like CR, was intended for human users whose attention should indeed be drawn to such contrasts. And keeping this information in, which is the option taken by the DEFI team, allows lexicographers to list in a few minutes all OH-recorded nouns sharing both statuses.

3.4. Semantic restrictions: indicator, s-feature, style register, field labels

Indicators are the main source of semantic information in the two dictionaries. Found mostly at the beginning of semantic (*i.e.*, second-level) subentries, they 'indicate' which sense of the lemma corresponds to the translations that follow. Consider the use of indicators in OH for *accommodate*:

accommodate I *vtr*
1 (*provide room, space for*) ...
2 (*adapt to*) ...
3 (*reconcile*) ...
4 (*satisfy*) ...
5 (*meet request*) ...

Most indicators are synonyms or paraphrases of the lemmas, as in all the examples above, and can be best compared to CIDE's guide words (Procter 1995). This guide-word value should not be taken for granted, however, since the indicator is probably the least formalized of all the DEFI fields. It can just as well feature usage restrictions, grammatical information or collocational information introducing a translational rather than semantic distinction. A few examples of such 'disparate' indicators:

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

flipper (*for swimmer*)
float (*on water, in air*)
floral pattern (*design: on material, wallpaper*)
greedy (*for money, power, information*)
get (*cause to be / gen {+} adj*)

Indicators are hints in natural language meant for human users, and therefore utterly unfit for computerized exploitation. Single-word indicators might be used to create a crude network of synonyms and semantically related words (*cf.* accommodate – reconcile – satisfy), and some indicators singled out for their more formal structure (*in + noun*, *of + noun*) have been turned into other types of information (domain labels, collocational restrictions). Apart from these few attempts at formalization, and from the s-features described below, indicators are geared too strongly to human users to be used by NLP systems.

S-features (semantic features) have been derived from the most frequent indicators in the two dictionaries. Different indicators conveying similar meanings were combined into a unique, more formal s-feature, for instance:

act, action, process → `act`
hen, butterfly, fish, fowl, bird → `animal`
person, people, man or woman, person in authority, student, some unspecified person → `hum`
tree, plant, fruit, vegetable, flower → `veg`

Choosing the indicators that were to produce s-features was done entirely by hand, on the basis of a list of all indicators occurring at least 10 times within either dictionary. The actual creation of s-features was then performed by a simple program running on both files. It must be admitted that s-features were created purely as a just-in-case procedure, and their semantic content remains too vague for use by our look-up programs. Their value for lexicographical research (e.g., listing all names of animals provided by the dictionaries) is dubious as well, since they depend on the somewhat haphazard distribution of indicators and the frequency requirements which we imposed, and thus can never lay claim to a comprehensive coverage of any topic (in the combined DEFIDIC, only 1 % of all records are provided with an s-feature).

Style labels, which are a traditional feature of dictionaries in general, are particularly important in learner's dictionaries, and to a lesser extent in bilingual dictionaries. In both cases it is vital that the user, who is not familiar with the language used, be made aware of the—especially negative—connotations carried by the word he/she plans to use. The

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

reason why this is less vital in bilingual dictionaries being, obviously, that the said connotation should appear clearly in the translation. While our two dictionaries provided style labels for both lemmas and translations (though not systematically so), we took the option of storing only source-language labels into a separate field. Target-language labels have been left where they were found, namely within the translations, where they retain a special tagging making them easily identifiable for future use. Some harmonization of OH and CR labels was required (e.g., `fml` and `frml`), yielding the following main DEFIDIC labels: `Br`, `US`, `Can`, `Austral`, `Scot`, `arch`, `babytalk`, `coll`, `dated`, `dial`, `euph`, `fml`, `hum`, `iro`, `liter`, `offensive`, `pej`, `sl`, `vulg`. The inclusion of regional distinctions (`Br`, `US`, `Can`) is debatable: they could probably have been regarded as field labels—since they restrict the 'contexts' in which the words are used—or stored into a separate 'geographic' field. We do not plan to make great use of them anyway, since regional distinctions are highly unreliable. As a matter of fact style labels as a whole are of little interest to the DEFI team, and this for two reasons.

The first one is that style labels do not always contribute to word sense distinction: a word like *fuck* is vulgar in all its uses, and a system programmed to choose the more formal sense (e.g., because it is meant to run on scholarly papers or serious newspaper articles) will be none the wiser for that. Even style labels that do signal a semantic distinction (e.g., `coll` with *bake* in the sense of 'lie lazily in the sun' or 'feel very hot') cannot really be exploited, because that distinction is mostly a slight one (`coll` as opposed to default) and the stylistically 'marked' sense is likely to be used—be it in a jocular way—in all contexts but the most formal. The style issue admittedly gets trickier when we deal with (sexual) slang words, such as *screw*, *cock* or *bitch*, which are often everyday words used with a very substandard meaning. Even for such words, though, style taggings are not reliable enough—witness *bitch*, which is more widely used in its derogatory slang sense than as 'female dog'.

The second, more down-to-earth reason for not making intensive use of style labels is lack of consensus: style labels are the result of individual lexicographers' choices, and some are both quite severe and inconsistent. A striking example of this can be found in the CR entry for *God*, where many everyday expressions based on 'God' are tagged as `sl`, probably because 'thou shalt not...', while some are only `coll` and others are neutral. This means that many words/word senses which the two dictionaries classify as far below standard may be regarded as quite

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

acceptable by the authors of our corpus texts, and thus that our system cannot reject them on the sole ground of their stylistic scarlet letter.

While style labels could have little use other than that of rejecting taboo words, ***field labels*** (or *domain labels*) offer the possibility of a more positive discrimination. Field labels, a staple of all monolingual and bilingual dictionaries, are used to specify in which particular domain(s) of activity a word takes on a given sense/translation. Consider the following field specifications for the adjective *compound*, as they are found in DEFIDIC:

compound (*ling, gram*): composé
compound (*ling, gen*): complexe
compound (*bio, bot, chem, gen*): composé
compound (*med*): multiple
compound (*math*): complexe
compound (*med*): compliqué
compound (*tech*): compound

Note that the apparent redundancies (some translations and labels appear twice) are due mainly to the presence of other discriminating elements in these records, and to the fact that these records stem from two different dictionaries (*cf.* below). One can easily imagine the potential of field labels, which can help the look-up system choose the translation that best fits the domain of the text worked on—*compound* (speaking of fractures) as *multiple* or *compliqué*, say, in an article from a medical journal. This presupposes of course that the system has been made 'aware' of the subject domain of that text. That information could quite easily be provided by the user or, with more uncertainty, determined by the system itself after a bootstrapping process consisting of a statistical analysis of the labels associated with all the words in the text (*cf.* Amsler and Walker 1986, Jansen 1989). Text categorization is not really part of the project, however, and we always assume that the system has been given the subject of the text by a human.

There are 268 different field labels in DEFIDIC, a number that remains too high even after harmonization. One reason for that superabundance of field labels is lack of formalization: although both dictionaries claim to list all their field labels in their prefatory material, new labels have apparently been created in the heat of the moment by individual lexicographers and apply only to a few records. This is particularly true of CR, where field labels have no specific tagging and were crammed with other types of information into the italics-between-brackets typographical mold—their only peculiarity being capitalization. And to top it all, we

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

decided to derive a few more from indicators after noticing that these are often labels in hiding. To achieve this we performed a frequency analysis of indicators similar to the one mentioned above for the creation of s-features, but applying only to indicators matching the pattern *in X*. The minimal required frequency was set to 4, yielding results like *in voting, in church, in class, in grammar, in religious matters, in engraving, in television programme, in race/election*, etc. Semantically relevant *in X* indicators produced a common new label, for instance:

in boarding school, in primary school,
in secondary school, in school/college → `scol`
in divorce, in will, in assizes court, in law,
from point of view of law → `jur`
in skiing → `skiing`
in tennis → `tennis`
in soccer → `soccer`

We also added to the field labels several tags that had no obvious classification, namely `gen` (general use, the zero-label), `fig` (figurative use), `lit` (literal meaning) and `r` (registered trademark). It could easily be argued that these are not really 'domain' labels, but on the other hand we can say that they (especially `gen`) help specify the contexts in which the words are used (`r` could be interpreted as meaning 'commercial'). The `gen` label's main task is often to temper the influence of other labels. While gen alone brings no new information to a record (zero-label), finding it within a list of other labels gives these labels a more 'indicative' value. Consider the 'compound' examples above: while *compliqué* and *multiple* translate *compound* **only** in its medical sense, *composé* is a general translation of *compound* that is **also** used for its biological, botanical and chemical senses. Indeed, `gen` was added to CR label lists to translate the *also* found in the original tape. Consider the following example from the CR *abstract* entry:

abstract (*also Chem: remove*) extraire

The content of the parentheses was reformatted into the label list `chem, gen` and an indicator field containing *remove*.

Many labels found in DEFIDIC are only marginal, *i.e.* they apply to such a small number of records that their contribution is insignificant. Of course they could easily be found out and dropped, and maintaining them is part of our general wait-and-see attitude. A better solution than dropping them, anyway, will probably be to regroup them: many of these labels are subordinates of other, more important labels that could be substituted for them (`sport` instead of `fencing, boxing, climbing, rowing`

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

and `wrestling`, for instance). Whatever the option we finally take, it will be necessary to instill some kind of semantic networking into our list of labels. Imagine an NLP system dealing with a text about archaeology (`archeo`) and having to choose between three translations of a word, the only discriminating element available being the label `hist` (history) associated with one of them. While a human user will instantly go for the 'historical' translation, a system that has not been taught to link related labels will be left to its own perplexed devices. To parry this we plan to establish (by hand) a number of 'semantic distances' separating or linking the main labels of the dictionary, letting our look-up program know for instance that `hist` is quite close to `archeo`, and indeed much closer than to `mus` (music).

3.5. Collocational restrictions: subject and object collocates, NP heads

Collocational restrictions, which made CR's fame among lexicographers if not among ill-informed students, are probably the most vital items of information provided by both OH and CR. The principle underlying the inclusion of collocate lists within dictionaries is that collocations play a major role in selecting the meaning of a word. To quote the ever so famous Firthian axiom: 'words shall be known by the company they keep'. Building on this, both OH and CR include in their metalinguistic apparatus lists of words (collocates) that typically appear together with the lemma in relation with a given translation. Consider the following examples (collocates are in italics, either alone or in square brackets, according to the CR typographical conventions):

cut *vt*

joint of meat découper; *abscess* inciser; *tobacco* découper; *steps* tailler; *channel* creuser, percer; *figure, statue* sculpter; *jewel, key, glass, crystal* tailler; *screw* fileter; *dress* couper; *hedge, trees* tailler; *corn, hay* faucher; *lawn* tondre; *class* manquer, sécher; *appointment* manquer exprès; *profits, wages* réduire, diminuer; *text, book, play* réduire, faire des coupures dans; *person* blesser (profondément); *cards* couper.

branch *n*

[tree, candelabra] branche; *[river]* bras; *[mountain chain]* ramification; *[road]* embranchement; *[railway]* bifurcation; *[pipe]* branchement; *[family]* ramification, branche; *[subject, science etc]* branche; *[store]* succursale; *[company]* succursale, branche; *[bank]* agence, succursale; *[police force]* antenne.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

What makes collocates even more vital for translation selection is that, more than just signalling semantic differences, they are often linked to translational differences that are independent of meaning distinctions in the source language. While Longman's Dictionary of English Language and Culture [Summers 1992] lists only two senses for the verb *screech*, collocational restrictions account for five translations in CR:

screech *vi*

[person] pousser des cris stridents, hurler; *[brakes]* grincer; *[tyres]* crisser; *[singer, owl]* crier; *[siren]* hurler.

There are six different kinds of collocates in CR and OH, whose interpretation depends partly on their typographical presentation and partly on the part of speech of the lemma they collocate with. Dictionaries in the DEFI format have only two collocate fields, which correspond to the original typographical distinctions in OH: `precoll` ('subject' collocates, which appeared to the left of translations in OH) and `postcoll` ('object' collocates, which appeared to the right of translations). While the field names refer to typographical conventions, the subject/object categorization relates more closely to the nature of the collocates in question.

Subject collocates are those deemed to be the 'major members' of their collocate/lemma pairs, *i.e.*:

— typical subjects of verbs (bracketed italics in CR),
— typical nouns modified by an adjective (naked italics in CR),
— typical adjectives modified by an adverb (naked italics in CR),
— typical verbs modified by an adverb (naked italics in CR).

Object collocates, on the contrary, are modifying elements:

— typical objects of verbs (naked italics in CR),
— modifying nouns in compounds (*N2* in *N1* of *N2*, or *N1* in *N1N2*, bracketed italics in CR).

Not all OH collocate lists were tagged as such in the original dictionary files. As with field labels, but this time only in OH, we soon discovered that *indicator* fields were a treasure trove of hidden collocational information. Consider the following simplified abstract from the OH entry for *branch*, to be compared with the CR *branch* entry above (all bracketed information was tagged as 'indicator' in the SGML file):

branch *n*

(of tree) branche; (of pipe, road, railway) embranchement; (of river) bras; (of candlestick, lamp) branche; (of antlers) ramure; (of family, language) rameau; (of study, subject) domaine; (of shop) succursale; (of bank) agence; (of company) filiale ...

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

The similarity between these indicators and the collocate lists in CR is striking, and it makes no doubt that nominal modifiers in OH were systematically tagged as *of+N* indicators. Re-formatting all such indicators into `postcoll` fields added a total of over 6,000 collocates to the OH DEFI dictionary.

Heads are a particular kind of subject collocates appearing most often in OH. They are the typical NP heads of nouns listed as modifiers (part of speech `modif`). So for instance, in OH:

abortion **I** *n* ... **II** *modif* [*law, debate*] sur l'avortement; [*rights*] à l'avortement; [*pill*] abortif/-ive.

The CR raw tape had no `modif` part of speech, and correspondingly no collocates directly interpretable as heads. We found out, however, that some lemmas listed as compounds were actually nominal modifiers, and thus that their collocates were heads. This is the case when the 'compound' lemma is actually not a compound, but the single-word headword in its modifier usage. Consider the following extract from the *Easter* entry:

Easter **1 n** ... **2 comp** *egg* de Pâques; *holidays* pascal, de Pâques; **Easter bonnet** chapeau de printemps ...

Easter bonnet is the first compound listed in its full form. In the first two 'compounds' of that list the lemma is each time *Easter*, and *egg* and *holidays* are NP heads in *Easter egg* and *Easter holidays*. Such half-compounds have been re-written as `modif` all over CR, and their collocates as heads.

Heads have not been stored together with the other subject collocates because the bond that unites them to their lemmas is much stronger. Translations of a *modif* lemma are not actually translations of that lemma, but of that lemma *when it appears as modifier in a compound whose head is one of those listed*. Consider the example of *abortion* as *sur l'avortement*: translating *abortion* as *sur l'avortement* is only possible if it appears in a compound where it means 'about/over abortion', like *abortion law* (*loi sur l'avortement*), *abortion debate* (*débat sur l'avortement*) or, possibly, *abortion bill* (*projet de loi sur l'avortement*). This close relationship means that our look-up systems will not even take *modif* lemmas into account if no head is found in the source text.

DEFIDIC boasts a total of 139,996 collocates or heads, spread over 76,967 different lists. In DEFIDIC 40 % of all records featuring a polysemous single-word lemma—the type of lemma that most urgently needs collocational restrictions—are provided with at least one collocate list.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

Homage be paid here to the contribution of OH, which managed to surpass its older brother's already impressive collocational wealth: OH has 93,942 collocates or heads in 49,753 lists, compared to 58,610 collocates in RC's 38,458 lists.

Many collocates/heads in the original files were unfit for computerized exploitation, either because they were pluralized or because they consisted of complex phrases. Pluralization is a sensible option for human-oriented dictionaries, where it is useful to indicate that some words typically appear in the plural (*cf. rights* in the *abortion* entry above). However this would be a serious handicap for our look-up programs, which are meant to deal with tagged, lemmatized and parsed text. We avoided that obstacle by automatically lemmatizing[5] all collocate and head lists, adding lemmatized forms at the end with a specific tagging. Here are a few examples of collocate lists containing plural forms, in their original and lemmatized versions:

conditions, circumstances → conditions, circumstances, θcondition, θcircumstance
truth, facts → truth, facts, θfact
blasphemies → blasphemies, θblasphemy
eggs, cream → eggs, cream, θegg

Collocates appearing under the guise of complex phrases could not be dealt with so easily. We decided to re-format automatically those matching a number of fixed patterns, and to leave the others in their original state. The phrase patterns that were simplified automatically are as follows:
— Determiner + noun: (*a cause, an order → cause, order*)
— Saxon genitive + noun (*person's qualities → quality*)

The potential of collocational restrictions for automated translation selection is obvious: if we find in the text that the user-selected word collocates with one of the words appearing in one of its collocate lists, the translation to be chosen is the one associated with that collocate list. Consider some of the collocational restrictions of *heavy* in CR, and the example sentences below:

heavy *adj*
weight, parcel lourd, pesant; *expense, movement* lourd; *payments, charges* important, considérable; *step* pesant, lourd; *crop* abondant, gros; *loss, fine* gros, lourd; *rain, shower* fort; *fog* épais, à couper

[5] Using ENGCG again.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

au couteau; *meal, food* lourd, indigeste; *defeat* grave; *odour* fort, lourd ...

The army has suffered a few heavy defeats recently. → heavy = *grave*
Heavy rain spoilt the second half of our holidays. → heavy = *fort*
I could recognize his heavy steps anywhere. → heavy = *pesant, lourd*
Heavy fog covered the morning landscape.
→ heavy = *épais, à couper au couteau*

A look-up system confronted with *heavy* in these four sentences will easily find the most apt translations simply by comparing the collocates of *heavy* in the text with those found in the dictionary. Of course one should keep in mind, as Fontenelle [1997a, 1997b] points out, that the collocates recorded in the dictionaries are to be understood as thesauric heads rather than as specific lexemes. A human user finding *fog* as collocate of *heavy* will always assume that the corresponding translation applies just as well to heavy *smoke, smog* and *mist*, for instance. Similarly, a heavy *meal* can be a heavy *breakfast, lunch, dinner, supper* or even *snack* (though this would be somewhat paradoxical). Since one cannot expect to find in the source text exactly the same collocates as are listed in the dictionaries, we have to provide our system with the tools needed to find out that *breakfast* is some kind of *meal*. The DEFI look-up programs use three different lexico-semantic databases in order to establish semantic relations between textual and dictionary collocates. Two of these databases are independent from the DEFI dictionaries, and will not be discussed at length here:

— WordNet (Miller *et al.* 1990) was the obvious choice, with its networks of semantic relations such as synonymy, hypernymy, and meronymy;
— Roget's Thesaurus of English Word and Phrases[6]: two words are considered as 'related' if they belong to the same Roget category.

Our third 'collocate recognition' database was derived automatically from DEFIDIC's own collocate lists, following an idea expressed by Montemagni *et al.* (1996) and based on what we call *metalinguistic slot sharing*. In that paper the authors argue that it is possible to establish the conceptual relatedness of two words using the metalinguistic information provided by dictionaries, and more specifically by comparing the contents of collocate lists. Their basic assumption can be summed up in the following, somewhat cavalier fashion: *two words that appear alongside each other in the same collocate list are likely to share certain*

[6] Raw text downloaded from the Project Gutenberg Web archives at (http:// ✂ wuarchive.wustl.edu/doc/gutenberg).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

semantic properties. The nature of that relationship can be extremely vague, and often falls outside the categories usually taken into account by semantic networks. Consider *weight* and *parcel* in the *heavy* entry above. *Weight* and *parcel* are related neither by synonymy, nor antonymy, nor hypernymy/hyponymy, nor meronymy, nor really a part-whole relation. And yet, somehow, they are similar in that they can both be said to be *heavy*—with the same meaning of *heavy*. This similarity can be easily exploited for word sense discrimination when it comes to establishing a relation between a collocate in the text and the collocates mentioned in the dictionary. Imagine a user-selected word W (*heavy*) collocating in the text with a word C (say, *load*). In order to disambiguate W, the look-up system must relate C to a C' (say, *weight*) mentioned as a typical collocate of W in the dictionary. To achieve this a system based on Montemagni *et al*'s approach will compute the number of co-occurrences of C and C' in all the collocate lists of the dictionary (in DEFIDIC's case 5 times). The higher the number of co-occurrences the stronger the link of course, and the translation considered to be the most likely will be the one whose C' co-occurs (one could say 'collocates') most often with C. This technique has already been implemented in the DEFI look-up programs, with encouraging results combining good coverage with surprisingly low noise. Consider the following collocational restrictions for *dependability* in DEFIDIC, and an example sentence extracted from John Le Carré's *The Little Drummer Girl* that was used in our preliminary tests:

dependability *n*
[equipment] fiabilité
[machine] sécurité de fonctionnement
[person] sérieux

Though again she mightn't, for under her scatty exterior she was cursed with a dependability of character that was often wasted on the company she kept.

Attempts to link *character* with the relevant dictionary collocate *person* in order to disambiguate the selected word *dependability* were made using WordNet, Roget's thesaurus and metalinguistic slot sharing. Metalinguistic slot sharing produced the best results, thanks to 63 co-occurrences of *person* and *character* in DEFIDIC's collocate lists.

Metalinguistic slot sharing computation offers the following advantages:

— Given a specifically organized database of collocate lists, the technique is computationally much less demanding than WordNet queries.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

— A database of collocate co-occurrences is much more intuitive and less formal than man-made semantic networks like WordNet. Such databases link only words that conform to a limited set of semantic relations, while collocates list combine words that, in a very precise context, share some unspecified properties that make them apt collocates of a third one.

Its potential drawback is that it requires a very high number of collocate lists to be efficient, but the combined resources of CR and OH certainly go a long way towards quenching that thirst for data.

3.6. Internal references: `xref` and `gothere`

Both dictionaries feature a ***cross-reference*** field, which will probably be of little use to our look-up programs. In the printed versions cross-references appear at the end of sub-entries, and refer the user to other entries where he/she is likely to find interesting information related to the entry now being considered. Basically, a cross-reference means nothing more than 'have a look at that other word, and you will find something interesting'. These other words are often synonyms of the one under consideration, but many are words that appear together with it in a multi-word lexeme. Here are a few examples of non-synonymic cross-references:

a → few, little, lot, many
cut → short
paint → coat, wet
panther → black

This lack of formalization hinders systematically exploiting cross-references (`xref`) by our look-up programs—although non-synonymic references might be valuable for collocation research. The `gothere` field on the other hand, which appears only in CR, refers explicitly to a synonym of the word being looked up. Consider the following `gothere` references in CR:

pantyhose → pantihose
radio contact → radiocommunication
econometrist → econometrician
railroad → railway
rambunctious → rumbustious

This will make it easy for a look-up program to follow the thread: a user who has selected *rambunctious* will get the translations of *rumbustious*, naturally with a notification of synonymy.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

3.7. If everything else fails: `tradratio`

The `tradratio` field is added to the dictionary as the element that will be used in the last resort to select one translation in preference to all others. `Tradratio` indicates how many times a given lemma has received a given translation, as compared to its total number of translations (*i.e.*, the number of records featuring that lemma and a translation, even though the same translation may appear several times). Consider the translation configuration of *cell* in OH:

cell *cellule*
cell *cellule*
cell *alvéole*
cell *élément*
cell *cellule*

All records featuring *cellule* as a translation of *cell* get a `tradratio` of 3–5, while *alvéole* and *élément* both get 1–5. It clearly appears that *cellule* is the most 'standard' translation of *cell*, and it is therefore the one that will be chosen if the system finds no other discriminating elements in the source text. This choice is debatable, and the relative translation frequency computed here does not have the legitimacy it would possess if it had been derived from corpus analysis. `Tradratio` is however a very simple way to compute automatically the 'default' translation of a word, and the relative frequency of a translation within the dictionary seems to reflect its relative frequency in the 'real world' of language use faithfully enough.

3.8. End-user information: full forms, translations, gloss

The fields mentioned here are those that are to be provided to the user by our look-up programs. They play no part in the sense discrimination process proper, of which they are simply the result.

The ***full forms*** of acronyms and abbreviations (`expand` field) complement or replace translations. They must be given to the user even when a translation is available, since translations of acronyms are often approximations and their exact meaning can only be understood by considering their full forms. Consider the following acronym / full form pairs:

A.A.U.P. → *American Association of University Professors*
CD → *Civil Defense*
GNVQ → *General National Vocational Qualification*

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

Some records featuring a full form have no translation at all, either because the full form was deemed explicit enough, or because it has its own entry elsewhere (full forms are often accompanied by cross-references):

Glos. → Gloucestershire
Jas. → James
LPN → Licensed Practical Nurse (`xref` to *license*)

Translations are of course the element we are looking for. Not all records have a translation, since it can be replaced either by a full form, a gloss or a cross-reference field. Since translations are not used for linguistic analysis, they have not been stripped of elements such as gender tags, style labels and grammatical indications. We decided to give it all to the user, who can best sort it out for him- or herself. ***Glosses*** are paraphrases or explanations in French of terms that cannot be translated. Many lemmas requiring a gloss are cultural or political terms that have no conceptual equivalent in the French-speaking world, for instance:

floater → *personne dont le rôle est de circuler dans une soirée*
punchball → *variante simplifiée du baseball, qui se joue sans batte*
fun run → *course à pied pour amateurs, souvent organisée pour collecter des fonds*
ginger beer → *boisson légèrement alcoolisée à base de gingembre*
pepperoni → *saucisson sec de porc et de bœuf très poivré*

Many glosses in the original files were used to complement existing but too imprecise translations. Such glosses received a gloss tagging in OH, while they were included within the translations in CR. CR complementary glosses are still included in the translations (with special markers $\perp$ and $_{\top}$), those stemming from OH have been stored separately like all other glosses.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

4. Merging the two DEFI dictionaries: *e pluribus unum*

While the two DEFI dictionaries were processed separately, our aim has always been to merge them into a single larger, richer bilingual database. The obvious drawback of such a merger is that it is bound to generate a cumbersome amount of redundancy: the two dictionaries are twin brothers in many respects, and just putting them together would provide the same information twice for many single-word lexemes. This section gives an overview of the strategies we used in order to eliminate that redundancy, creating a merged dictionary that was, in size and number of records, 'smaller but better than the sum of its parts': DEFIDIC.

One of the first options we took was *not* to get rid of 'internal' redundancy, *i.e.* of redundancy occurring within the same dictionary. From a translation selection perspective, internal redundancy may be of two kinds: exact repetitions introduced for user-friendliness, or near-repetition denoting semantic distinctions irrelevant to translation selection. The exact repetition of multi-word lexemes at different places in the dictionary is as a help for human users: an MWL *like to be conspicuous by one's absence*, for instance, is found both under *conspicuous* and under *absence* in OH. This is redundancy in the strictest sense of the word, since the two occurrences are equivalent and a computerized look-up system does not need two access keys for the same lemma. The greatest source of internal redundancy, however, is the repetition of the same lemma/translation pair caused by differences in the metalinguistic information. Consider the following *absolute* records in CR (only the relevant fields are mentioned):

```
LEMMA=|absolute
POS=|adj
INDICATOR=|unlimited
PRECOLL=|monarch
TRANS=|absolu


LEMMA=|absolute
POS=|adj
INDICATOR=|unqualified
PRECOLL=|refusal, command, majority
TRANS=|absolu
```

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

```
LEMMA=|absolute
POS=|adj
LABELS=|math, phys
PRECOLL=|value, temperature, zero
TRANS=|absolu
```

These records account for three slightly different meanings of *absolute*, and thus the repetition of *absolute/absolu* cannot really be called redundancy. For the purposes of computerized translation selection, however, it might be expedient to regroup these three records into a single one:

```
LEMMA=|absolute
POS=|adj
INDICATOR=|unlimited / unqualified
LABELS=|math, phys, gen
PRECOLL=|value, temperature, zero, refusal, command, majority, monarch
TRANS=|absolu
```

Still the elimination of internal redundancy is irrelevant for the merging of our two DEFI dictionaries (it should be performed last, taking DEFIDIC as basis), and we chose to 'wait and see' rather than taking the risk of oversimplifying our lexical database. Furthermore, internal redundancy (of any kind) is actually quite seldom: a program meant to deal with it has been written and tested, but was not able to reduce the number of records in DEFIDIC by more than 3 %.

Since internal redundancy is not dealt with, merging the two dictionaries means merging *pairs* of records coming the one from OH, the other from CR. Merged records inherit and combine the information from their two 'parents', and are added to the resulting dictionary file, while all records that find no equivalent in the other dictionary are added to the new file as such. The following example illustrates the merging of two *absolute* records:

```
IDNUM=|2                          IDNUM=|2
HEADWORD=|absolute                HEADWORD=|absolute
LEMMA=|absolute                   LEMMA=|absolute
LEMMATYPE=|standard               LEMMATYPE=|standard
POS=|adj                          POS=|adj
INDICATOR=|unlimited              LABELS=|pol
PRECOLL=|monarch                  PRECOLL=|monarch, power
TRANS=|absolu                     TRANS=|absolu
ORIGIN=|rcef                      ORIGIN=|ohef
```

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

```
IDNUM=|2
HEADWORD=|absolute
LEMMA=|absolute
LEMMATYPE=|standard
POS=|adj
LABELS=|pol, gen
INDICATOR=|unlimited
PRECOLL=|monarch, power
TRANS=|absolu
ORIGIN=|efm
```

The *sine qua non* condition for merging two records is that their lemmas be perfectly identical. Our merging program thus compares batches of OH and CR records sharing a given lemma and computes, for all possible record pairs between these batches, their 'merging potential'[7]—taking into account such criteria as translation overlap and the compatibility/equivalence of other fields. Record pairs with a merging potential higher than zero are then merged on a 'best match first' principle, as illustrated in the following merge array (fig. 1) for all records featuring *advance* as lemma.

In its present version, our merging program processes about 32 % of the records of each dictionary, creating 65,080 merged records out of two dictionaries with 210,000-odd records each. The resulting DEFIDIC file is made up of 354,078 records, instead of the 419,158 a straightforward concatenation would create. Although the merging rate of 32 % might not seem high, and although the final space gain is only half that figure, there is little doubt as to the necessity and expediency of the merging process. Merged records are mainly single-word lemmas, which are the most polysemic and whose meaning is thus the most difficult to select in a dictionary. Merging such records creates richer ones, notably with longer collocate lists increasing the efficiency of metalinguistic slot sharing computation. The bulk of the entries that could not be merged consists of more complex lemmas and example sentences, and their variety increases the lexical wealth of DEFIDIC.

[7] For more details about the computation of merging potentials please refer to Dufour (1997).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

	oh1	oh2	oh3	oh4	oh5	oh6	oh7	oh8	oh9	oh10	oh11	oh12	oh13	oh14	oh15	oh16	oh17
cr1			1														
cr2		3															
cr3	3		1														
cr4																	
cr5	3		4														
cr6					3	3	15			3	3						
cr7					54	3	3			3	3						
cr8																	
cr9																	
cr10																	
cr11																	
cr12					3	3	3			9	3						
cr13																	
cr14																	
cr15																	
cr16																	
cr17																	
cr18												2					
cr19													9	3			
cr20																	
cr21															36		
cr22																	
cr23																	
cr24																12	

Fig. 1.– Empty cells indicate a merging potential of zero, resulting for instance from incompatible parts of speech or translations. Darkened cells indicate pairs of records that were actually merged, such as oh1 and cr3 (the first *advance* record in OH with the third *advance* record in CR).

5. Conclusions

DEFIDIC is the best we could get out of two excellent machine-readable dictionaries, and we are convinced that its potential for translation selection will be demonstrated in the later phases of the project. One of its main advantages is certainly its size: by combining the strengths of two dictionaries into one, it allows more extensive lexicographical research and is a safer basis for NLP applications. Another strong point is its flexibility: although it is specifically intended for translation selection, it can be re-formatted with a few lines of codes and made to match the needs of a wide range of applications. Its domain labels, for instance, have been used as the backbone of a prototype for text categorization and

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

vocabulary analysis. Leaving aside NLP applications, the data it contains is of paramount importance for more fundamental lexicographic work like collocational research or the creation of semantic networks [*cf.* Fontenelle (1997)].

Bibliography

BAUER (Daniel), SEGOND (Frédérique) and ZAENEN (Annie): 1994, "Enriching an SGML-Tagged Bilingual Dictionary for Machine-Aided Comprehension", Rank Xerox Research Center Technical Report, MLTT, 11, Meylan.

BAUER (Daniel), SEGOND (Frédérique) and ZAENEN (Annie): 1995, "Locolex: the translation rolls off your tongue", Rank Xerox Research Center Technical Report, MLTT, Meylan.

BOGURAEV (Branimir), BRISCOE (Ted), CARROLL (John) and COPESTAKE (Ann): 1990, "Database Models for Compuational Lexicography", in *EURALEX'90 Proceedings*, Barcelona, Biblograf, p. 59–78.

CORRÉARD (Marie-Hélène), and GRUNDY (Valérie), eds: 1994, *Oxford-Hachette French Dictionary* (Oxford: O.U.P.).

DUFOUR (Nicolas): 1996, "Turning the Oxford-Hachette SGML tape into a DEFI dictionary", DEFI technical report, Liège, available from http:// ✂ engdep1.philo.ulg.ac.be/michiels/defi.htm.

DUFOUR (Nicolas): 1997, "Merging two DEFI dictionaries", DEFI technical report, Liège, available from http://engdep1.philo.ulg.ac.be/michiels/defi.htm.

DUVAL (A.) and SINCLAIR (L.S.), eds.: 1993, *Collins-Robert French Dictionary* (Glasgow: HarperCollins).

FONTENELLE (Thierry): 1997a, "Using a bilingual dictionary to create semantic networks", *International Journal of Lexicography, X*, 4 (forthcoming).

FONTENELLE (Thierry): 1997b, *Turning a bilingual dictionary into a lexical-semantic database* (Tübingen: Max Niemeyer Verlag, Lexicographica Series Maior) (forthcoming).

GUO (Cheng Ming), ed: 1995, *Machine Tractable Dictionaries: Design and Construction* (Norwood, N.J.: Ablex Publishing Co.).

JANSEN (Jacques): 1989, "Apport contrastif des dictionnaires généraux de la langue au problème de l'indexation automatique dans le discours techno-scientifique", *Meta, XXXIV*, 3.

MILLER (George A.), BECKWITH (Richard), FELLBAUM (Christiane), GROSS (Derek) and MILLER (Katherine J.): 1990, "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography, III*, 4, p. 235–244.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

MONTEMAGNI (Simonetta), FEDERICI (Stefano) and PIRRELLI (Vito): 1996, "Example- based Word Sense Disambiguation: a Paradigm-driven Approach", in *EURALEX'96 Proceedings*, University of Göteborg, p. 151–159.

PROCTER (Paul), ed.: 1978, *Longman Dictionary of Contemporary English* (2nd Edition edited by Della Summers) (Harlow: Longman Group Ltd.).

PROCTER (Paul), ed.: 1995, *Cambridge International Dictionary of English* (Cambridge: Cambridge University Press).

SINCLAIR (John), ed.: 1987, *Collins COBUILD English Language Dictionary* (Glasgow: HarperCollins).

WALKER (D. E.) and AMSLER (R. A.): 1986, "The use of machine-readable dictionaries in sublanguage analysis". In GRISHMAN (L. R.) and KITTREDGE (R.), eds: *Analyzing language in restricted domains.* (Hillsdale, N.J.: Lawrence Erlbaum).

WILKS (Yorick A.), SLATOR (Brian M.) and GUTHRIE (Louise M.): 1996, *Electric Words – Dictionaries, Computers, and Meanings* (Cambridge, Ma – London: MIT Press).

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.