[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

DEFI, a Tool for Automatic Multi-Word Unit Recognition, Meaning Assignment and Translation Selection

Archibald Michiels & Nicolas Dufour
University of Liège
Place Cockerill, 3,
B-4000 Liège, BELGIUM
[amichiels@ulg.ac.be]

Abstract

This paper provides a brief description of the DEFI project, a project aiming at word sense discrimination and translation selection in English-French machine-tractable dictionaries. The guiding principles are described, and preliminary results presented and commented in some detail.

The DEFI Project

DEFI is a five-year basic research project in the field of word sense discrimination and translation selection. It started in October 1995, and is therefore due do end in 2000. The project's general objective is to create a prototype that would provide the reader of a text in a foreign language *(in casu* English) with the best possible translation *(in casu* into French) of any word he/she selects online, depending on its environment in the source text. The look-up system, a Prolog-based prototype, works as a 'text-dictionary matcher' that attempts to find the lexical database entry (featuring the appropriate 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 are 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, Segond and Zaenen, 1995), albeit with a different, more semantically oriented approach.

Textual Data

The textual data provided to the DEFI matcher consist of 'textual chunks', i.e. syntactically independent bits of text extracted according to major punctuation barriers. Each textual chunk is submitted to LingSoft's ENGCG surface parser[1], whose results are reformatted and enhanced by various heuristics before being turned into Prolog structures and fed in to the matcher. In addition to the parser's output, the *txt-clauses* (textual clauses) contain the following pieces of information:

- a list of NPs;
- a list of syntactic attachments;
- a polarity value (negative/affirmative);
- a voice (active/passive);
- a structural hypothesis (NP, VP, PP, whole clause...).

Note that textual chunks are normally always 'clauses', other values being mainly of interest for the treatment of multi-word expressions in our dictionary (cf. below).

Lexical Data

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. DEFl's lexical resources are the following:

- the complete Collins-Robert and Oxford-Hachette English/French dictionaries (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 learner's dictionaries of English (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 now 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.

The two bilingual dictionaries (amounting to 2,000 pages of small print in their 'paper' versions) make up the main lexical resource used by the matchers. They were provided to us by their owners in the form of two typesetting tapes, and the transformation of these unwieldy files into a single machine-tractable dictionary —Defidic— has taken up a good part of the first eighteen months of the project. In its present form Defidic has about 350,000 'records', each record consisting of a source-target pair plus all the

[1] ENGCG was developed at the University of Helsinki and is marketed by LingSoft Inc. (www.lingsoft.fi).

relevant linguistic and metalinguistic information (part of speech, grammatical environment, semantic markers, field labels, collocational constraints, etc.) provided by the original dictionaries.

The bilingual lexical data base used by our text-dictionary matchers is made up of two Prolog trees, one for single-word lexemes (about 130,000 records) and one for multi-word units (henceforth MWUs, 220,000 records). The structure of our database for single-word lexemes is rather straightforward, being nothing more sophisticated than a Prolog-readable representation of all the information present in Defidic.

For the purposes of the project, any source item in our dictionary that features more than one word is regarded as a 'multi-word unit'. MWUs thus range from compound nouns and phrasal verbs to proverbs and example sentences. As part of their re-formatting into a Prolog database, all MWUs are submitted to the same parser which we use to prepare our textual data, and submitted to the same enrichment process on the basis of the parser's output. Note that the access to MWUs in our Prolog database is not dependent on their storing order in the original 'paper' dictionaries: since we cannot expect the user to know which word to select, and since MWU storing conventions are always too complex and rarely observed to the letter by lexicographers themselves, the matcher's database contains an additional index tree allowing the retrieval of each MWU via any of its content words (including prepositions).

A well-known approach to MWU tagging is the finite-state technique adopted by Rank Xerox for their LOCOLEX machine: each MWU in the dictionary is provided —by hand— with a local grammar specifying the location and nature of each component, and the places where the MWU is likely to receive external modifiers (adjectives, adverbs) in natural text. This technique has the advantage of being very fast at run time, but its drawbacks are numerous. First, it requires extensive human intervention in the coding of MWUs, and inter-annotator agreement is admittedly low. Second, strict local grammars are incompatible with dictionary MWUs listed only as example sentences —a frequent situation in bilingual dictionaries, since the 'canonical' (i.e., infinitive) form of verbal idioms is not always translatable as such (cf. *to miss sb*, which has no direct translation in French; the best way to illustrate that MWU is through an example: *I miss you* - 'tu me manques'). Finally, we do not think that such *ad hoc* local grammars are flexible enough. Natural language plays around with MWUs to an extent which lexicographers tend to underestimate, adding modifiers where nobody expects them and using idiomatic expressions in non-idiomatic contexts. All these reasons explain why MWUs in DEFI'S databases have no 'grammar' of their own: they are simply described as they were recorded in the original dictionaries, and the task of dealing with variations, expected and unexpected, is left to the matcher.

The matcher also draws upon the resources of three 'thesauric' databases for the exploitation of the collocational constraints provided by the bilingual dictionaries:

- WordNet (Prolog package);
- Roget's thesaurus (as downloaded from *Project Gutenberg* Web site[2]);
- a database providing, for each pair of dictionary collocates, the number of slots they share in Defidic; this database allows us to put into practice the hypothesis put forward by Montemagni *et al.* 1996.

Main Working Principles

The DEFI matcher attempts to match the user-selected word in three phases:

- as a terminological item, or as part of one;
- as part of a general-language MWU;
- as a single-word lexeme;

If the word is found in the relevant terminological data base (for our test bed, a database covering the field of Aegean archaeology), the 'terminological' translation is returned to the user and the matching process stops. This is by far the simplest case, for which we assume that the text being considered belongs to a certain specialized field where terminological items are not used with their possible 'everyday' meaning.

In a second stage, the selected word is regarded as part of a multi-word expression and the system tries to find oat which dictionary MWU best matches the textual chunk around the selected word.

In a third stage, and only if the first two matching procedures have failed, the selected word and its (grammatical and collocational) environment in the user's text are matched against the single-word entries of our database.

An important issue in the matching procedure is the choice between MWU and single-word lexeme treatment: how is the system to determine that it should look for an MWU? how can we have it decide that the search for MWUs has lasted long enough and should give way to a single-word type of analysis? We have so far avoided confronting this problem by having two matchers instead of one: the single-word matcher only looks for single-words, and the MWU matcher for MWUs. Test sentences are fed in to one or the other matcher according to which is deemed more appropriate in each case. In the future we will have to find ways of solving that 'decision' problem automatically, but so far the strict division has allowed us to work separately on the problems of translation selection (for single-word items) and MWU recognition (translation

[2] http://wuarchive.wustl.edu/doc/gutenberg

selection is rarely a problem for MWUs, with the notable exception of phrasal verbs).

The working principles of the DEFI matchers can only be summed up here, as the present paper is regarded primarily as a report on preliminary results and a complement to the system's demonstration. More in-depth description can be found in Michiels 1998 and Dufour 1998b.

Commented Results: MWUs

In this section we present and discuss various test results for multi-word expressions. The first sub-section deals with a series of tests revolving around much-discussed idioms containing *brunt* and *havoc,* while the second emphasizes various points of interest on the basis of example sentences extracted from literary and journalistic sources.

Bearing the Brunt of... the Havoc

Looking for multi-word units containing *brunt* in Defidic (i.e., in two mid-size general language dictionaries), one comes up with the following results:

the brunt
to take the brunt of
to bear the brunt of
to bear the brunt of the assault
to bear the brunt of the expense
to bear the brunt of the work

Apart from the stand-alone np *the brunt,* idioms containing *brunt* would thus seem to conform to the following pattern:

bear	the	brunt	of
take			

A random extraction of about 80 instances of *brunt* from the BNC, however, shows the need to extend it to the following:

bear	the	main	brunt	of
take	a	full		—
carry	—	heaviest		
catch		real		
feel		considerable		
face				

Bear and *take* are the most frequent verbs, but they are clearly semantically depleted and can be replaced at will by any verb meaning 'face', 'be confronted with'. Similarly, although *the brunt* could be expected to be a frozen entity, it is subject to seemingly endless manipulations —the determiner was dropped in a newspaper headline, a case local grammar writers can hardly afford to take into account. The prepositional phrase coming after *brunt* is very often dropped, and the whole idiom is frequently passivized: these are all modifications that are not recorded as such in the dictionary, and that could not be dealt with if the dictionary MWUs were forced into the straitjacket of a local grammar. Note moreover that *brunt* does not appear as a single word in the dictionary, so that the matcher must be able to make one of the MWUs match whatever textual chunk it has to face. In this case the 'mock MWU' *the brunt* will match any instance of *brunt,* because the matcher has a grammar rule according to which any determiner matches any determiner, including zero —a heresy to proponents of local grammars, although such laxness has thus far not been shown to provoke noise.

The DEFI matcher readily deals with all manipulations and/or additions of adjectives and determiners, since they are all allowed by its internal grammar, as is the omission of the preposition (in this case *of)* found at the end of dictionary MWUs. A sentence featuring *catch the brunt* will return *to take the brunt of,* since both *catch* and *take* belong to the matcher's lists of semantically depleted verbs (together with *get, put, set, come...)* It cannot however tackle the substitution of *bear* or *take* by *carry/feel/face,* since none of them belongs to these lists, and will return only *the brunt* in such cases. Another —and more serious— weakness is its inability to deal with passivization or topicalization: the matcher compares dictionary MWUs with the textual chunk from left to right only, and is not yet able to identify superficial structure changes as such. Here are a few examples of raw results (the first number is the preference score), with *brunt* as selected word:

1.Walter Zenga in goal, Franco Baresi the libero, and Gianluca Vialli in attack, will carry the brunt of the responsibility today.
127- the brunt = le (plus gros du) choc
123- the brunt = le poids

2. The doctor took the full brunt of Moran 's resentment.
260- to take the brunt of = être le plus touché par
260- to take the brunt of = subir le plus fort de
260- to take the brunt of = subir tout le poids de
153- the brunt = le poids
123- the brunt = le (plus gros du) choc

3...forced on us here in London , who will certainly bear the brunt.
209- to bear the brunt of = être le plus touché par
209- to bear the brunt of = subir le plus fort de
209- to bear the brunt of = subir tout le poids de
123- the brunt = le (plus gros du) choc
123- the brunt = le poids

4.and that the real brunt of the war was being borne by the men on the battlefield.
131- the brunt = le (plus gros du) choc
127- the brunt = le poids

5.Sussex bore a considerable brunt of the next stage of the Conquest and was the first area to be systematically "Normanised".
221 - to bear the brunt of = être le plus touché par
221- to bear the brunt of = subir le plus fort de
221 - to bear the brunt of = subir tout le poids de
87- the brunt = le (plus gros du) choc
87- the brunt = le poids

6. The Prince has a fierce temper, which Colborne often caught the brunt of.
198- to take the brunt of = être le plus touché par
198- to take the brunt of = subir le plus fort de
198- to take the brunt of = subir tout le poids de
123- the brunt = le (plus gros du) choc
123- the brunt = le poids

7.He commanded the only Indian Parachute Battalion (152) which in March 1944 bore the brunt of the Japanese assault east from the River Chindwin .
396- to bear the brunt of the assault = soutenir / essuyer le plus fort de l'attaque
260- to bear the brunt of = être le plus touché par
260- to bear the brunt of = subir le plus fort de
260- to bear the brunt of = subir tout le poids de

8.MAINFRAME BUSINESS BEARS BRUNT OF SWINGEING NEW IBM CUTS.
211- to bear the brunt of = être le plus touché par
211- to bear the brunt of = subir le plus fort de
211- to bear the brunt of = subir tout le poids de
87- the brunt = le (plus gros du) choc
87- the brunt = le poids

A similar comparison of MWUs containing *havoc* in Defidic and the BNC yields the following results, in the same order (dictionary occurrences - dictionary pattern - rough corpus pattern):

it wrought havoc
this wreaked havoc with their plans
to cause havoc
to make havoc of
to play havoc with
to wreak havoc in
to wreak havoc on sth
to wreak havoc on
to wreak havoc

wreak	havoc	— with on
cause	havoc	—
play	havoc	with
make	havoc	of

wreak	— such	horrible financial more	havoc	— in on
		untold seasonal		with among against to
play	—	—	havoc	with
cause	— some such	—	havoc	— at to for in around
make	— much	—	havoc	of
create	— such	—	havoc	— in
do	such	—	hayoc	—

As in the case of *brunt*, it seems that the choice of verbs taking *havoc* as object is far less limited than is recorded in the dictionary. Except perhaps in the case of *play havoc with*, where the choice of 'play' may induce a slight change in meaning (introducing the idea of purpose?), the verb is relatively irrelevant —*wreak* and *cause* are more frequent, and seem to combine in a restricted collocation with *havoc*. Similarly, the presence of determiners or adjectives, as well as the type of prepositional phrase that can follow *havoc,* are hardly predictable. The noun havoc was more than once found to function as antecedent of of a relative clause (*... and the havoc his hooligans wrought),* and passivization is relatively frequent as well.

The performance of the DEFI matcher in identifying MWUs containing *havoc* is similar to that recorded for *brunt:* adjectives and determiners (or quantifiers) are never a problem, but differences in the nature of the support verb or in word order are not yet tackled.

9. Claiming that no-fault divorce was first introduced by Nazi Germany and that it had since wreaked "more havoc on the Allied countries than any German army or air force ever did".
135- to wreak havoc on sth = dévaster qch
108- to wreak havoc on = dévaster

10.Countless independent "free house" owners have copied brewers' fashions and wrought untold havoc with unassuming old country pubs before moving on to pastures new.
Note: the failure of the matcher in this case (no result at all is returned), as often, is due to a weakness of the otherwise robust ENGCG parser: *wrought* is lemmatized only as *work.* Note that we consider the parser's results as a given, which we try to enrich but never to correct.

11. Whatever the Navy's intentions, their shells were landing in the Commando positions and causing some havoc.
127- to cause havoc = provoquer des dégâts
127- to cause havoc = tout mettre sens dessus dessous

12. Worrying not so much about the effects of heavy-duty exercise on their bodily contours, but with the havoc it is wreaking on their faces.
No results due to word order.

13.Direct sunlight plays havoc with the varnish.
213- to play havoc with = désorganiser complètement # chambouler
213- to play havoc with = abîmer, bousiller {coll}

14.Claret, chocolate or turkey curry stains can wreak seasonal havoc.
164- to wreak havoc = faire des ravages, dévaster # infliger des dégâts

The last two examples prove the point: for a lexicographer to decide intuitively which variations an idiom can undergo is impossible in practice. Moreover, in a recognition perspective, the expediency of such severity is also questionable: what is the need of specifying that a given variation is impossible? If it is really impossible it will not occur anyway, and if it occurs then it must be possible. It might be argued that beyond a certain level of manipulation, an idiom should not be read as an idiom any more, and the literal meaning should be forwarded. This is doubtful, however, since native speakers tend to play with idioms and to 'disguise' them while hoping that their readers/hearers will get the hint anyway (cf. Michiels, 1998; Dufour, 1998a).

Miscellaneous MWU results

Results presented in this subsection were extracted from a battery of tests based on articles from *The Economist* newspaper and on extracts of John Le Carré's novel *The Little Drummer Girl.* **Clwlist** is the list of words (mostly just one) selected by the user online.

Note that 'peripheral' results (i.e., results with relatively low preference scores) may differ according to the selected word, as shown by the first two examples.

l.The summer of 1995 may be remembered as the moment when Heisenberg Tourism achieved a sort of global critical mass.

Clwlist - [critical]
152- critical mass = masse {f} critique
65- to be critical of = critiquer, trouver à redire à

Clwlist = [mass]
152- critical mass = masse {f} critique
87- the mass = la masse, le peuple, les masses
65- masses of = des masses de {coll}, des tas de {coll}
65- the masses = la foule
65- the masses = les masses {fpl}
65- the masses = la masse, le peuple, les masses populaires
60- Mass. = nil

Example 2 shows how 'to be X' MWUs, in which an adjectival phrase is inserted in an infinitive clause for ease of translation through rephrasing, are recognized outside infinitive clauses *(plain sailing* is of course the MWU we are looking for here):

2.For a Grand Old Man of Letters it had become fairly plain sailing.
Clwlist = [plain]
169- to be plain sailing = marcher comme sur des roulettes
70- she's rather plain = elle a un visage quelconque, elle n'a rien d'une beauté
70- she's very plain = elle a un visage ingrat, elle n'a rien d'une beauté
65- it's plain madness = c'est pure folie, c'est de la folie toute pure
65- the Plains = les Grandes Plaines
65- the Plains = les Prairies {fpl}, la Grande Prairie

Example 3 shows how, in exceptional cases, a noun phrase can be 'ignored' in dictionary MWUs (in this case, because *whiff* is recognized as the major element, while *chloroform, garlic* etc. are little more than slot fillers). The user is thus provided with examples that do not exactly match his/her sentence, but will nonetheless help him/her understand the text. In the first result, 'get' matches 'has', both being semantically depleted.

3.And while this may be true, it also has the whiff of elitism, not to mention a thinly disguised hostility to those who are less than adept with computers.
Clwlist = [whiff]
198- to get a whiff of = sentir l'odeur de
191- a whiff of chloroform = une bouffée / petite dose de chloroforme
191- a whiff of garlic = une bouffée d'ail
191- a whiff of seaweed = une bouffée de varech
82- what a whiff! = ce que ça sent mauvais!

In 4 we are confronted with MWU polysemy: the MWU 'to come unstuck' is identified perfectly, but the machine cannot decide between its proper or figurative sense *(se décoller* vs. *tomber à l'eau).*

4.Attempts to introduce western political models into poor countries have a habit of coming unstuck.
Clwlist = [unstuck]
110- to come unstuck = se décoller
110- to come unstuck = tomber à l'eau {coll}

In 5 a whole NP is ignored in the dictionary MWU, allowing the matcher to return *the issue at stake* as a twin brother of *the other interests at stake* (because *stake* is semantically heavier).

5.Depending on the egregiousness of the offence and the other interests at stake, supporting human rights may mean anything from armed intervention to a statement in parliament.
Clwlist = [stake]

152- the issue at stake = ce dont il s'agit, ce qui est en jeu, ce qui se joue ici
150- to be at stake = être en jeu

Just as 'to be X' clauses are reduced to 'X', any dictionary MWU matching the pattern Pron+be+X will fit 'X' alone on the textual side:

6.1n November 1996 they retained most of their winnings in both houses, no mean feat.
Clwlist = [feat]
184- it was no mean feat = cela a été un veritable exploit, ce n'a pas été un mince exploit
184- that's no mean feat! = ce n'est pas un mince exploit!
87- a feat of = une prouesse de

In the same way that parser errors are inevitable, in many cases our dictionary just doesn't provide the right translations. In the following example, *to go soft* is translated only as 'perdre la boule', which means 'to go nuts' rather than 'to become too indulgent'. Note that the second choice, *you're too soft!,* is much closer to the intended meaning.

7.It is also to be lamented because if the Republicans go soft, there is no earthly reason why the Democrats, all decked out as they are in borrowed and often ill-fitting Republican clothes, should keep their resolution for a moment longer.
Clwlist = [soft]
152- to go soft = perdre la boule {coll}
93- you're too soft! = tu es trop indulgent! / trop bon!
77- he's soft = c'est une mauviette / un mollasson, il n'a pas de nerf
77- the brakes are soft = il y a du mou dans la pédale de freins
77- the market is soft = le marché est lourd

Commented Results: single words

The DEFI matcher for single word lexemes is much simpler than the MWU matcher, since it does not have to deal with structure on the dictionary side. This matcher compares the selected word with all the relevant entries in the dictionary, and simply chooses those whose metalinguistic information best matches its environment in the user's text. The types of information taken into account so far are the following:

- Part of speech;
- grammatical (i.e. clausal and prepositional) environment;
- collocational constraints;

While the exploitation of parts of speech and the grammatical environment is rather straightforward (apart from parser errors regarding parts of speech, of course), collocational constraints are dodgier: we cannot expect to find in the user's text exactly the collocates listed in the dictionary, so that various strategies are implemented to establish semantic relationships between collocates in the text and collocates in the dictionary. In brief, we use 3 different resources:

- *WordNet:* a search through the WordNet taxonomy is the most obvious way to find out that two words are related.
- *Roget's* thesaurus: two words are related if they belong to the same Roget category, with various degrees of closeness according to their closeness within the category.
- *Metalinguistic slot sharing* (cf. Montemagni *et al.*, 1996; Dufour, 1998b): two words are regarded as semantically close if they appear together in one or more of the 38,000 multi-collocate lists of Defidic. The underlying assumption is that words that are 'listed' together as collocates (preferably several times) must share some semantic properties.

There are still many cases when none of these strategies really works, and in the cases of highly polysemous words the results can be disastrous. Consider the following example, where the correct translation would have been 'facile' *(soft* meaning 'easy'):

1.The countries singled out for a bashing are often soft targets, like Myanmar, which offer few economic opportunities and have little power to hit back.
Clw = soft
31 - soft = flasque, avachi
31 - soft = (trop) indulgent
28 - soft = instable à la baisse
20 - soft = doux # doux/douce
20 - soft = doux # doux/douce
20 - soft = stupide, bête, débile {coll}
20 - soft = doux [{f} douce], moelleux
20 - soft = souple, doux
20 - soft = doux, fin, satiné # doux/douce
20 - soft = doux [{f} douce], léger
20 - soft = doux, mélodieux, harmonieux # doux/ douce
20 -, soft = tendre
20 - soft = mou, (r)amolli # mou/molle
20 - soft = souple
20 - soft = tendre, compatissant
20 - soft = doux, estompé, flou
20 - soft = ouaté, feutré
20 - soft = mou [{f} molle] {pej}
20 - soft = mou
20 - soft = mou, malléable
20 - soft = doux, tendre
20 - soft = mou
20 - soft = doux, soyeux, satiné
20 - soft = doux
20 - soft = soyeux
20 - soft = doux [{f} douce], aimable, gentil
20 - soft = aimable, gentil
20 - soft = doux, facile, tranquille
20 - soft = facile
20 - soft = doux [{f} douce], léger
20 - soft = doux
20 - soft = doux [{f} douce], pâle
20 - soft = doux, pastel {inv}
20 - soft = faible

20 - soft = mollasson, qui manque de nerf
20 - soft = meuble
20 - soft = lourd
20 - soft = léger/-ère
20 - soft = doux/douce
20 - soft = apaisant
20 - soft = diplomatique
20 - soft = modéré
20 - soft = privilégié
20 - soft = doux/douce
20 - soft = peinard {coll}
20 - soft = trouillard {coll}

There are of course many cases where collocational constraints work much better, providing the relevant translation with a preference score so high that other possibilities, lagging too far behind, are not even displayed:

2.They will almost certainly not go to war and they are generally reluctant to disrupt trade.
Clw = disrupt
339 - disrupt = perturber

3.He's like a greedy child grabbing all the cakes on the plate.
Clw = greedy
194 - greedy = gourmand

4.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.
Clw = dependability
279 - dependability = sérieux {m}

A few other examples of successful collocational constraint analyses:

[war]
5.Dictatorships unleashed the first and second world wars, and most wars before and since.
Clw = unleashed
243 - unleash = déclencher
69 - unleash = libérer
69 - unleash = lancer
51 - unleash = déchaîner

[criminal, person, case]
6.All very well, the sceptics reply, but even with a global economy the world is not a global country with a global set of laws, a global police force to enforce them and a global judiciary to try wrongdoers.
Clw = try
94 - try = juger
42 - try = mettre à l'épreuve, éprouver
42 - try = prendre {qn} à l'essai
42 - try = demander à
35 - try = essayer
[...]

7.he lacked the elitist background from the kibbutzim, the universities and the crack regiments that, to his dismay, increasingly supplied the narrowing aristocracy of his service.
Clw = crack
250 - crack = d'élite

Part of speech matching is used primarily to prune away totally irrelevant translations. A perfect POS match brings a score of 20 (i.e., not a lot), an absolute mismatch brings 0 (often preventing the display of the corresponding translations). This can spell disaster in cases of parser error, like in the following example (where *bear* is parsed as noun only):

8.But they will not, in Jack Kennedy's words, "pay any price, bear any burden" to promote liberty.
Clw = bear
22 - bear = ours {m} (mal léché)
22 - bear = ours(e) {m(f)}
22 - bear = baissier {m}

In this case the erroneous parse of *bear* also prevents the recognition of *burden* as object of *bear* by our parse enriching programs, although it would have made the choice of the best translation all but certain. In the following example *don* is parsed both as a noun and a verb, thus causing the matcher to return the translation of *don* as Oxbridge professor alongside *don* 'put on'.

9. In the highlands of Papua New Guinea, in a village near Goroka, the warriors, all but naked, smear their bodies with a pale mud and don surreal mud masks.
Clw = don
22 - don = professeur {m} d'université -surtout à Oxford et à Cambridge-
22 - don = don {m}
22 - don = mettre
20 - don = chef {m} de la Mafia
20 - don = revêtir, mettre

The analysis of grammatical (prepositional) environments often helps to forward just one or two translations of a polysemous word, and can in some cases compensate for parser errors. This is the case in the first of the following examples, where *bother* was parsed as a noun (a perfect preposition match is worth 50):

10.But why bother to object ?
Clw = bother
50 - bother = se donner la peine
20 - bother = ennui {m}, embêtement {coll} {m}
20 - bother = ennuis {mpl}
20 - bother = casse-pieds {coll} {mf inv}, enquiquineur/-euse {coll} {m/f}
20 - bother = ennui {m}, barbe {coll} {f}, scie {coll} {f}
20 - bother = mal {m}

11.Devotion to tax cuts at any price (one reason for the present rift between Mr Gingrich and his party) is not a virtue.

Clw = rift
70 - rift = désaccord {m}
70 - rift = rapture {f}
20 - rift = trouée {f}, éclaircie {f}
20 - rift = fissure {f}, fente {f}, crevasse {f}
20 - rift = haut-fond {m} créant des rapides
20 - rift = division {f}
20 - rift = division, désaccord

As with *to go soft*, there are still many cases where *Defidic* (despite its size) just does not feature the right translation. For instance in example 12, where *conceit* means 'delusion, naive belief but is translated only in its senses of 'self-conceit' or 'elaborate metaphor'.

12.The idea that tourism inevitably strips off some holiness of place, some magic, may be descended from the primitive conceit that a camera steals the soul of the person photographed.

Clw = conceit
22 - conceit = suffisance {f}, vanité {f}, prétention {f}
20 - conceit = afféterie {f} {literary}
20 - conceit = métaphore {f} élaborée
20 - conceit = compliment {m}
20 -conceit = trait {m} d'esprit, expression brillante

References

Bauer, D., Segond, F. and Zaenen, A. (1994). Enriching an SGML-Tagged Bilingual Dictionary for Machine-Aided Comprehension. Rank Xerox Research Center Technical Report, MLTT, 11, Meylan.

Bauer, D., Segond, F. and Zaenen, A. (1995). Locolex: the translation rolls off your tongue, Rank Xerox Research Center Technical Report, MLTT, Meylan.

Corréard, M-H and Grundy, V., eds (1994). *Oxford-Hachette French Dictionary*. Oxford: O.U.P.

Dufour, N. (1998a). A Database for Computerized Multi-Word Unit Recognition. In *Proceedings of the 3rd International Symposium on Phraseology,* Stuttgart, forthcoming.

Dufour, N. (1998b). Matching Collocational Constraints for Translation Selection: Defi's Combined Approach. In *Proceedings of Euralex'98*, Liège, forthcoming.

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

Michiels, A. (1998). The Defi Matcher. In *Euralex'98 Proceedings.* Liège, forthcoming.

Miller, G. A., *et al.* (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography, III*, 4, pp. 235-244.

Montemagni, S. *et al.* (1996). Example-based Word Sense Disambiguation: a Paradigm-driven Approach. In *EURALEX"96 Proceedings*, University of Goteborg, *pp.* 151-159.

Procter, P. ed. (1978). *Longman Dictionary of Contemporary English.* (2nd Edition edited by Delia ; Summers) (Harlow: Longman Group Ltd.).

Procter, P. ed(1995). *Cambridge International Dictionary of English.* Cambridge: C.U.P.

Sinclair, J. ed. (1987). *Collins COBUILD English Language Dictionary.* Glasgow: HarperCollins.