LEXDIS, a tool to measure lexical proximity

A. Michiels, University of Liège, Belgium (amichiels@ulg.ac.be)

The measure of the lexical distance between items *x* and *y* can be conceived of as the computing of the number and strength of the lexical links between *x* and *y*. Such links can hold both on the horizontal, syntagmatic axis and on the vertical, paradigmatic one. An example of the former is the link between *grenadier* and *parade* as exemplified in *He was standing there like a grenadier on parade*, link which can be exploited to work out the word senses to assign to both items (grenadier: soldier, not fish[1]; parade, military, not shopping mall[2]). An example of the latter link is that between *ritual* and *ceremony*, link which can be used in the selection of the right word sense of *go through* in *She was expected to go through such a daily ritual*, since *ceremony* is a member of the collocate list for the prepositional argument of *go through* under the reading 'perform_rehearse' of *go through*, and the strength of such a link exceeds that of the other links that hold between *ritual* and the other members of the collocate lists, be they associated with the 'perform_rehearse' word sense or with any of the other readings assigned by the dictionary used in the disambiguation task. We shall come back to both these examples in more detail in the remainder of this paper.

The hypothesis underlying LEXDIS is that such **lexical** links can be explored, and their strength assessed, on the basis of the **lexicographical** links that can be established by exploring a sufficient body of lexicographical tools, dictionaries and thesauri, and specific, task-oriented data bases derived from these basic tools.

As a matter of fact, we would be hard put to define lexical distance without referring to lexicographical links. We know that lexical similarity, e.g. as measured by humans in the Rubenstein-Goodenough word pairs, is not the whole story. It tends to privilege paradigmatic relations over syntagmatic ones, making it easier to relate *car/train* than *car/journey*, or *jury/decision*, or *fury/ fit*.

LEXDIS performs well on the Rubenstein-Goodenough word pairs, achieving a correlation factor of 0.869 with the human assessments (Pearson correlation index), but surely the business of such a tool as LEXDIS is not to ape humans in similarity assessment games, but to be a building block in the solving of various tasks, such as word sense assignment (monolingual) or target equivalent selection (bilingual or multilingual). For practical purposes, we can do with a **nearer-than** relation, so that the weights returned by LEXDIS need not be meaningful in isolation, which they aren't.

There has been a lot of research recently on tools for the computation of lexical distance (see Budanitsky et al. 2006 and the references there; Banerjee and Pedersen 2003; Li et al. 2003) and it is easy to understand why. The exploitation of the metalinguistic information provided by dictionaries for purposes such as translation relies on the user being able to assess the lexical distance of a word in the text he is confronted with with pieces of information often boiling down to lists of lexical items presented as candidates for being 'the nearest equivalent' of the textual item. Consider *collocate lists* and *indicators*. They are undoubtedly among the most valuable metalinguistic information items provided by monolingual and bilingual dictionaries to enable the

[1] From *semdic*, a LEXDIS dictionary : mono(lem(**grenadier**), ori(wn), idnum('grenadier%1:05:00::'), pos(n), lab([]), gw([]), deflex([rattail, 'rattail fish', 'deep-sea', fish, large, head, body, tapering, tail]), exlex([]), def('**deep-sea fish with a large head and body and long tapering tail**')).

[2] From *semdic* : mono(lem(**parade**), ori(ci), idnum(ci52657), pos(n), lab([]), gw([]), deflex([row, shops]), exlex([]), def(' **a parade is also a row of shops**')).

user to distinguish between the various word senses or translations associated with a given lexical item. Most of the time, though, the user's text does not display as head of the relevant syntactic slot (subject, object, etc.) a morphosyntactic word whose lemmatization yields a member of one of the dictionary's collocate lists for that syntactic position. The reader (or the computer program) has to go through the items to be found in the collocate lists associated with the relevant syntactic position, measuring the proximity of the textual element to each of the members of every collocate list. The winner (the selected word sense or translation) should be the bearer of the collocate list one of whose members features the best proximity measure with respect to the textual item.

In a very large number of cases, indicators and collocates are the **only** information made available by dictionaries to enable the reader to distinguish among a range of word senses or translations. It stands to reason that NLP systems that are unable to use that information will equally be bound to stop short of an analysis sufficiently fine-grained to satisfy the requirements of a large number of applications, among which, very obviously, machine or machine-assisted translation.

An example will help to clarify matters. Using **Defidic** (an in-house merge of the Robert-Collins and Oxford-Hachette English-French dictionaries), in

(1) *From his beloved sister Fatmeh, who had given him the gilded* ***charm*** *to* ***wear*** *round his neck*

(John Le Carré*, The Little Drummer Girl,* Pan Books edition, 1984, p.210)

the analysis of the pair WEAR-CHARM should lead to the selection of the translations **porter** or **mettre** (which both display *jewellery* in their associated collocate lists for the object) rather than **arborer**, **accepter** or **user** whose collocate lists do not provide as good a proximity match with *charm* as does *jewellery.* At the same time, *charm* itself can be disambiguated on the basis of the word sense providing the match, i.e. *ci11034* in our CIDE database, namely:

a small esp gold or silver object worn on a chain as jewellery

Computing a lexical proximity factor also comes in useful in relating **indicators** and **collocates**. Consider 2:

(2) *And now that little bogy had been exorcized with the rest!*

(Angus Wilson, *Hemlock and After*, Penguin ed., pp. 9-10 ; French tr. by Marie Tadié (1954), *La ciguë et après*, Robert Laffont, 10-18, domaine étranger, p. 8 : *Et voilà que ce petit* ***croquemitaine****-là avait été exorcisé comme les autres!*)

In order to translate *bog(e)y* by **croquemitaine** (in preference to **crotte de nez**, **bogée**, **épouvantail**, **démon**, **spectre** and **bête noire**, the other translations offered by DEFIDIC), we need to explore the collocate list for the object of *exorcize* (a single list), and attempt to find the adequate reading/translation for *bogey* by measuring how well the indicators for *bogey* (namely **in nose** for *crotte de nez*, **in golf** for *bogée*, **frightening** for *épouvantail/démon*, **evil spirit** for *croquemitaine*, **imagined fear** for *spectre*, and **bugbear** for *bête noire*) match against the collocates for the object of *exorcize* (a single, three-item list: **demon**, **memory**, **past**). The task is simple enough for the human user, but we claim that there is no single lexical resource that can be consulted in order to assess the quality of the match, assessment which proves indispensable for the selection of the right translation, in so far as in the case of the *exorcize-bogey* pair the indicators and collocates are the only elements provided by the dictionaries to discriminate between the six candidate translations for *bogey.* The winner is :

```
QUERY : [demon, pos:n, spirit, pos:n, w:none, m:g].
demon with POS=n is related to spirit with POS=n with weight=7.42019 as follows:
Shared Labels: [my] -> weight: 4
Shared words in definition: [considered, force, activity, believed, peoples, energy] -> weight: 38
Cooccurrence in Roget's thesaurus  -> weight: 1
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/5.795
```

Here, the query concerns the two nouns *demon* and *spirit*, no minimal weight is specified (w:none), the mode is global (m:g), which means that we are attempting to match lexemes and not word senses. LEXDIS tells us that a proximity factor of 7.42019 links the two nouns. The strength of the link is assessed on the basis of a shared label (my=*Mythology and Legend*), six words common to the definitions of the two items, and co-occurrence in a slot of Roget's thesaurus (a broad slot, the weight being only 1) . The combined lexicographical weight of the two items leads to an adjustment with respect to less heavy items, so that the returned weight should reflect the quality of the link rather than the sheer quantity of common elements between the lexical worlds of the two items.

LEXDIS measures the lexical distance between any two English words belonging to the following POS: adj, adv, n, v. It works across POS, so that the following is a licit query: [wear, pos:v, slipper, pos:n, w:t, m:l, noadjust]. The elements of the query are Prolog variables, and need not be instantiated within the query, so that [wear, pos:POS, slipper, pos:n, w:t, m:l, noadjust], where POS is left uninstantiated, is a licit query, too.

The LEXDIS lexical resources are given in Appendix B, and the LEXDIS algorithm in Appendix C.

We claim that lexical proximity cannot be reduced to the measure of relatively well-defined relations such as those to be found in a standard thesaurus such as Roget's or in a WordNet type thesaurus. Although these relations are relevant and contribute to the total assessment of lexical proximity, they are too restrictive in the horizontal dimension (synonymy and antonymy providing the bulk of the matches) and tend to wander too far from the pivot along the vertical axis (hyponymy and hypernymy soon provide catch-all categories, or else – as in WordNet– explore a scientific terminology that is of little use for the analysis of lexical relations in the general language). We argue that lexical proximity is best conceived of as the result of an inherently heuristics-based exploration of various lexical associations derivable from (suitably massaged) available lexical resources, dictionaries as well as thesauri. The justification for building such a tool lies wholly in its discriminatory power, i.e. its power to select the right translation or word sense in context.

We have therefore opted for the use of a highly specific type of corpus, namely a corpus based on lexicographical resources, because we believe that good lexicographical resources themselves result from a careful analysis of textual corpora, and incorporate in a nutshell the best lexical information that can be derived from a study of raw textual data. An entry in a dictionary (especially a learner's dictionary, such as the three we have used, namely CIDE, LDOCE and COBUILD) can be looked at as the description of the lexical world of the word being characterized. Definitions and examples tend to capture in as short a piece of text as possible the main elements to be found in the environment of the word, both along the paradigmatic (definition) and syntagmatic (both definition and example) axes. The network of relationships captured in definitions and examples goes far beyond the exploration of the horizontal and vertical thesauric relations embodied in the listing of hyponyms, hypernyms, synonyms and antonyms. For instance, the relation between an instrument and what it is used for (or the other 'qualia' of Pustejovsky's generative lexicon – see Pustejovsky 1995) will very often be part of the information provided by definitions and examples.

It should be clear by now that if lexical distance is computed on the basis of data gathered from a textual corpus (as is the case in Church and Hanks 1989 and in the considerable body of derived and associated research), the syntagmatic axis is privileged. Such a bias is suitable for the design of lexicographer's workbenches and similar tools, but is not so helpful when the task is to measure the proximity of textual data and lexicographical metalinguistic information such as indicator and collocate lists, as in the example outlined above.

Nor is an emphasis on the paradigmatic axis more suitable for such a task, as the indicators and collocates are not to be read as thesauric heads any more than they are to be read as standing for individual lexical items. More subtly, they give an idea of the lexical 'world' in which the item described 'belongs' or 'feels at home', to use metaphors that reflect how difficult it is to pinpoint the relationship between a collocate in a dictionary collocate list and the textual items it is supposed to 'stand for'.

Among the resources tapped by LEXDIS, some are clearly paradigm-oriented and some are clearly syntagm-oriented. The syntagmatic axis is reflected in *dictionary examples*, *collocate lists* and *environments*; the paradigmatic axis comes to the fore in *indicators*, *guidewords*, [*WordNet Synsets and Relations*[3]] and *Roget's Categories*. But LEXDIS also makes use of *dictionary definitions*, and these, whenever they are well designed, offer in themselves a balance between syntagmatic and paradigmatic information, whether they espouse the traditional Aristotelian format of genus+differentiae (where the genus is in paradigmatic relation with the definiendum, the differentiae more often in the syntagmatic one) or conform to the context-setting environment of the COBUILD definition type.

LEXDIS does not relate *word forms* (the morpho-syntactic words making up a textual corpus), but either *lemmas* (global mode) or *word senses* (local mode). Whatever the ontological status of the latter (see Kilgarriff 1999), it stands to reason that they are essential to the monolingual organisation of the semantic space covered by a lemma, just as the division of this semantic space into areas covered by the various translations of the source item in the target language is at the very heart of bilingual lexicography. We need not look for justifications for such practice: they are all over the place.

We do need the two LEXDIS modes. The problem is that words do not carry labels indicating the word sense they illustrate, neither in the user's text (the task – Word Sense Discrimination – is precisely to provide such labels), nor in the lexicographical resources (where it would be reasonable to expect them, e.g. we can conceive of collocate lists where it would be clear that the *watch* that is offered as a collocate for the object of *wear* is not the lemma *watch* but a word sense or set of closely related word senses associated with that lemma). Consequently, when we attempt to assess the proximity of a textual element to a collocate, we are working with two lemmas, and the global mode is the one we have to use.

However, LEXDIS is also happy to work in local mode, trying to constrain the hypotheses as to what area of the semantic space is covered by a lexical item be found in a piece of text. Consider:

(3) *He was wearing an expensive red tie and a gold watch.*

from which a parser should enable us to derive the triplet: *t(wear, tie, watch),*i.e. *t(ArgBearer, FirstArg, SecondArg).*

Wear, **watch** and **tie** are all three polysemic, as are all reasonably frequent and reasonably heavy lexical items. By using LEXDIS on the triplet, we are able to reduce the number of word senses that ought to be taken into consideration to account for a 'standard' reading of (3). LEXDIS computes the triplets of best matches for the following relations:

ArgBearer – FirstArg
ArgBearer – Second Arg
FirstArg – SecondArg

and selects the word senses (if any) that occur in more than one relation. (QUERY: t(wear, tie, watch – see results in Appendix A, last query). It would seem that – in this particular case at least – the simple algorithm outlined above is powerful enough to enable LEXDIS to cash in on the word sense connectivity assumed by the connector linking the two arguments to the argument bearer.

[3] LEXDIS comes in three flavours : the one used in this paper does not make use of WordNet at all; a second flavour uses the WordNet glosses, but not the WordNet relations, besides the standard LEXDIS resources (for which see Appendix B); a third flavour uses the WordNet relations as well as the WordNet glosses and standard LEXDIS resources.

LEXDIS is written in SWI-Prolog (see Wielemaker 2003) and runs on standard PCs under the various operating systems for which there exists a SWI-Prolog interpreter/compiler (Windows, Linux, Mac-OS). Appendix A gives the protocol of a short LEXDIS session.

LEXDIS should be used in conjunction with a parser that is able to retrieve the syntactic relations that are targeted by the collocate lists of standard monolingual or bilingual dictionaries. Besides, such a parser should preferably be lexicon-centred, so that multi-word units can be recognized as such. Multi-word units have their own argument structure, in which some of the material can be lexically described, i.e. described at the level of individual lexical entries rather than broad grammatical categories. For example, in *go through the motions*, we identify the various components of the multi-word unit by providing descriptions that go as deep as the specification of individual lexical entries (*go*, *through*, *the*, *motions*), either as lemmas (*go*) or as textual forms (*motions*).

Thus the parser (namely VERBA – see Michiels 2009) will work with entries such as the following two for *go through* (two among a dozen):

A.

```
% they finally went through the marriage ceremony for the sake of their children
% they went through the scene over and over again
verb([v(goes,go,went,gone,going,go_through_3_perform_rehearse)],v_mwu_prep,
arglist:[subject:[type:np, canon:0,gappable:yes, oblig:yes,
constraints:[sem:[hum]]],

     athematic:[type:prep,canon:2,gappable:no,

                   oblig:yes,constraints:[lex:through]],

     arg_prep:[type:np,canon:3,gappable:yes,oblig:yes,

                     constraints:[c_str:[head:[lex:Lex]]]]],
ft:[pc:[coll(arg_prep,Lex,[marriage, initiation, scene, lesson, programme,

              ceremony, formality, procedure])]]).
```

B.

```
% go through the motions
verb([v(goes,go,went,gone,going,go_through_the_motions)],v_mwu_prep,
arglist:[subject:[type:np, canon:0,gappable:yes, oblig:yes,
constraints:[sem:[hum]]],

     athematic:[type:prep,canon:2,gappable:no,

                   oblig:yes,constraints:[lex:through]],

         arg_prep:[type:np,canon:3,gappable:no,oblig:yes,
constraints:[c_str:[det:[lex:the]],

       c_str:[head:[txt:motions]]]]],

    ft:[]).
```

In the first (A) we have one constraint specified down to the level of lexis, namely the lexeme value of the preposition (*through*), whereas in the second (B), we also reach the lexical level in specifying the constraints on the argument of the preposition: namely, that the determiner should be *the* and the textual form of the head noun phrase should be *motions*. In (A), the constraints on the argument of the preposition target a variable Lex which will be passed on to the *coll* procedure. This procedure will call on LEXDIS to measure the proximity of this lexeme with respect to each of the members of the collocate list associated with the argument of the preposition, namely marriage, initiation, scene, lesson, programme, ceremony, formality and procedure.

It should be stressed again that here as in innumerable similar cases, the only way to keep the various word senses apart is to exploit the collocate lists. And it is very uncommon for the lemma of the head of the textual argument to happen to be one of the collocates, however 'typical' the collocates are supposed to be[4]. We do need such a tool as LEXDIS to be able to make full use of the information on the environment of the argument bearer captured by the collocate lists.
Obviously, the parser must be able to keep track of the disruptions to the canonical order of the arguments brought about by various types of 'transformations'. In
I know the rituals inspectors are expected to go through.

whose parse is given in appendix F, VERBA must be able to retrieve the syntactic link between *through* and *rituals* despite the disruption brought about by relativization. Similarly, the subject raising due to *expect* should not prevent the parser from assigning '*inspectors*' as subject of '*go through*'. Otherwise, the constraints specified in the lexical entry could not be met (constraints requiring that the subject of *go through* should bear a +HUM semantic feature, and that the head of the prepositional object should be available for LEXDIS to compute its proximity with respect to the eight collocates).

Similarly, in our second entry, we need to recognize the lexical chain *the motions* in the appropriate syntactic slot, but we also need to be able to keep track of the subject, which can be embedded as far down as in the possessive adjective of a deverbal noun, as in:
She knew his refusal to go through the motions.
where the *his* of *his refusal* yields the subject of *go through* (third person personal pronoun, singular masculine).
Of course, in
I know the rituals inspectors are expected to go through.
the head of the textual argument of the preposition will be matched, not only against the eight elements of the collocate list associated with the 'perform_rehearse' reading of *go through*, but with all the collocate lists that belong to all the readings of *go through* that are posited by the dictionary we use (the Oxford Dictionary of Current Idiomatic English, with additional collocates taken over from the Oxford/Hachette and Robert/Collins bilingual dictionaries). The list numbers 44 items, given below in alphabetical order:
[apprenticeship, argument, beer, ceremony, clothes, cupboard, document, drink, edition, experience, experiment, fact, file, fire, food, formality, fortune, initiation, lesson, list, luggage, mail, marriage, money, operation, ordeal, pain , paper, phase, pocket, printing, procedure, process, programme, room, scene, stage, stock, store, subject, suitcase, text, trunk, wardrobe]

The task of running through all these collocates, matching them all against the textual candidate, is likely to be heavy on computer resources, as indeed it is[5]. This is even more the case if LEXDIS is embedded in a parser such as VERBA, whose design features are dictated entirely by perspicuity and ease of use by a linguist and/or lexicographer. VERBA is indeed purely incremental, building

[4] In the case of our *go through* entries, such identity between textual filler of the argument and corresponding collocate is most likely in the very restricted reading of *go through* as **be published,** where the collocates *printing* and *edition* will often be found as textual exponents of the argument of the preposition.

[5] LEXDIS takes only about 4 seconds to answer all the 13 queries of the **quicktest** file given in Appendix A, but the full parsing of *I know the rituals inspectors are expected to go through* (Appendix F), producing five parses corresponding to five 'readings' of *go through,* takes about 20 seconds cputime (cputime : 21.6217). The parsing of the parallel sentence *I know the books inspectors are expected to write,* yielding a single parse and making no call on LEXDIS, takes 2.5 seconds cputime.

structure on top of structures established in a previous pass, and making use of feature unification all through.

However, we have got accustomed to seeing the availability of computer resources increase dramatically over time, so that an emphasis on clear design principles rather than efficiency is a reasonable choice to make, in a world where you can't have your cake and eat it, to end this paper on a multi-word unit (... avoiding a near miss).

References

Lexicographical resources

CIDE = *Cambridge International Dictionary of English*. Cambridge Press, Cambridge, England, 1995
COBUILD = *The Collins COBUILD English Language Dictionary,* edited by J. Sinclair *et al.,* HarperCollins, London and Glasgow, 1987
LDOCE = *Longman Dictionary of Contemporary English* edited by Paul Procter, Longman, Harlow, 1978
ODCIE = *Oxford Dictionary of Current Idiomatic English* (Vol.1: Verbs With Prepositions and Particles), edited by A.P. Cowie and R. Mackin, Oxford University Press, London, 1975
OH = Oxford/Hachette English/French pair (*The Oxford-Hachette French Dictionary French-English English-French* edited by Marie-Hélène Corréard and Valerie Grundy, Oxford University Press, Oxford, Hachette, Paris, 1994)
RC = Le Robert and Collins English/French pair (*Collins Robert French/English, English/French Dictionary*, Unabridged, Third Edition, edited by Beryl T. Atkins, Alain Duval and Rosemary C. Milne, Harper-Collins Publishers, 1993 (First ed. 1978)
WordNet = WordNet 3.0 Prolog files (see Miller 1990)

Other references

Banerjee, S. and Pedersen, T. (2003). 'Extended gloss overlaps as a measure of semantic relatedness'. In Proceedings of the Eighteenth International Conference on Artificial Intelligence (IJCAI-03), Acapulco,Mexico.
Budanitsky, A., and Hirst, G., 'Evaluating WordNet-based Measures of Lexical Semantic Relatedness', *Computational Linguistics***,** March 2006, Vol. 32, No. 1, 13-47
Church, K., and Hanks, P. (1989). 'Word Association Norms, Mutual Information and Lexicography,' *Association for Computational Linguistics,* Vancouver, Canada
Ide, N. and Véronis, J. (1998). 'Word Sense Disambiguation: The State of the Art', *Computational Linguistics*, 1998, 24(1)
Kilgarriff, A. (1999), ' “I don't believe in word senses” ', ITRI-97-12, *Information Technology Research Institute Technical Report Series*, ITRI, University of Brighton, Brighton, March, 1999
Lesk, M. (1986). 'Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from anIce Cream Cone.' *Proceedings of the 1986 SIGDOC Conference,* Toronto, Canada, June 1986, 24-26.
Li, Y., Bandar, Z.A. and Mclean, D. (2003), 'An approach for measuring semantic similarity between words using multiple information sources', *IEEE Transactions onKnowledge and Data Engineering*, Volume 15, Issue 4, 871- 882
Michiels, A. (1999). 'An Experiment in Translation Selection and Word Sense Discrimination', in Tops, Guy, Devriendt, Betty and Geukens, Steven (eds), *Thinking English Grammar,* Orbis/Supplementa, Tome 12, Peeters, Leuven-Paris, 1999, pp. 383-407

Michiels, A. (2000). 'New Developments in the DEFI Matcher', *International Journal of Lexicography,* Vol. 13, No 3, 2000, 151-167
Michiels, A. (2001).' DEFI, un outil d'aide à la compréhension**',** in *Actes du Congrès TALN 2001*, Tours, 2001, 283-293
Michiels, A. (2002). 'Le traitement de la phraséologie dans DEFI', Van Vaerenbergh, Leona (ed), *Linguistics and Translation Studies. Translation Studies and Linguistics*. Linguistica Antverpiensia. New Series 1/2002, 2002, 349-364
Michiels, A. (2006). 'Les lexies en TAL', in Bracops,M., Dalcq, A.-E., Goffin, I., Jabé, A., Louis, V. and Van Campenhoudt, M., eds., 2006 : *Des arbres et des mots. Hommage à Daniel Blampain*, Bruxelles, Éditions du Hazard, ISBN : 2-930154-14-4. (http://hdl.handle.net/2268/1884)
Michiels, A (2009). ', a Multi-word-unit-oriented Feature-unification-based Parser', Unpublished paper, University of Liège, 2009 (available as

http://promethee.philo.ulg.ac.be/engdep1/download/prolog/lexdis/verba.pdf)
Miller, G. (1990). 'Wordnet: An on-line lexical database.' *International Journal of Lexicography (special issue)*, 3(4):235–312.
Montemagni, S., Federici, S. and Pirrelli, V. (1996). 'Example-based Word Sense Disambiguation: a Paradigm-driven Approach', in *Euralex'96 Proceedings*, Göteborg University, 151-160.
Pustejovsky, J. (1995). *The Generative Lexicon* The MIT Press 1995.
H. Rubenstein, H. and Goodenough, J.B. (1965). 'Contextual correlates of synonymy.' *Communications of the ACM*, 8:627–633.
Wielemaker, J. (2003). 'An overview of the SWI-Prolog Programming Environment', in Fred Mesnard, F. and Serebenik, A., eds, *Proceedings of the 13th International Workshop on Logic Programming Environments,* Katholieke Universiteit Leuven, Heverlee, Belgium, 2003

APPENDICES

A. *Sample LEXDIS Session*

```
QUERY: [gun, pos:n, hunter]
gun with POS=n is related to hunter with POS=n with weight=4.53074 as follows:
Shared Labels: [hfzh] -> weight: 4
Shared words in definition: [metal, sport, hunting] -> weight: 9
Shared words in examples: [big] -> weight: 1
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/3.09

QUERY: [gun, pos:n, hunger]
gun with POS=n is ***not*** related to hunger with POS=n (Adjusted Weight = 0.0).

QUERY: [bread, pos:n, hunger]
bread with POS=n is related to hunger with POS=n with weight=1.58416 as follows:
Shared words in definition: [food, formal] -> weight: 4
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/2.525

QUERY: [bread, pos:n, gun]
bread with POS=n is ***not*** related to gun with POS=n (Adjusted Weight = 0.0).

QUERY: [cat, pos:n, dog]
cat with POS=n is related to dog with POS=n with weight=9.38104 as follows:
Shared Labels: [am] -> weight: 4
Shared words in definition: [animal, man, woman, four-legged, pet, animals, life] -> weight:
21
Shared words in examples: [dogs, pet] -> weight: 2
Cooccurrence in collocate lists  -> weight: 10
Cooccurrence in Roget's thesaurus  -> weight: 3
Cooccurrence in R/C-Oxf/Hach indic db  -> weight: 4
Cooccurrence in R/C-Oxf/Hach extended lemma db  -> weight: 4.5
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/5.17

QUERY: [cat, pos:n, fox]
cat with POS=n is related to fox with POS=n with weight=7.98065 as follows:
Shared Labels: [am] -> weight: 4
Shared words in definition: [animal, belonging, family, woman, fur, small, furry, tail] ->
weight: 24
Cooccurrence in Roget's thesaurus  -> weight: 1
Cooccurrence in R/C-Oxf/Hach indic db  -> weight: 4
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/4.135

QUERY: [serpent, pos:n, snake]
serpent with POS=n is related to snake with POS=n with weight=16.7488 as follows:
Shared Labels: [am] -> weight: 4
Shared words in definition: [large] -> weight: 22
Cooccurrence in Roget's thesaurus  -> weight: 8
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/2.03
```

```
QUERY: [serpent, pos:n, bird]
serpent with POS=n is related to bird with POS=n with weight=1.46789 as follows:
Shared words in definition: [creature] -> weight: 2
Cooccurrence in Roget's thesaurus  -> weight: 2
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/2.725

QUERY: [pool, pos:n, water]
pool with POS=n is related to water with POS=n with weight=12.5761 as follows:
Shared words in definition: [large, ground, swimming, liquid, amount, surface, area, supply,
goods, thin] -> weight: 70
Shared words in examples: [swimming, pounds, men] -> weight: 13
Cooccurrence in R/C-Oxf/Hach indic db  -> weight: 5
Cooccurrence in R/C-Oxf/Hach collocates db  -> weight: 5
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/7.395

QUERY: [pool, pos:n, house]
pool with POS=n is related to house with POS=n with weight=2.1102 as follows:
Shared words in definition: [large, small, money, goods, business] -> weight: 15
Shared words in examples: [football, won, cheap] -> weight: 3
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/8.53

QUERY: [pool, pos:n, football]
pool with POS=n is related to football with POS=n with weight=12.3214 as follows:
Shared Guide Words: [game] -> weight: 12
Shared words in definition: [large, filled, game, try, win, players, played, games] -> weight:
44
Shared words in examples: [garden, local, want, play] -> weight: 9
Cooccurrence in Roget's thesaurus  -> weight: 1
Cooccurrence in R/C-Oxf/Hach extended lemma db  -> weight: 3
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/5.6


QUERY: [pool, pos:n, tennis]
pool with POS=n is related to tennis with POS=n with weight=5.18135 as follows:
Shared words in definition: [small, game, area, players, played, hit] -> weight: 18
Shared words in examples: [play] -> weight: 6
Cooccurrence in Roget's thesaurus  -> weight: 1
Penalty for Heavy Lexical Items  -> adjusted_weight= weight/4.825

QUERY: [pool, pos:n, football, pos:n, w:t, m:l]
pool  n  [football_pools]  []  co44102
Def: if you do the pools you take part in a type of gambling competition in which people try to win money by
guessing correctly the results of football matches used in british english
is related to
football  n  [soccer]  [fb]  ci27328ci69956*co22637#co22638/lg29686&football%1:04:00::
Def: [ a game in which two teams of  players try to kick or use their heads to send a round ball into the goal
of the opposing side, any of various games played with a ball (round or oval) in which two teams try to kick or
carry or propel the ball into each other's goal, any of various similar games played with a round or oval ball
which is kicked thrown or carried in an attempt to score goals or reach the opposing teams goal line rugby
american football and australian rules football are all types of football,  any of several games for  teams in
which a ball is kicked andor thrown about a field in an attempt to get goals esp bre soccer any of several
games for  teams in which a ball is kicked andor thrown about a field in an attempt to get goals esp bre rugby
any of several games for  teams in which a ball is kicked andor thrown about a field in an attempt to get
goals esp ame american football any of several games for  teams in which a ball is kicked andor thrown
about a field in an attempt to get goals esp austre australian rules football, football is a game played between
two teams of eleven players who kick a ball around a field in an attempt to score goals,  football br informal
footie or footy esp am soccer is a game played with a large round ball between two teams of esp eleven
people where each team tries to win by kicking the ball into the other teams goal]
with weight=24 as follows:
Shared words in definition: [try, win] -> weight: 24

QUERY: [pool, pos:n, water, pos:n, w:t, m:l]
pool  n  []  []  lg43110
Def:  a small area of still water in a hollow place usu naturally formed
is related to
water  n  []  []  ci83118
Def:  water often refers to an area of water such as the sea a lake or a swimming pool
with weight=52 as follows:
Shared words in definition: [area] -> weight: 42
Shared words in examples: [] -> weight: 10

QUERY: t(wear, tie, watch)
[-3-p(wear, ci83351, tie, ci77397/lg53208#co60228&tie%1:06:01::), -2-p(wear, ci83339, tie,
co60229), -2-p(wear, co64915, tie, co60233)]
[-22-p(wear, ci83339, watch, ci83082#co64601), -20-p(wear, ci83340, watch,
ci83082#co64601), -20-p(wear, ci83347, watch, ci83082#co64601)]
[-2-p(tie, ci77397/lg53208#co60228&tie%1:06:01::, watch, ci83082#co64601), -2-p(tie,
ci77397/lg53208#co60228&tie%1:06:01::, watch, lg55831), -2-p(tie,
ci77397/lg53208#co60228&tie%1:06:01::, watch, lg55833)]

wear  v  [body]  []  ci83339
Def:  to have clothing or jewellery on your body
tie  n  [necktie]  [cl]  ci77397/lg53208#co60228&tie%1:06:01::
Def: [neckwear consisting of a long narrow piece of material worn (mostly by men) under a collar and tied in
knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie", a tie is a long
narrow piece of cloth that is worn round the neck under a shirt collar and tied in a knot at the front ties are
worn mainly by men see also bow tie old school tie,  also esp ame necktie a band of cloth worn round the
neck usu inside a shirt collar and tied in a knot at the front,  a tie also esp am necktie is a long thin piece of
material that is worn under a shirt collar esp by men and tied in a knot at the front]
watch  n  [clock]  []  ci83082#co64601
Def: [ a small clock which is worn on a strap around the wrist or sometimes connected to a piece of clothing
by a chain, a watch is a small clock which you wear on a strap on your wrist or on a chain]

QUERY: nadamas

cputime : 4.02483
```

B. What is Lexdis based on ?

LEXDIS calls on the following lexical resources, available to us mainly through research contracts:

semdic : dictionary *clauses* (i.e. Prolog clauses) derived from CIDE, COBUILD, LDOCE and the WordNet Synsets and Synset Glosses (the dictionary clauses feature the lexical items in both definitions and examples as *word bags*, to the exclusion of a specially designed list of stopwords, both tool words and words specific to lexicographic practice, such as *especially*)
Here is one of the dictionary clauses for the entry CAT (derived from CIDE):

```
mono(lem('cat'), ori('ci'), idnum('ci10278'), pos('n'), lab([]), gw([]),
deflex(['small', 'four-legged', 'furry', 'animal', 'tail', 'claws', 'pet', 'catching', 'mice', 'member',
'biologically', 'similar', 'animals', 'lion']),
exlex(['pet', 'stray', 'feed', 'holiday']),
def(' a small four-legged furry animal with a tail and claws usually kept as a pet or for
catching mice or any member of the group of biologically similar animals such as the
lion')).
```

mt : data base of RC/OH collocates - the pivotal property is co-presence within the same collocate field (cf. the hypothesis put forward in Montemagni et al. 1996). The co-occurrence lists are assigned as early as possible in the alphabetical ranking of the lexical items and it is therefore the 'smaller' word that should be explored. An mt line looks like the following:

```
mt(digestion, [ [growth,1], [machine,1], [mind,1], [movement,1], [reaction,1], [recovery,1],
[stomach,4]]).
```

this means that the word 'digestion' co-occurs 1 time with 'growth' in a collocate list ... and 4 times with 'stomach'; the sharing of 'digestion' with a word preceding 'digestion' should be looked for under that word

roget : database of Roget's Thesaurus Categories (three levels)
Connectedness is established through the sharing of Roget's categories; three levels of delicacy in thesaurus organisation are catered for. A *r* line looks like the following:

```
r( 'antiquarian', [ ['n','122','4','4'],  ['n','492','4','2'] ])
```

which means that the word *antiquarian* is a noun that belongs to the two category triples *122/ 4/ 4* and *492 /4 /2* where the broadest category comes first (492), followed by sub-category(4) and sub-sub-category(2).

indic: data base of RC(Robert/Collins)/OH(Oxford/Hachette) indicators (in these two bilingual E-F/ F-E dictionaries, only the E->F direction is explored). Here are the indicators for the lexeme CAT:

```
ind(lemma('cat'),pos(n),indic(['catalytic', 'converter'])).
ind(lemma('cat'),pos(n),indic(['domestic'])).
ind(lemma('cat'),pos(n),indic(['feline', 'species'])).
ind(lemma('cat'),pos(n),indic(['female'])).
ind(lemma('cat'),pos(n),indic(['guy'])).
ind(lemma('cat'),pos(n),indic(['man', 'woman'])).
ind(lemma('cat'),pos(n),indic(['man'])).
ind(lemma('cat'),pos(n),indic(['woman'])).
```

coll : data base of RC/OH collocates - the pivotal element is the collocate bearer
Connectedness is established through collocate sharing in RC/OH collocate data base. A sample line:

```
coll( lemma('abandonment'), pos(n), coll(['property', 'right'])).
```

Here, the two items are related if they POSSESS common elements in their collocate lists, whereas in metameet it is the co-presence within a collocate list (associated with whatever item) that is significant.

envir : data base of environments derived from RC/OH 'extended' lemmas i.e. including phrases and examples

```
e( hdwd('dative'), envir(['case','ending'])).
e(hdwd('cat'), envir(['big', 'cats', 'burglar', 'cat-basket', 'cat-lick', 'cat-onine-tails', 'catbird',
'seat', 'catfood', 'catgut', 'cathouse', 'catmint', 'bag', 'dogs', 'mice', 'play', 'cats-cradle', 'cats-
eye', 'cats-paw', 'cats-whisker', 'catsuit', 'door', 'family', 'fight', 'dog', 'flap', 'give', 'grin',
'hardly', 'room', 'swing', 'hot', 'bricks', 'tin', 'roof', 'jump', 'kill', 'laugh', 'lead', 'life', 'let', 'look',
'brought', 'dragged', 'king', 'pigeons', 'cat-and-mouse', 'game', 'mouse', 'rain', 'see', 'jumps',
'skin', 'take', 'catnap', 'think', 'meow', 'pajamas', 'whiskers', 'thinks', 'wait'])).
```

pesi : data base recording the lexical weight of lemmas
We keep track of the lexicographical space occupied by lexical items for weighting purposes. We can thus decrease the factor of computed proximity in the case of ‘heavy’ lexical items. Lexical weight is computed by LEXDIS itself; it is the weight that LEXDIS assigns to the link between a *Word,Pos* pair and itself. By leaving Word and Pos as variables and executing the query *[Word,Pos,Word,Pos,w:none,m:g]*, we obtain the weights of all the lexical items making up **semdic**.

```
w(cat, n, 512).
w(dog, n, 522).
```

For the purposes of this paper we have used Lexdis in its WordNet-free variety so as to enable comparison with WordNet-based and text corpus-based methods. But LEXDIS also features a variety where the WordNet glosses are part of semdic, and an additional one where the following WordNet or WordNet-derived data bases are used:

s : Synset WordNet database in Prolog format downloadable from the WordNet website. The WordNet 's' predicate yields the Synset to which a Word-Pos pair belongs, as in

```
s(105441468,1,'suppressor',n,2,0).
```

paths : database of paths derived from various WordNet data bases. The 'path' predicate' is built on the basis of a recursive exploration of the following Synset-to-Synset relations: **cs, ent, hyp, ins, mm, mp,ms**. A path line looks like this

```
path(105441468,
[105436752,108459252,108457976,108456993,107938773,100031264,100002137,10000
1740]).
```

i.e. a synset identifier followed by a hierarchical list of hypernym synsets of various types (hypo-hyper stricto sensu, part-whole, etc.)

C. The LEXDIS algorithm

(Global mode, WordNet-free dictionary)

Let Weight be the weight computed by LEXDIS to reflect the proximity of two lexical items Wx and Wy.
Let Ci(S1,S2) be the cardinality of the intersection of two sets S1 and S2, i.e. the number of elements that they have in common.

Computation

Weight = GlobalWeight/LW
GlobalWeight = DicW + IndicW + EnvirW + CollW + MetaW + RogetW

1) From SemDic, a merge of CIDE, LDOCE and COBUILD

DicW = LabW + GuideW + DefW + ExW

a) labels

LabW= Ci(Lab-x,Lab-y) * 4
Lab-x, Lab-y : set of Labels associated with Wx, Wy

b) guidewords

GuideW= Ci(Guide-x,Guide-y) * 12
Guide-x, Guide-y : set of Guide-Words associated with Wx, Wy

c) definitions

Let Ldef= Ci(Def-x,Def-y).
Def-x, Def-y : set of Def-Words associated with Wx,Wy (the words occurring in the definitions, except those belonging to a stoplist of high-frequency words, including lexicographical toolwords such as *especially*)
If Ldef > 0 then If Ldef in [1,2] then Mdef = 2 else Mdef = 3
Let BonusDefx be 20 if Word-x occurs in Def-y, 0 otherwise.
Let BonusDefy be 20 if Word-y occurs in Def-x, 0 otherwise.
DefW = (Ldef * Mdef) + BonusDefx + BonusDefy

d) examples

Let Lex= Ci(Ex-x,Ex-y).
Ex-x, Ex-y : set of Ex-Words associated with Wx,Wy (the words occurring in the examples, except those belonging to a stoplist of high-frequency words)
Let BonusExx be 5 if Word-x occurs in Ex-y, 0 otherwise.
Let BonusExy be 5 if Word-y occurs in Ex-x, 0 otherwise.
ExW = Lex + BonusExx + BonusExy

2) From Indic, the database of indicators used in defidic (E-F pairs of RC-OH)

Let Indic= Ci(Indic-x,Indic-y)
Indic-x, Indic-y : set of indicators associated with Wx,Wy
Let BonusIndicx be 5 if Word-x occurs in Indic-y, 0 otherwise.
Let BonusIndicy be 5 if Word-y occurs in Indic-x, 0 otherwise.

```
If Indic > 0 then
        If Indic in [1,2] then
                    IndicW = (Indic * 2) + BonusIndicx + BonusIndicy
                    else
                    IndicW = (Indic * 4) + BonusIndicx + BonusIndixy
```

3) From Envir, the database of environments extracted from an extended lemma db

Let Envir= Ci(Envir-x,Envir-y)
Envir-x, Envir-y : set of environments associated with Wx,Wy
Let BonusEnvirx be 2 if Word-x occurs in Envir-y, 0 otherwise.
Let BonusEnviry be 2 if Word-y occurs in Envir-x, 0 otherwise.

```
If Envir > 1 then
        EnvirW = Envir/2  + BonusEnvirx + BonusEnviry
        else
        EnvirW = BonusEnvirx + BonusEnviry
```

4) From Coll, the database of collocates used in defidic (E-F pairs of RC-OH)

Let Coll= Ci(Coll-x,Coll-y)
Coll-x, Coll-y : set of collocates associated with Wx,Wy
Let BonusCollx be 5 if Word-x occurs in Coll-y, 0 otherwise.
Let BonusColly be 5 if Word-y occurs in Coll-x, 0 otherwise.

```
If Coll > 0 then
        If Coll in [1,2] then
                    CollW = (Coll * 2)  + BonusCollx + BonusColly
                    else
                    CollW = (Coll * 4) + BonusCollx + BonusColly
```

5) From Mt, a database of collocate sharings derived from Coll

Let M be the number of times Word-x is used by the side of Word-y in collocate lists derived from Defidic, a merge of the RC and OH English->French dictionaries.

```
If either Word-x or Word-y is in [person,object]  then
               MetaW is M//16
               else
               MetaW is M//2
```

(// is used for integer division)

6) From Roget, a db of Roget's categories derived from Roget's thesaurus

Let Rwx and Rwy be the lists of Roget's triples associated with Word-x and Word-y respectively.

Each member i of Rwx is matched with all members of Rwy, yielding Rw_i

RogetW is the sum of all Rw_is

Rw_i is itsef a sum of matches between Roget's triples, which are roget-matched to yield a RMW, a weight reflecting the quality of the match.

Roget-matching yields a RMW of 3 if all three tiers (class, category, sub-category) are matched in the two triples considered, a RMW of 2 if only class and category match, and a RMW of 1 if the match holds only between the two classes.

7) From pesi, a db of lexicographical weights

Let Lwx and Lwy be the lexicographical weights associated with Word-x and Word-y in the pesi database.

Lwxy = Lwx+Lwy

If Lwxy >= 200 then LW=Lwxy/200 else LW=1

D. LEXDIS at grips with the Rubenstein and Goodenough word pairs

WORD PAIRS	HUMAN SIMILARITY ASSESSMENT	LEXDIS Global Mode
[cord,n,smile]	0,02	0
[rooster, n, voyage]	0,04	0
[noon, n, string]	0,04	0,42
[fruit, n, furnace	0,05	0,32
[autograph, n, shore]	0,06	0
[automobile, n, wizard]	0,11	0
[mound, n, stove]	0,14	0,86
[grin, n, implement]	0,18	0
[asylum, n, fruit]	0,19	0,63
[asylum, n, monk]	0,39	0
[graveyard, n, madhouse]	0,42	0
[glass, n, magician]	0,44	0
[boy, n, rooster]	0,44	0,8
[cushion, n, jewel]	0,45	0
[monk, n, slave]	0,57	0
[asylum, n, cemetery]	0,79	0
[coast, n, forest]	0,85	4,47
[grin, n, lad]	0,88	0
[shore, n, woodland]	0,9	2,42
[monk, n, oracle]	0,91	1,12
[boy, n, sage]	0,96	0,95
[automobile, n, cushion]	0,97	0
[mound, n, shore]	0,97	1,6
[lad, n, wizard]	0,99	0,88

[forest, n, graveyard]	1	2,06
[food, n, rooster]	1,09	0
[cemetery, n, woodland]	1,18	2
[shore, n, voyage]	1,22	0,63
[bird, n, woodland]	1,24	0,39
[coast, n, hill]	1,26	17,07
[furnace, n, implement]	1,37	1,29
[crane, n, rooster]	1,41	9,12
[hill, n, woodland]	1,48	5,68
[car, n, journey]	1,55	2,99
[cemetery, n, mound]	1,69	1,67
[glass, n, jewel]	1,78	10,22
[magician, n, oracle]	1,82	0
[crane, n, implement]	2,37	0
[brother, n, lad]	2,41	5,27
[sage, n, wizard]	2,46	1,19
[oracle, n, sage]	2,61	2,33
[bird, n, crane]	2,63	7,3
[bird, n, cock]	2,63	10,04
[food, n, fruit]	2,69	7,61
[brother, n, monk]	2,74	16,6
[asylum, n, madhouse]	3,04	5,52
[furnace, n, stove]	3,11	14,25
[magician, n, wizard]	3,21	14,69
[hill, n, mound]	3,29	12,37
[cord, n, string]	3,41	13,39
[glass, n, tumbler]	3,45	12,79

[grin, n, smile]	3,46	26,74
[serf, n, slave]	3,46	23,87
[journey, n, voyage]	3,58	22,55
[autograph, n, signature]	3,59	25,15
[coast, n, shore]	3,6	19,39
[forest, n, woodland]	3,65	15,48
[implement, n, tool]	3,66	15,29
[cock, n, rooster]	3,68	31,07
[boy, n, lad]	3,82	15,06
[cushion, n, pillow]	3,84	23,42
[cemetery, n, graveyard]	3,88	26,02
[automobile, n, car]	3,92	50,31
[midday, n, noon]	3,94	58,22
[gem, n, jewel]	3,94	27,59

Correlation factor of 0,869 as computed by Pearson function of Excel-type spreadsheet

E. Exploring both the syntagmatic and the paradigmatic axes.

Carriage and attitude, grenadier and parade in Henry James' THE AMERICAN

His usual **** attitude and carriage **** were of a rather relaxed and lounging kind, but when, under a special inspiration, he straightened himself, he looked like a **** grenadier on parade *****. **Henry James, THE AMERICAN, Penguin Books edition, 1986 [1876-1877], p.34-35**

The *attitude/carriage* word pair leads to easy identification of the relevant word senses for the two items in any measure of lexical distance that is able to explore the paradigmatic axis, such as measures based on the WordNet relations. LEXDIS has no trouble with them either, as it explores both axes. In the *grenadier/parade* pair, however, paradigmatic exploration is of no avail; we do need an exploration of the other axis, the syntagmatic one: the sharing of a label reflects the redundancy along the syntagmatic axis ('maximisation des lectures isotopiques').

QU**ERY:** [grenadier, pos:n, parade, pos:n, w:t, m:l]

```
grenadier  n  []  [mi]  lg31458
Def:  a soldier formerly one who threw grenades
is related to
parade  n  []  [mi]  lg40875/lg40879
Def: [ esp of soldiers an example of a gathering together in ceremonial order for the purpose of being officially looked
at or for a march esp in the phr on parade,  also parade ground -- - a large flat area where soldiers parade ]
with weight=4 as follows:
Shared Labels: [mi] -> weight: 4
```

Charge, grenadier and bayonet in JM Coetzee's 'YOUTH'

as the French **** grenadiers **** came *** charging *** at him with their grim **** bayonets *****. **JM Coetzee, YOUTH, Vintage edition, 2003 [2002], p.85**

QUERY: t(charge,POS, grenadier, n, bayonet, n)

[Here we see the use of an uninstantiated variable for the POS slot (Prolog variables have a capital letter as first character), so as to explore the lexical information associated with *charge* both as a noun and as a verb. The algorithm used on triplets leads to the selection of the word senses that participate in more than one relation.]

[-4-p(charge, **lg21167**, grenadier, **lg31458**), -4-p(charge, lg21167, grenadier, lg31459), 0-p(charge, ci10969, grenadier, ci31434)]
[-4-p(charge, **lg21167**, bayonet, **bayonet%1:06:00::/lg17834**), -2-p(charge, ci10997, bayonet, bayonet %1:06:00::/lg17834), -2-p(charge, ci10998, bayonet, bayonet%1:06:00::/lg17834)]
[-4-p(grenadier, **lg31458**, bayonet, **bayonet%1:06:00::/lg17834**), -4-p(grenadier, lg31459, bayonet, bayonet %1:06:00::/lg17834), 0-p(grenadier, ci31434, bayonet, bayonet%1:06:00::/lg17834)]

```
charge  n  []  [mi]  lg21167
Def:  a rushing forceful attack
grenadier  n  []  [mi]  lg31458
Def:  a soldier formerly one who threw grenades
bayonet  n  []  [mi, kl]  bayonet%1:06:00::/lg17834
Def: [a knife that can be fixed to the end of a rifle and used as a weapon,  a long knife fixed to the end of a soldiers gun
rifle]
```

(the reference to *bayonet%1:06:00::* in a WordNet-free LEXDIS is readily explained: the WordNet entry participated in the merge process designed to streamline the dictionary used by LEXDIS, and got linked to the LDOCE entry for *bayonet*.)

F. A Verba parse with WSD

1. String :

I know **the rituals** inspectors are expected to **go through**.

2. WordList:

[0/i, 1/know, 2/the, 3/rituals, 4/inspectors, 5/are, 6/expected, 7/to, 8/go, 9/through, endpos(10)]

3. Pretty-printed parse

```
cat:pred
   voice:active
   weight_coll:0
   c_str
      head
         cat:vg
         pos:v
         lex:know
         tense:present
         voice:active
      subject
         cat:np
         sem:[hum]
         lex:i
         index:i(0, 1)
         c_str
            head
               lex:i
               sem:[hum]
      object
         cat:np
         weight_coll:12.3153-ceremony-4.39294
         index:i(2, 10)
         sem:[abstract]
         lex:ritual
         c_str
            head
               cat:np
               sem:[abstract]
               lex:ritual
               index:i(2, 4)
               c_str
                  det
                     pos:det
                     lex:the
                  head
                     pos:n
                     lex:ritual
                     sem:[abstract]
```

```
rel_clause
   index:i(2, 4)
   sem:[abstract]
   weight_coll:0
   c_str
      head
         auxgroup:[tense:present]
         prop:[voice:passive]
         pos:v
         lex:expect
         tense:untensed
         voice:passive
      subject
         cat:np
         sem:[hum]
         lex:inspector
         index:i(4, 5)
         c_str
            det
               det
               zero
            head
               pos:n
               lex:inspector
               sem:[hum]
      object
         cat:pred
         voice:active
         weight_coll:0
         c_str
            head
               auxgroup:[tense:untensed]
               pos:v
               lex:go_through_3_perform_rehearse
               tense:untensed
               voice:active
            subject
               e:i(4, 5)
            arg_prep
               e:i(2, 4)
```

G. WSD on 'go through' as a prepositional verb (Lexdis in global mode)

We give the collocate that yields the highest score, and also the mean weight of the matches with the other elements of the same collocate list : 12.31-ceremony- 4.39 means that the best match of the textual element 'ritual' (from the test sentence 'She was expected to go through a daily ritual') in all the *go-through* collocate lists is the collocate *ceremony* with weight 12.31, to be found in the collocate list **associated with the perform-rehearse WS of go through; 4.39** is the mean match yielded by *ritual* as against all the members of the relevant collocate list : [*marriage, initiation, scene, lesson, programme, ceremony, formality, procedure*].

Perhaps a remark is in order regarding the relationship between the highest value for proximity and the average one. If the two values diverge widely (let's say something like a 10:1 ratio), we might enquire whether we do not have a problem, either with the collocate list or with LEXDIS itself. The various elements of a collocate list should describe a lexical universe displaying a certain degree of coherence; this hypothesis is at the very basis of using co-presence in a collocate list in order to relate two lexical items (cf. Montemagni et al. 1996), a property that LEXDIS itself uses (by exploiting the MT database). In fact, the relationship between the top value and the average, once the coherence of LEXDIS itself is better established, could be used as a marker for a need to revise the collocate lists, and perhaps the division of the lexical space covered by the item the collocate lists are attached to.

Collocate lists:

Word sense 'consume' :	[money, food, drink]
Word sense 'search' :	[room, pocket, clothes, cupboard, wardrobe, luggage, suitcase, trunk]
Word sense 'perform_rehearse':	[marriage,initiation,scene, lesson,programme, ceremony, formality, procedure]
Word sense 'examine':	[fact, argument, subject, file, mail, text, list, document]
Word sense 'endure_experience':	[operation, pain, ordeal, apprenticeship, fire, phase, stage, process, experience, experiment]
Word sense 'be published':	subject : [book, title, article] prepositional object : [printing, edition]

In all the sentences of our little corpus of simplified sentences (they are derived from genuine BNC sentences, but edited so as to be parsable by VERBA), the VERBA-LEXDIS tandem is able to pick out the WS that would be preferred by a human reader (in ***bold italics*** in the following table).

GO THROUGH sentences submitted to Verba	consume	search	perform_rehearse	examine	endure_experience	be_published
He thought that she had gone through the **bins**.	1.1-drink-0.49	***13.01-pocket-4.01***	1.21-ceremony-0.48	4.3-file-1.63	1.15-fire-0.35	_
The inspector wanted the teachers to go through his **report**.	2.58-money-1.45	1.83-pocket-0.37	4.17-lesson-1.7	***8.51-document-3.44***	2.61-process-1.15	_
She was expected to go through a daily **ritual**.	0.87-drink-0.72	2.91-pocket-0.65	***12.31-ceremony-4.39***	2.10-text-0.63	3.21-process-1.57	_
We know the **suffering** you went through.	1.12-drink-0.92	2-clothes-0.57	1.28-lesson-0.55	2.85-subject-0.82	***16.62-pain-3.49***	_
It is the **training** the students will be expected to go through.	4.81-money-3.03	6.82-pocket-1.85	5.02-lesson-2.07	2.35-subject-0.59	***7.25-experience-2.67***	_
You don't know the **torture** I have gone through.	1.27-food-0.84	1.47-room-0.61	1.11-scene-0.58	2.25-subject-1.05	***9.95-pain-1.86***	_
I think that they went through five **bottles**.	***13.42-drink-5.14***	5.81-trunk-2.68	0.81-ceremony-0.37	3.66-file-0.96	1.99-pain-0.67	0.30-printing-0.15
The report went through five **editions**.	_	_	_	_	_	subj***:6.95-article-4.07*** *pobj**:1000-edition-503.52***