Computers and the Humanities **34:** 49–60, 2000.

Framework and Results for French

FRÉDÉRIQUE SEGOND*
Xerox Research Centre Europe, Meylan, France

1. Setting Up the French Exercise

To make the evaluation exercise valuable and useful it is important to prepare the evaluation material according to a rigorous methodology. This includes having clear criteria for choosing the words, being aware of the consequences on the evaluation of the dictionary and corpus choice. Also, because sense disambiguation is a difficult task even for human beings, it is important to provide comparison figures with human tagger agreement. In the following sections we present the adopted methodology together with the material used.

1.1. CHOOSING THE CORPUS

The corpus used for the ROMANSEVAL[1] exercise is the same as the ones used within the ARCADE project.[2] It is a parallel corpus comprising nine European languages[3] (ca. 1.1 million words per language). This corpus has been developed within the MLCC-MULTEXT projects.[4] It is composed of written questions asked by members of the European parliament on a wide variety of topics and of corresponding answers from the European Commission. The format is just plain text. Sentences are relatively long and the style is, unsurprisingly, rather administrative.

Although we did not use yet the parallel aspect of this corpus we plan to use it in order to study, for instance, relationships between sense tagging and translation.

1.2. CHOOSING THE WORDS

The choice of test words is particularly difficult and cannot be left to intuition. Frequency criteria have a lot of drawbacks. If it is true that frequent content words of a text are very polysemous, it also has been shown that a large number of the words tend to be mostly monosemous in a given corpus. As such, a list of frequent words does not permit a proper evaluation of automatic WSD systems. Choosing the most polysemous words of a dictionary has also some drawbacks: chances are high that few of these senses appear in a corpus.

Table I. Average polysemy across four dictionaries.

	verbs	adjectives	nouns
French dictionary	12.6	6.3	7.6
Italian dictionary	5.3	4.7	4.9
English dictionary[7]	5.1	4.4	5.0
WordNet	8.63	7.95	4.74

We used a combination of these two methods. We extracted 60 words (i.e. 20 nouns, 20 verbs, 20 adjectives) from 3 lists of 200 non part-of-speech ambiguous[5] words obtained according to frequency criteria. The words chosen had word forms with comparable frequencies in the corpus, around the desired number of 50, so that, for each test word, all its contexts will be tested. These words were then proposed to 6 human judges who had to decide, for each of them, whether or not they were polysemous in the evaluation corpus.[6] A score was then attributed to each word by summing up the answers and the 20 words with the highest grade were selected. Altogether, full agreement on polysemy was achieved on only 4.5% of the words. Conversely, 40.8% of words were unanimously judged as having only one sense; the rest received mixed judgement.

The words are presented below. The numbers in brackets are firstly, the full number of senses (where each sense or subsense is treated as distinct), and second, the number of "top-level" sense distinctions. Petit Larousse dictionary entries are often hierarchical, and it is likely that, for many NLP tasks, top-level disambiguation is sufficient.

nouns barrage (6;2), chef (7;6), communication (4;2), compagnie (8;4), concentration (4;4), constitution (6;4), degré (17;4), détention (2;2), économie (8;2), formation (13;9), lancement (3;3), observation (7;3), organe (5;5), passage (12;2), pied (15;5), restauration (7;2), solution (4;2), station (7;3), suspension (8;3), vol (9;2)

adjectives biologique (3;3), clair (9;2), correct (3;3), courant (6;6), exceptionnel (2;2), frais (8;3), haut (10;3), historique (4;3), plein (11;9), populaire (4;4), régulier (12;2), sain (6;2), secondaire (10;3), sensible (11;9), simple (11;4), strict (4;4), sûr (5;5), traditionnel (2;2), utile (2;2), vaste (3;3)

verbs arrêter (8;3), comprendre (4;2), conclure (4;3), conduire (6;4), connaître (9;4), couvrir (16;3), entrer (9;4), exercer (6;6), importer (5;2), mettre (20;5), ouvrir (16;10), parvenir (4;4), passer (37;9), porter (26;8), poursuivre (5;5), présenter (13;4), rendre (12;3), répondre (9;3), tirer (30;9), venir (12;3)

Because the chosen words are the same ones as the one chosen within ARCADE it will be possible to adopt a multilingual perspective on WSD systems.

1.3. CHOOSING THE DICTIONARY

For French we used the Petit Larousse (Larousse95) dictionary. It is a monolingual dictionary of 54,900 entries which is widely available on CD-ROM. Most French speakers are familiar with this dictionary and therefore no particular training was required for human taggers.

There are many differences in the lexical resources used for the different languages. One difference is the average number of senses that are given by each dictionary for each part of speech (see Table I).

All else being equal, the more senses, the more difficult the disambiguation task.[8] Another difference concerns the way these resources have been built. For instance the Oxford English dictionary used within SENSEVAL is corpus and frequency based, while the Petit Larousse is a traditional dictionary with a clear encyclopedic bias. Corpus and frequency based dictionaries first display senses which have the highest frequency in corpora. This influences evaluation results in terms of comparison with the baseline as well as in terms of inter-tagger agreement.[9]

Also of importance is the fact that, unlike for the English exercise, for French and Italian there was no particular adequacy of the dictionaries to the corpora. Indeed the English experiment in SENSEVAL was in an especially favorable situation: contexts from the HECTOR corpus were tagged with the HECTOR dictionary based on the same corpus. The high inter-tagger agreement reached is in a accordance with Kilgarriff's (1998b) hope that such particular context would ease the taggers' task.

None of the French participants had the advantage of using their own dictionary/ontology. They all had to map them to the Larousse dictionary. This mapping has a lot of consequences on system evaluation, especially when participating systems had to map fine-grained dictionaries with the Petit Larousse.

1.4. TAGGING TEXT

In order to create an evaluation corpus, six human informants[10] were asked to semantically annotate the corpus. Each of the 60 words appeared in 50 different contexts which yielded 3000 contexts to be manually sense-tagged.[11]

Annotators were instructed to choose either zero, one, or several senses for each word in each context. (A question mark was used when none of the senses matched the given context. The question-mark sense was treated as an additional sense for each word, taking together all meanings not found in the dictionary.)

Because the Petit Larousse encodes more senses for verbs than for adjectives and nouns, annotators gave more senses per context for this part of speech. Still, it appeared that the average number of senses (used by a single judge in a given context) per part of speech is not very high. The average number of answers per word ranged from 1 to 1.3. Annotators used up to six senses in a single answer for a given context.

Table II. Inter-tagger agreement for French

	Full Max.	Full Min.	Pair Max.	Pair Min.	Pair Wei	Agree cor.
Nouns	44%	45%	72%	74%	73%	46%
Verbs	29%	34%	60%	65%	63%	41%
Adjectives	43%	46%	49%	72%	71%	41%

Agreement among annotators was computed according to the following measures:

– Full agreement among annotators. Two variants have been computed:

  - *Min*: counts agreement when judges agree on all the senses proposed for a given context
  - *Max*: counts agreement when judges agree on at least one of the senses proposed for a given context

– Pairwise agreement. Three variants have been computed:

  - *Min*: counts agreement when judges agree on all the senses proposed for a given context
  - *Max*: counts agreement when judges agree on at least one of the senses proposed for a given context
  - *Weighted*: Accounts for partial agreement using the Dice coefficient ($Dice = 2\frac{|A \cap B|}{|A|+|B|}$)

– Weighted pairwise agreement corrected for chance: using the Kappa statistic:[12]

$$k = \frac{Pobserved - Pexpected}{1 - Pexpected}$$

A kappa value of 1 indicates perfect agreement, and 0 indicates that agreement is no better than chance. (It can also become negative in case of systematic disagreement).

According to each of the above measures the inter-tagger agreement for French is an shown in Table II. The kappa values here are low, and indicate an enormous amount of disagreement between judges. Looked at word-by-word, the values range between 0.92 and 0.01; for some words, agreement was no better than chance.

This semantically hand-tagged corpus has been used for evaluation purposes only. Participating systems did not benefit from a training corpus either to train their system, or to tune their sense mappings. For training they were given an untagged corpus containing the test words. This was due to lack of time and resources.

2. Participating Systems and Evaluation Procedure

Four institutions participated with five systems in the French ROMANSEVAL exercise. They were:

EPFL Ecole Polytechnique Fédérale de Lausanne
IRISA Institut de recherche en informatique et Systèmes Aléatoire, Rennes
LIA-BERTIN Laboratoire d'informatique, Université d'Avignon, and BERTIN, Paris
XRCE Xerox Research Centre Europe, Grenoble

The first three systems are briefly described in the Appendix. The fourth has a paper of its own in this Special Issue.

The test procedure followed the steps described below:

- Each site received well in advance the raw corpus in order to get familiar with the format, and to interface, tune and train their systems as much as possible,
- a dry run was organised in order to check the procedures and evaluation programs,
- each site received the test words,
- each site returned the semantically-tagged test words.

Then each system was evaluated according to the metrics described in the next section.

3. Evaluation Metrics and Results

The measure of human inter-tagger agreement set the upper bound of the efficiency measures. It would be unrealistic to expect WSD systems to agree more with the reference corpus than human annotators among themselves.

Given the low human inter-tagger agreement, we tried to be as generous as possible. We treated the gold standard as the union of all answers given by all human taggers and adopted the following metrics:

- *Agree* counts matches between the system and gold standard, weighted by the number of proposed senses: $\frac{human \cap system}{system}$
- *Kappa* which is as above, corrected for chance agreement

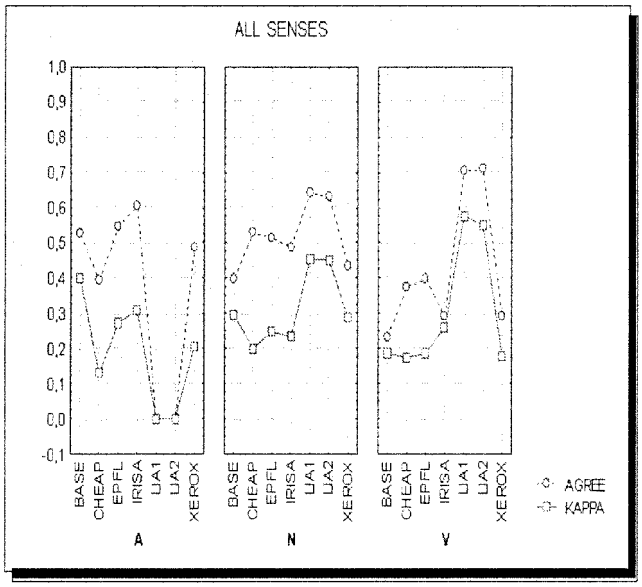

Figure 1. Results for adjective, nouns, verbs, all sense.

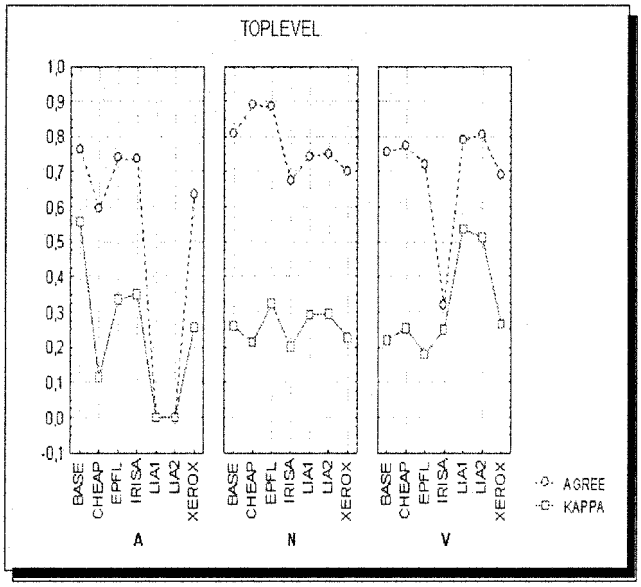

Figure 2. Results for adjective, nouns, verbs, top-level senses only.

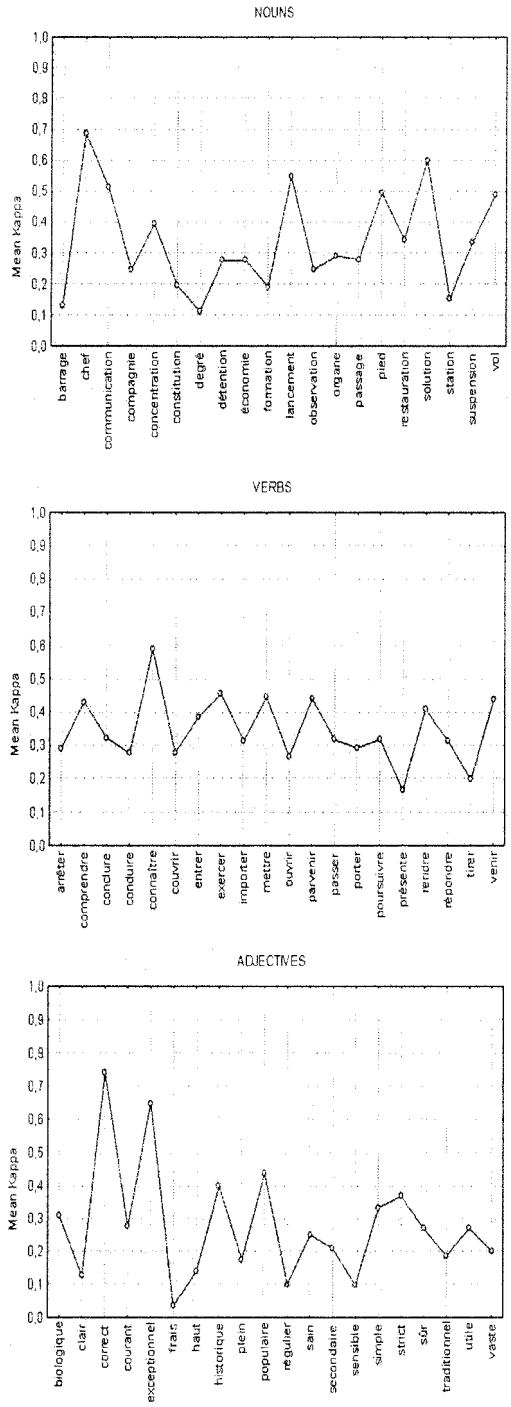
NOUNS
Mean Kappa
1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
barrage
chef
communication
compagnie
concentration
constitution
degré
détention
économie
formation
lancement
observation
organe
passage
pied
restauration
solution
station
suspension
vol
VERBS
Mean Kappa
arrêter
comprendre
conclure
conduire
connaître
couvrir
entrer
exercer
importer
mettre
ouvrir
parvenir
passer
porter
poursuivre
présente
rendre
répondre
tirer
venir
ADJECTIVES
Mean Kappa
biologique
clair
correct
courant
exceptionnel
frais
haut
historique
plein
populaire
régulier
sain
secondaire
sensible
simple
strict
sûr
traditionnel
utile
vaste

Figure 3.

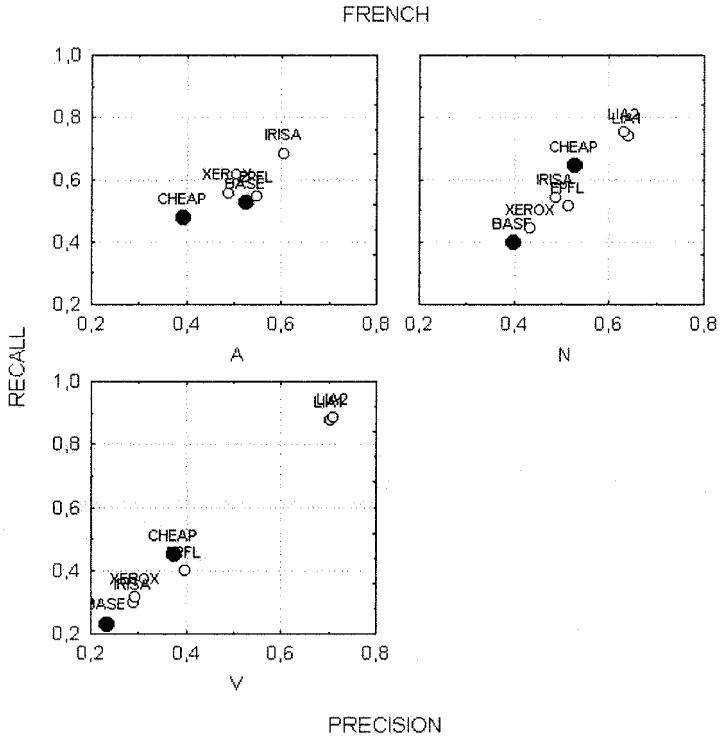

Figure 4. Results according to Precision and Recall.

In order to provide a line of comparison we also computed results for two baseline "trivial" systems which we called *Base* and *Cheap. Base* always chooses the first sense proposed in the Petit Larousse dictionary. (As already noted, one cannot assume that the first sense is the most common). *Cheap* is a variant of Lesk's method (Lesk86) which relies on finding the best overlap between a word in context and a dictionary definition.

The results are presented in Figures 1 and 2. The first considers all senses and subsenses as distinct. The second looked only at "top level" sense distinctions. For this calculation, all subsenses were treated as equivalent to the top level sense they fell under. Consider the case where, at the first level of the hierarchy, a word has senses 1 and 2, and sense 1 has subsenses *a* and *b*. Then, if the Gold Standard answer is 1*a* and a system response is 1*b*, then, in the top level calculation, the system response is correct, since both Gold Standard and system responses are equivalent to 1. (All other results figures are calculated on the basis of all-senses).

It is also interesting to explore which words were easier and which harder. Figure 3 shows, for each word, the average Kappa score for agreement between the system and the human taggers, for all seven systems. The graph indicates that some words presented far more problems than others.

All metrics have their own advantages and we decided to use the usual precision and recall figures as a secondary source, for ease of comparison with the English exercise. In our case precision is correct senses/total senses proposed and recall is correct senses retrieved/total senses in reference. The precision/recall results are shown in Figure 4.

The quantitative results still need to be refined (for example in terms of metrics) and discussed among participants. A qualitative study still needs to be undertaken, asking, for instance: what are the difficult words for systems, why are they difficult, what is the impact of sense mapping, what is the impact of the evaluation metrics, and what are the multilingual issues involved and the relationship with translation? We invite readers to participate in this process.

The overall exercise went very well thanks to the dedication and the motivation of all participants. We have been able to achieve a great deal in a little time and with few resources. We have laid the methodology and groundwork for a larger scale evaluation. Further experiments can include: the addition of new texts, the use of different dictionaries, and running an all-word tagging exercise as well as measuring efficiency of WSD in real tasks.

Notes

* I am especially grateful to Jean Véronis with whom I organised the ROMANSEVAL exercise. This paper is mainly a compilation of previous publications by Jean Véronis (see in particular Véronis 1998, Véronis et al. 1998, and Ide and Véronis 1998). Many thanks also to Marie-Hélène Corréard, Véronika Lux and Corinne Jean for comments on previous versions of the paper.

1 See http://www.lpl.univ-aix.fr/projects/romanseval

2 See http://www.lpl.univ-aix.fr/projects/arcade

3 The languages are: Dutch, Danish, English, French, German, Greek, Italian, Portuguese and Spanish.

4 MLCC stands for Multilingual Corpora for Cooperation; see MLCC, 1997.

5 This was to eliminate the need for POS tagging of the corpus, and the associated hand-validation.

6 The question asked was “According to you, does the word X have several senses in the following contexts?” They had three possible answers: “yes”, “no” and “I don’t know”.

7 These figures do not take into account the four POS ambiguous words.

8 This holds for both humans (according to Fellbaum, 1997) and automatic systems.

9 Fellbaum (1997) reports higher inter-tagger agreement when senses in dictionary entries are ordered according to their frequency of occurrence in the corpus, with the most frequent sense placed first.

10 The informants were linguistic students at Université de Provence.

11 We would like to thank Corinne Jean and Valérie Houitte for their help in coordinating the task.

12 The kappa statistic (Cohen, 1960; Carletta, 1996) measures the “true” agreement, i.e. of the proportion of agreement above what would be expected by chance. The extension of kappa for partial agreement, as proposed in Cohen (1968), was used.

Appendix: Brief Descriptions of Three ROMANSEVAL WSD Systems for French

IRISA WSD SYSTEM

Ronan Pichon and Pascale Sébillot

The WSD system that we have developed is based on a clustering method, which consists of associating a contextual vector with each noun, verb and adjective occurrence in the corpus (not only with the 60 words of the test) and in aggregating the most "similar" elements at each step of the clustering. The contents (the words and their frequencies) of the clusters in which test occurrences appear are then used to choose the Petit Larousse most relevant sense(s).

Some problems

Concerning verbs, results are not very good. In fact, we have stopped the search of the meanings of the test occurrences. One explanation: there are greedy clusters which "swallow" a lot of verbs; therefore, the interpretation of the class is impossible. This greedy cluster phenomenon also happens for other categories, but it is very accentuated for the verbs. A "normal" class contains about 30–50 elements (that means about 6 to 8 distinct lemmas); a greedy cluster can contain 2000 elements; the maximal cluster for verbs that we have found had 20000 elements.

Different contexts for nouns, verbs and adjectives will probably improve the results. For example, we think that for adjectives, it will be better to consider a closer context (better than the whole sentence).

WSD System of Laboratoire Informatique D'Avignon and Bertin Technologies

Claude de Loupy, Marc El-Bèze and Pierre-François Marteau

Due to the lack of a training corpus in ROMANSEVAL, it was impossible to use the automatic method we have implemented for English SENSEVAL (see our full paper in this volume for a description of the SCT method). This has led us to perform a semi-automatic experiment for the French task. This procedure makes use of the test corpora.

For each word to be tagged, the set of sentences was submitted to the same automatic preprocessing as for the English task. We then manually extracted some patterns and assigned them to one or more senses, where possible. When more than one sense could be attached to a corpus instance, the instance was duplicated for each sense.

Some omissions in the definitions caused problems for the manual assignment of sense. For instance, the very frequent *chef-d'oeuvre* was not represented.

This work was done for the French corpus and the English counter-part. Moreover, samples have been extracted from the definitions. The confidence of a sample depends both on the number of times it appears and an arbitrary score given by a human judge.

The very good results we have obtained in that way may be considered as an upper bound of French WSD performances for an automatic system using the SCT method and a very large coverage bilingual corpus.

WSD System of EPFL, Swiss Federal Institute of Technology

Martin Rajman

The EPFL team proposed a disambiguation model based on Distributional Semantics (DS), which is an extension of the standard Vector Space (VS) model. The VS model represents a textual document d_n as a vector $(w_{n1}, \ldots, w_{nM})$, called *lexical profile*, where each component w_{nk} is the *weight* (usually the frequency) of the *term* t_k in the document (terms are here various predefined textual units, such as words, lemmas or compounds). The DS model further takes the co-frequencies between the terms in a given reference corpus into account. These co-frequencies are considered to provide a distributional representation of the "semantics" of the terms. In the DS model, each term t_i is represented by a vector $c_i = (c_{i1}, \ldots, c_{iP})$ (*co-occurrence profile*), where each component c_{ik} is the frequency of co-occurrence between the term under consideration t_i and the indexing term t_k. The documents are then represented as the average vector of the co-occurrence profiles of the terms they contain

$$d_n = \sum_{i=1}^{M} w_{ni} c_i$$

In the DS-based disambiguation model, the context of any ambiguous word and each of its definitions is first positioned in the DS vector space. Then, the semantic similarity between a context (represented by a vector C) and each of the definitions (represented by a vector D_i) is computed according to a similarity formula such as cosine similarity ($cos(C, D_i) = \frac{C.D_i}{\|C\|\|D_i\|}$) and the definition corresponding to the higher similarity is selected.

References

Carletta, J. "Assessing agreement on classification tasks: the kappa statistic" *Computational Linguistics*, 22(2) (1996), 249–254.

Cohen, J. "A coefficient of agreement for nominal scales" *Educational and psychological Measurement*, 20, (1990), 37–46.

Cohen, J. "Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit" *Psychological Bulletin*, (70)4 (1968), 213–220.

Fellbaum, C., Grabowski, J. and S. Landes. *Analysis of Hand-Tagging Task* in: Proceedings of ANLP, Workshop on Tagging Text with Lexical Semantics, Why, What and How?, Washington D.C., April 1997.

Ide, N. and J. Véronis, "Introduction to the special issue on word sense disambiguation: the state of the art" *Computational Linguistics*, 24(1) (1998), 1–40.

Kilgarriff, A. "SENSEVAL: an exercise in evaluating word sense disambiguation programs" Proceeding of LREC, Granada, May 1998, pp: 581–588.

Kilgarriff, A. "Gold standard datasets for evaluating word sense disambiguation programs" *Computer Speech and Language*, 12(4) (1998b), 453–472.

Le Petit Larousse illustré – dictionnaire encyclopédique Edited by P. Maubourguet, Larousse, Paris, 1995.

Lesk. *Automated sense disambiguation using machine-readable dictionaries: how to tell a pine cone from an ice-cream cone* in: Proceedings of the 1986 SIGDOC Conference. Toronto, June 1986, New York: Association for Computing Machinery, pp. 24–26.

Multilingual Corpora for Co-Operation. *Distributed by ELRA* 1997.

Jean Véronis. *A study of polysemy judgements and inter-annotator agreement* in : Programme and advanced papers of the SENSEVAL workshop, Herstmonceux Castle, September 1998.

Véronis, J., Houitte, V. and C. Jean. *Methodology for the construction of test material for the evaluation of word sense disambiguation systems* in : Workshop WLSS, Pisa, April 1998.