

SENSE TAGGING: DON'T LOOK FOR THE MEANING BUT FOR THE USE

Jean Véronis

Université de Provence

29, Avenue Robert Schuman, 13100 Aix-en-Provence (France)

Jean.Veronis@up.univ-mrs.fr

<http://www.up.univ-mrs.fr/~veronis>

ABSTRACT

Automatic sense-tagging is one of the next challenges that corpus linguists have to face. So far, results are modest, despite numerous efforts, and sense-tagging appears as a vexing task. Difficulties stem from various sources, and in particular the extraction of disambiguating information from the context. However, one of the main problems comes upstream of the disambiguating process and lies in the sense inventory itself. Most tagging efforts use traditional dictionaries as the reference sense list, or machine-oriented resources such as *WordNet* which do not differ significantly from traditional dictionaries in terms of sense division. This paper shows that human taggers perform very poorly when they are given a traditional dictionary as reference, and that machines should therefore not be expected to perform better using the same kind of resource. A detailed analysis reveals the lack of distributional criteria in dictionary entries: traditional dictionaries are chiefly concerned with the definition of meaning, and not with the surface clues (syntactic, collocational, etc.) that are required to match a given sense with a given corpus occurrence. It is argued that no fundamental progress can be made until large-scale lexical resources have been built that incorporate extensive distributional information.

Keywords: sense-tagging, polysemy judgements, interannotator agreement, dictionaries, distributional information

1. INTRODUCTION

Now that part-of-speech (POS) tagging has become a well-mastered technique, and that POS-tagged corpora are available in increasingly large quantities, sense tagging is the next challenge that corpus linguists have to face. Applications of

sense tagging are numerous, in fields such as information retrieval or machine translation, and sense-related information constitutes a bottleneck for further analysis, such as syntactic parsing or anaphora resolution.

However, sense tagging is a much more vexing problem than POS tagging. Word sense disambiguation (WSD) has been recognised as a central (and difficult) issue in the very first paper on computer treatment of language, Weaver's memorandum [17]. Since then, there has been continuous research on WSD and an impressive array of methods has been proposed — and occasionally rediscovered over the years — but we are still very far from all-purpose sense-tagging programs with 95% accuracy or more, as it is the case for POS tagging (for a recent survey, see [11]).

Despite this quite substantial body of research, we have very little sense of how well humans perform on the sense-tagging task. Only a handful of researchers have tackled the subject [1, 2, 3, 4, 5, 12], and these studies are informal and/or involved only a few words or annotators. [9] is more detailed, but difficult to interpret due to the lack of comparison of raw figures with a chance baseline (see discussion below).

I find however extremely surprising that an entire field of research can develop without a clear view of human performance in the area. If it turned out that humans perform poorly on sense-tagging, the entire idea of mimicking or replicated the task by machines would have to be reconsidered. It is true that machines can perform some tasks better, and mainly faster, than humans; it is however rarely the case when meaning and context are involved, and in general in all areas where intelligence is required. In sense-tagging even more than in other linguistically-oriented tasks, human performance (of course in terms of accuracy, not speed) can be

safely taken as an upper bound for any kind of automation.

In this paper, I will precisely show that humans do poorly on sense annotation – or at least that they disagree widely in their judgements, which I take to amount to the same thing. I will first report on an experiment which shows that humans differ greatly in their assessment of whether a word is polysemous or not in a given corpus (Section 2). I will then report on a sense-tagging task in which annotators had to assign senses from a common dictionary to corpus examples (Section 3). Agreement between judges is also very low, virtually not greater than chance for some words¹.

These results, obtained on reasonably large-scale data (36000 corpus examples in the first experiment, 3724 in the second), seem to indicate that it is pointless to expect machine sense-tagging to reach 95% accuracy or more, as for POS tagging, under the same experimental conditions. In Section 4, I analyse the main reasons that lead to the observed disagreement, and show that ordinary dictionaries are inappropriate for the task. The main criticism consists in the lack of distributional criteria in dictionary entries: traditional dictionaries are chiefly concerned with the definition of meaning, and not so much with the surface clues (syntactic, collocational, etc.) that are required to match a given sense with a given corpus occurrence. This is also true of recent computer-oriented resources such as *WordNet*, and I argue that no fundamental progress can be made until large-scale lexical resources have been built that incorporate extensive distributional information.

2. EXPERIMENT ONE

2.1. Material

The first experiment aims at evaluating the polysemy judgements given by human subjects. 600 different word types (200 adjectives, 200 nouns and 200 verbs) were selected from the one-million word French part of the “JOC” corpus composed of written questions asked by members of the European Parliament on a wide variety of topics (health, education, environment, economy, etc.) and corresponding answers from the

¹ These experiments have been carried out in the ROMANSEVAL project, the Romance language counterpart of the SENSEVAL evaluation exercise [13]. See: <http://www.up.univ-mrs.fr/~veronis/romanseval>, <http://www.itri.bton.ac.uk/events/senseval>

European Commission².

The 600 test words were drawn from the corpus on frequency criteria, so that each word has around 60 occurrences, which yields ca. 36000 corpus examples in total. The set of examples for each word was organised in the form of concordance lines printed on a separate page. The concordances were bound in three volumes of 200 pages each (one for each POS).

2.2. Procedure

The concordance volumes were given to 6 different informants. The question asked to them was “According to you, does the word X have one sense or several senses in the following contexts?”, and they were invited to tick the corresponding box or a “don't know” box. Informants were fourth-year linguistic students, but had never received any specific lexicographic training. They received a small payment for the task, and could accomplish the task in free time, but had to give the results back within a week.

2.3. Results

Somewhat to my surprise, none of the informants found the task difficult. This is confirmed by the rate of “don't know” responses, which is particularly low (4.05%). Most words were judged as having only one sense (73.0%), but there are substantial differences among categories: nouns are judged more polysemous than verbs, in turn judged more polysemous than adjectives (Figure 1). This difference is statistically significant at $p < 10^{-4}$ ($\chi^2 = 67.87$; $v = 4$).

However, despite the low rate of “don't know” responses, agreement between informants is also low. It seems that individual informants have no difficulty in making spontaneous judgements, but different informants tend to make different judgements. Altogether, all six judges agreed only in 45 % of the cases (Table 1). Full agreement on polysemy was achieved on only 4.5% of the words. Conversely, 40.8% of words were judged as having only one sense by all judges — the rest receiving mixed judgements. This measure is striking, but is biased with the number of judges: it tends to decrease asymptotically towards zero as the number of judges increases. I therefore used

² This corpus, collected and prepared within the MLCC-MULTEXT projects, was chosen because it exists in nine parallel versions, which enabled simultaneously a study of relationships between sense tagging and translation in the ARCADE project [16].

also pairwise agreement, which is a stable measure. Table 1 shows that pairwise agreement reaches 73%.

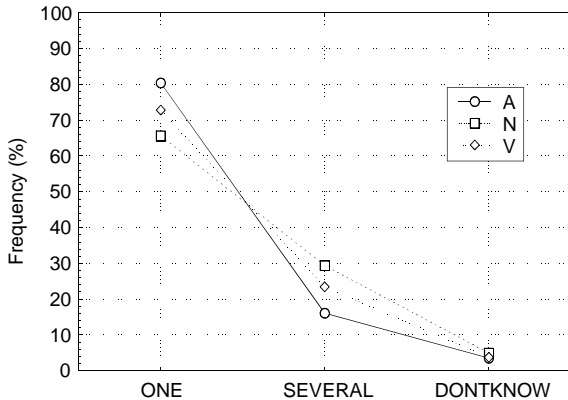


Figure 1. Polysemy judgements per POS category

POS	Full	Pairwise	k
A	0.54	0.78	0.67
N	0.37	0.68	0.36
V	0.45	0.74	0.37
All	0.45	0.73	0.49

Table 1. Measures of agreement on polysemy judgements

Agreement measures should however always be corrected for chance agreement: it is obvious that some agreement would be reached even if annotators were responding at random. The kappa statistics, proposed by [7] (see also [6]) is a standard measure of “true” agreement, i.e. of the proportion of agreement above what would be expected by pure chance:

$$k = \frac{p_{\text{observed}} - p_{\text{expected}}}{1 - p_{\text{expected}}}$$

where p_{expected} is computed on the basis of the marginal frequencies. This coefficient ranges between 0 when agreement is no better than chance and 1 when there is perfect agreement (it can also become negative in case of systematic disagreement).

The average k values are low, since they range from 0.37 to 0.67 depending on the POS category, with a value of 0.49 for all categories combined. Such values are generally considered in the literature as indicative of poor agreement (see [14]). The correction due to k is particularly noticeable: adjectives and verbs have very close absolute values of pairwise agreement, although

“true” agreement is almost twice as important for adjectives as for verbs (this is due to the higher rate of “one sense” responses for adjectives).

These differences among POS categories remain to be explained. It is interesting to note that the dictionary does not have the same perception of polysemy differences among POS. Table 2 gives the average number of senses for the test words in the *Petit Larousse* grouped by POS. Adjectives have less senses than other categories, which is consistent with the polysemy judgements. However, verbs have more senses than nouns, although they were felt less polysemous by informants.

POS	Senses
A	2.4
N	4.6
V	5.8

Table 2. Average number of senses of the test words in the dictionary

3. EXPERIMENT TWO

3.1. Material

A score was then attributed to each of the 600 words of Experiment One by summing up the responses (1=several senses, 0=don’t know, -1=one sense). The 20 words with the highest score within each POS were selected as test words for Experiment Two (i.e. 60 words in total). These words occurred 3724 times altogether in the JOC corpus.

3.2. Procedure

The 3724 occurrences of the 60 selected words were sense-tagged independently by six different annotators according to the *Petit Larousse* dictionary sense list. The *Petit Larousse* is probably the most widespread French dictionary, known and used by most educated people. The annotators were fourth-year linguistic students with no specific training in lexicography, different from the set of informants used in the previous task, in order to avoid cross-task biases. They were also paid for the task.

For each of the 3724 occurrences of the selected words in the corpus, the corresponding paragraph of context was displayed in a spreadsheet, with the word to tag highlighted. All occurrences of the same word were grouped on the same spreadsheet,

and annotators were asked to mark the senses in additional columns. They had therefore all occurrences of the same word available on the screen. They could mark them in any order, and revise their judgement as they were going along.

Annotators were instructed to chose either one sense, or several if they felt that more than one were appropriate in the given context. They could also choose no sense at all, if they felt that no sense in the dictionary was appropriate in the context. In the latter case, they were instructed to write down a question mark in the sense column. In the subsequent study, the question mark was treated as an additional sense for each word, grouping all meanings that were not found in the dictionary.

3.3. Results

The annotators gave more senses per context for verbs than for adjectives and nouns (Table 3, column *Nsen*). This is likely to be a result of the larger number of senses offered for verbs in the dictionary (see discussion above and Table 2). The average number of senses (used by a single judge in a given context) per POS category is not very high, which shows that annotators have a tendency to avoid multiple answers (as said above, the “no sense” answer is counted as a special sense). However, the average per POS category masks important differences between words: the average number of responses per word ranges from 1 to to 1.311 (verb *comprendre*). In some cases, annotators used up to six senses in a single response for a given context.

Agreement was computed according to several measures (summarised in Table 3):

(1) Full agreement among the six annotators.

Two variants were computed:

<i>Min</i>	Counts agreement when judges agree on all senses proposed for a given context
<i>Max</i>	Counts agreement when judges agree on at least one of the senses proposed for a given context

The difference between the *min* and *max* measures is not very important, apart from for a few words (*sûr, comprendre, importer*). This is due to the fact that the average number of senses given by judges is close to 1 (Table 3, column *Nsen*). Of course, these measures are biased with the number of judges, as mentioned above. It is however striking to note that for some words (*correct, historique, économie, comprendre*) there was full

agreement on none of the contexts for that word.

(2) Pairwise agreement.

Three variants were computed:

<i>Min</i>	Counts agreement when judges agree on all senses proposed for a given context
<i>Max</i>	Counts agreement when judges agree on at least one of the senses proposed for a given context
<i>Weighted</i>	Accounts for partial agreement using the Dice coefficient: $Dice = 2 \frac{ A \cap B }{ A + B }$

Again, there is not much difference between the measures, apart from for a few words, interestingly enough not exactly the same as before (*chef, comprendre, connaître*).

(3) Agreement corrected for chance.

The measures above are not completely satisfactory, because they do not enable comparison of observed agreement and agreement that would be obtained by pure chance. The *k* statistics mentioned above enables such a comparison. In order to account for partial agreement, *k* was computed on the weighted pairwise measure using the extension proposed in [8].

It is interesting to note that *k* ranges between 0.92 (noun *détention*) and 0.007 (adjective *correct*). In other terms, there is no more agreement than chance for some words. The average *k* values are low, below 50%, which indicates a great amount of disagreement among judges.

<i>POS</i>	<i>Nsen</i>	<i>Full</i>		<i>Pairwise</i>			<i>k</i>
		<i>min</i>	<i>max</i>	<i>min</i>	<i>max</i>	<i>wgh</i>	
A	1.013	0.43	0.46	0.69	0.72	0.71	0.41
N	1.009	0.44	0.45	0.72	0.74	0.73	0.46
V	1.045	0.29	0.34	0.60	0.65	0.63	0.41

Table 3. Agreement measures per POS category

It is possible that the sense divisions contained in dictionaries are too fine-grained for NLP purposes. This argument has been made many times, and many WSD systems have been restricted to homograph level or broad sense distinctions.

In order to test this hypothesis, I have computed the degree of inter-annotator agreement when their responses are reduced to the top-level distinctions made in the dictionary (French dictionaries are much more hierarchical than English ones, due to different lexicographic traditions). The

improvement was measured as the reduction of disagreement once corrected for chance, i.e.:

$$\Delta = 1 - \frac{1 - k_2}{1 - k_1}$$

The results are disappointing: the disagreement reduction is only of 8% for adjectives and 9% for verbs. It is higher for nouns, but reaches only 25% (Table 4).

POS	Nsen	Full		Pairwise			k	Δ (%)
		min	max	min	max	wgh		
A	1.010	0.55	0.57	0.78	0.80	0.79	0.46	7.9
N	1.003	0.70	0.70	0.86	0.86	0.86	0.60	25.2
V	1.018	0.54	0.56	0.77	0.80	0.79	0.46	8.9

Table 4. Agreement on top-level divisions

4. DISCUSSION

4.1. Summary of results

Experiment One showed that judges disagree widely on whether a given word is polysemous or not in a corpus. Experiment Two showed that they also disagree enormously when they have to tag corpus examples according to the sense list provided by a common dictionary. The rate of disagreement is so important that for some words, there was no more agreement than what would be obtained by mere chance. It cannot be argued that sense distinctions are too fine-grained for WSD, since, somewhat surprisingly, most disagreement between annotators spans across the top-level divisions of entries.

These results shed a new light on automated sense-tagging. The dictionary chosen (*Petit Larousse*) is not at fault. It is a very respectable medium-size dictionary which builds on a century and a half of lexicographic tradition. I am convinced that the results would be similar with any other traditional dictionary.

4.2. An example of difficulty

The word *degré* (=degree) exemplifies the type of difficulty that annotators are faced with. At the top level, the dictionary gives the following divisions and definitions (I translate roughly and skip the sub-senses for lack of space):

DEGRÉ. **I.** Literary: step/stair. **II.** each of the intermediary state leading from one state to another. **III.** relative intensity (of an affective, moral or pathological state). **IV.** each of the divisions, corresponding to a unit,

of a scale of measurement.

If divisions I and IV are (almost) straightforward, the distinction between II and III is extremely confusing for annotators. In sentences such as:

...les trois principaux **degrés** de cette élimination étatique: le génocide, la déportation en masse et l'assimilation forcée... (*...the three main steps/levels of this state elimination: genocide, mass deportation and forced assimilation...*)

Ils s'inquiètent de ce qu'ils perçoivent comme un **degré** croissant d'anarchie... (*they point out their concern about what they perceive to be an increasing level of lawlessness...*)

it is very unclear whether *degré* refers to “an intermediary state leading from one state to another” or a “relative intensity of an affective, moral or pathological state”³.

In this example, it would however be very easy to split uses according to syntactic criteria. A first set of uses accepts cardinal determiners (*un, deux, trois* / =one, two, three, etc.) as well as ordinal qualifiers such as *premier, second, dernier* (=first, second, last):

...les trois principaux **degrés** de cette élimination étatique → le **premier** degré, le **second** degré, etc. (*the first step, the second step, etc.*)

On the other hand, another, disjoint, set of uses accepts intensifying qualifiers whose prototype is the *fort/faible* (=high/low) pair:

un **degré** croissant d'anarchie → un **faible** degré, un **fort** degré d'anarchie (*a low level, a high level of lawlessness*)

Other adjectives in the paradigm are *alarmant* (=alarming), *élevé* (=high), *minimal* (=minimal), *différent* (=different), *croissant* (=increasing), etc.

In other words, one set of uses is discrete and countable, the other set is continuous and intensifiable. Annotators would have little trouble using these tests, and machines could use the presence of the appropriate adjectives or determiners as a reliable disambiguating clue. However, none of the French dictionaries that I examined use or mention this rather simple

3 *WordNet* 1.6 proposes a similar distinction for the English *degree*, resulting in exactly the same kind of indecision: **1.** a position on a scale of intensity or amount or quality (e.g. : “a moderate degree of intelligence” etc.) **2.** a specific identifiable position in a continuum or series or especially in a process (e.g. : “a remarkable degree of frankness” etc.). It is hard to see why “degree of intelligence” and “degree of frankness” should be treated differently.

syntactic property. Worse yet, the *Petit Larousse* definitions II and III which at first glance could correspond to this division are in fact at odds with it, as the examples and sub-senses reveal.

4.3. From meaning to use

It is always easy to point out weaknesses and errors in entries, in any dictionary. However, my criticism is of a different nature. I am not trying to spot occasional flaws, but questioning the very style and organisation of entries. In almost all of the 60 words used in the Experiment Two, the definitions (which are after all the only information that annotators have at their disposal in order to match individual senses with corpus contexts) do not contain enough clues to perform the task safely. Worse yet, the division of entries itself rarely takes into account (and is often contradictory with) distributional facts. Annotators all commented on the vagueness of definitions and lack of clear-cut distinctions among senses, which they had never fully realised until they were confronted with the systematic tagging task. This vagueness is particularly apparent in abstract, very polysemous words, such as *degré*, *économie* (=economy, economics, saving, etc.), *communication* (=communication, report, telephone call, etc.), *formation* (=education, training, forming, formation, etc.), which constitute a large part of most texts.

The reason for this is probably to be found in a lexicographic tradition that has its roots in the Aristotelian approach to meaning and definition. For several centuries, dictionaries have primarily tried to give an account of meaning, not of usage (apart from occasional indications of register or domain). As a result, they rarely provide the surface distributional clues that would enable sense discrimination. Only recently some dictionaries (e.g. *Cobuild*, *LDOCE*, *OALD*) have started incorporating detailed syntactic, collocational and paradigmatic information, using corpus evidence instead of lexicographer's introspection. This trend is however very new compared to the four-century dictionary building tradition, and distributional information in modern dictionaries is still very far from being systematic and precise enough for computer use. More computer-oriented resources such as *WordNet* unfortunately also almost totally lack this type of information.

A major departure from traditional lexicography has to be made if we want to accomplish

significant progress in sense tagging and other sense-related activities. We have to radically shift from the description of meaning to that of the uses. The dictionaries cited above go one step in that direction, but distributional information is still very much conceived as an add-on on top of traditional foundations. I will take the radical stance that distributional information can provide the very foundations of dictionary organisation, and that entries can be divided up into coherent *usage classes* — that one can think about as *senses* — on the sole basis of that information, with no resort to meaning analysis and the more or less introspective or psychological considerations that such analysis usually requires.

Although never implemented fully and systematically in lexicographic work and computer applications, this point of view is not entirely new. It can be tracked back at least to Meillet [15]:

“Le sens d'un mot ne se laisse définir que par une moyenne entre [ses] emplois linguistiques.” (*The sense of a word is defined only by the average of its linguistic uses.*)

Wittgenstein [18] popularised a similar position in the well-known aphorism⁴:

“Don't look for the meaning, but for the use”,

and Harris made it part of his linguistic programme, by defining “meaning as a function of distribution” [10:155-158].

4.4. Distributional information

In this section, I will show that entries can be divided up using various types of distributional information with no resort to meaning analysis. At the same time, this information is of primary importance for human annotators and tagging systems. I will use the word *barrage* (=dam, blocking, roadblock, barrier, etc.) as an example, since while being polysemous, it is not too complex for the space constraints of this paper.

4.4.1. Syntactic information

Syntax provides an extremely powerful tool for splitting entries. For example, some uses of *barrage* are an active nominalisation of the verb *barrer*, others are not. By active, I mean that the nominalisation is a strict synonym of the verb, by which it can be replaced by changing the

⁴ in the *Philosophische Untersuchungen* – he had previously defended the opposite view in the *Tractatus*.

construct. At the same time, the valency of the verb is kept in the noun; in particular, the noun has (or can have) an agent (corresponding to the verb subject) :

le barrage de la rivière [par les castors] (*the blocking up of the river [by the beavers]*) → les castors ont barré la rivière (*the blocking up of the river [by the beavers]*) → *the beavers blocked up the river*

This use, although given as the core sense by most dictionaries (“the act of blocking”) is in fact very rare in corpora. It must not be confused with the other uses of *barrage* which, although etymologically formed through a nominalisation, have lost the direct relationship to the verb.

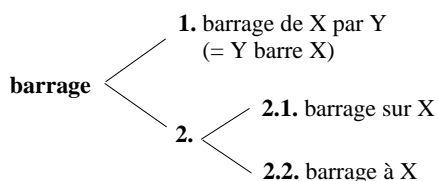
le barrage sur le Rhône (*the dam on the Rhône river*)
→* quelqu'un barre le Rhône

This second set of uses has developed its own valency over the centuries, in different ways: a first subset takes a complement introduced by the preposition *sur* (=on) and a second subset a complement introduced by *à* (=to):

le barrage sur le Rhône, sur l'autoroute (*the dam on the Rhône river, the roadblock on the highway*)

le barrage à la loi sur l'avortement (*the opposition to the abortion law*)

At this stage, the entry is structured as follows:



4.4.2. Paradigmatic information

Another type of information is of paradigmatic nature. For example, one set of uses of *barrage* has a hypernym, *ouvrage* (≈civil engineering structure, no exact translation), while the others have no hypernym. This assertion may seem odd, since a long Aristotelian tradition, and the recent upsurge of ontology development, have contributed to the widespread feeling that all words, and all senses of these words, have a hypernym and that the lexical space is organised as a giant taxonomy. It all depends, of course, on what we want to call hypernym, and how lax we want this definition to be. In the distributional perspective that I am advocating here, I will adopt a very strict view of hyperonymy and restrict it to

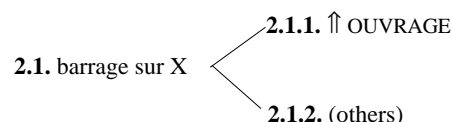
the only cases where there is syntagmatic evidence of the relationship, for instance in enumerations or anaphoras:

les **ouvrages** "lourds" du GAP, comme le **barrage** Atatürk ou les tunnels jumeaux d'Urfa... (*the "heavy" structures built by the GAP, such as the Atatürk dam or the twin tunnels of Urfa...*)

le **barrage** d'Assouan ... cet **ouvrage** géant, monstrueux (*the Hassouan dam ... this giant, monstrous structure*)

The usual tests such as “is a kind of” simply do not work on most senses, unless we accept to distort language use in the laxest and most unnatural way. For example, it is impossible to find a natural filler for the pattern “is a kind of ...” for the sense “roadblock” of *barrage*. Despite extended search in large corpora, I was unable to find any syntagmatic evidence of a term that could be a satisfactory hypernym.

This differential behaviour with respect to hypernyms enables us to subdivide further the use 2.1 (*barrage sur*):



Other types of paradigmatic information can be used as well, such as the presence or absence of synonyms. Here again, I restrict the notion to strict synonyms, i.e. which can be substituted with no change or loss in a given context. The substitutability can be established by assessing whether the contexts of the candidate synonyms are similar in terms of distribution (valency, etc.). For example, use 2.2 of *barrage* (*barrage à*) accepts a strict synonym, *obstacle*:

la volonté de faire **barrage** (=obstacle) à une probable expansion du communisme (*the desire to block a probable expansion of communism*)

while no other subset has any strict synonym. This does not enable us to subdivide classes further, but confirms that 2.2 should be a separate class.

4.4.3. Collocational information

Collocational information is at the crossroads of syntagmatic and paradigmatic information. On one hand it has a syntactic base, since it expresses the “preferences” of syntactically bound terms (verb-object, etc.); on the other hand, it enables the grouping of words in paradigms that can fulfil a given syntactic place (e.g.: *read a <book, newspaper, letter, report, ... >*). This information

can relatively easily be extracted from corpora using grammatical and statistic filters, and manual checking. In the *barrage* example, this information is quite productive. It does not impose further dividing, but strongly confirms the classes established so far. For instance, frequent verbs with *barrage 2.1.1* as object are *construire* (=build), *édifier* (=edify), *démolir* (=demolish), etc., while verbs associated with *barrage 2.1.2* are a totally disjoint subset: *dresser* (=put up), *franchir* (=cross), *démanteler* (=dismantle), etc.

Figure 2 shows the most frequent collocations associated with the various classes of uses for *barrage*, roughly grouped by syntactic category. Glosses are provided between square brackets only for the sake of readability. It is important to note that the meanings that they are referring to were not used in the splitting process, which was done only on distributional grounds. However, interestingly enough, the classes of uses obtained this way are also coherent from a cognitive point of view.

5. CONCLUSION

In this paper, I have shown that interannotator agreement is very low in a straightforward sense-tagging task, using a traditional dictionary. For some words, agreement was no better than chance. A careful analysis reveals that the main difficulties come from the lack of distributional information in traditional dictionaries. Building on several centuries of lexicographic tradition, dictionaries mainly attempt to describe and define meaning, and rather marginally give information about word uses and distributional data. Only very recently lexicographers have started making systematic use of corpora, and dictionaries still do not contain systematically the surface clues (syntactic, collocational, etc.) that are required to match a given sense with a given corpus occurrence. I tried to show that distributional information can provide the very foundations of dictionary organisation, and that entries can be divided up into coherent *usage classes* — that one can think about as *senses* — on the sole basis of that information, with no resort to meaning analysis and the more or less introspective or psychological considerations that such analysis usually requires. I am convinced that large scale lexicons organised this way, and containing detailed distributional information are necessary in order for fundamental progress to be made in sense tagging and other sense-related language processing.

BARRAGE

1. [act of blocking] *barrage de X par Y* (+Nomin. < BARRER)

Barrage de qqchose (par qqchose/qqun)

2. (–Nomin.)

2.1. *barrage sur X*

2.1.1 [dam] (Ÿ OUVRAGE)

Barrage sur un fleuve, une rivière.

Grand, futur, gigantesque, coûteux barrage.

Barrage hydraulique, hydroélectrique, gigantesque, monumental, ultramoderne.

Barrage X, barrage de X: (e.g. *le barrage Atatürk, le barrage d'Assouan*).

Le lac, les eaux, les turbines, les lâchages, le chantier, la construction, l'achèvement, le financement, l'inauguration, la rupture d'un barrage.

Un projet, un programme, un système, une série de barrage(s).

Construire, édifier, financer, détruire un barrage.

Le barrage engloutit (des forêts, des bâtiments...), contrôle (le débit, les inondations), alimente (qqchose en énergie, en électricité), fournit (de l'énergie, de l'électricité à qqchose), se rompt, se brise.

En amont, en aval, en contrebas du barrage.

2.1.2 [roadblock]

Barrage sur une route, un chemin, une voie ferrée.

Barrage de police, de l'armée, des douanes.

Barrages routiers, policiers, militaires, filtrant, volant.

Dresser, établir, démanteler, rencontrer, éviter, franchir un barrage.

(Route) être hérissé(e) de barrages.

Des barrages se dressent, foisonnent sur (une route)

2.2. [opposition] *barrage à* (= OBSTACLE)

Barrage à qqchose.

Barrage idéologique, systématique.

Constituer, dresser un barrage à (qqchose).

Faire l'objet d'un barrage.

Faire barrage à qqchose.

Figure 2. Example of entry based on distributional information

6. REFERENCES

- [1] Ahlswede, T. E. (1993). Sense Disambiguation Strategies for Humans and Machines. *Proceedings of the 9th Annual Conference on the New Oxford English Dictionary*, Oxford, England, September, 75-88.
- [2] Ahlswede, T. E. (1995). Word Sense Disambiguation by Human Informants.

Proceedings of the Sixth Midwest Artificial Intelligence and Cognitive Society Conference, Carbondale, Illinois, April 1995, 73-78.

[3] Ahlswede, T. E., & Lorand, D. (1993). The Ambiguity Questionnaire: A Study of Lexical Disambiguation by Human Informants. *Proceedings of the Fifth Midwest Artificial Intelligence and Cognitive Society Conference*, Chesterton, Indiana, 21-25.

[4] Amsler, R. A., & White, J. S. (1979). *Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries*. Final report on NSF project MCS77-01315. University of Texas at Austin, Austin, Texas.

[5] Bruce, R., & Wiebe, J. (1998). Word sense distinguishability and inter-coder agreement. *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98)*. Association for Computational Linguistics SIGDAT, Granada, Spain, June 1998.

[6] Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249-254.

[7] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

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

[9] Fellbaum, C., Grabowski, J., & Landes, S. (1998). Performance and confidence in a semantic annotation task. In C. Fellbaum (Ed.), *WordNet: An electronic database* (pp. 217-237). Cambridge, Massachusetts: The MIT Press.

[10] Harris, Z. S. (1954). "Distributional Structure." *Word*, 10, 146-162.

[11] Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1), 1-40.

[12] Jorgensen, J. (1990). The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19, 167-190.

[13] Kilgarriff, A. (1998). SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. *Proceedings of the Language Resources and Evaluation Conference* (pp. 581-588). Granada, Spain.

[14] Krippendorff, K. (1980). *Content Analysis: An introduction to its Methodology*. Sage

Publications.

[15] Meillet, A. (1926). *Linguistique historique et linguistique générale*. Vol. 1. Champion, Paris, 351pp. (2nd édition).

[16] Véronis, J. (2000). Evaluation of parallel text alignment systems: the ARCADE project. In J. Véronis (Ed.), *Parallel text processing: Alignment and use of translation corpora* (pp. 369-388). Dordrecht: Kluwer Academic Publishers.

[17] Weaver, W. (1949). *Translation*. Mimeographed, 12 pp., July 15, 1949. Reprinted in Locke, William N. and Booth, A. Donald (1955) (Eds.), *Machine translation of languages*. John Wiley & Sons, New York, 15-23.

[18] Wittgenstein, L. (1953). *Philosophische Untersuchungen [Philosophical Investigations]*, translated by G.E.M. Anscombe, New York, Macmillan].