Computers and the Humanities **34:** 61–78, 2000.

Senseval/Romanseval: The Framework for Italian

NICOLETTA CALZOLARI and ORNELLA CORAZZARI
Istituto di Linguistica Computazionale (ILC) – CNR, Via della Faggiola 32, Pisa, Italy
(E-mail: {glottolo,corazzar}@ilc.pi.cnr.it)

Abstract. In this paper we present some observations concerning an experiment of (manual/automatic) semantic tagging of a small Italian corpus performed within the framework of the SENSEVAL/ROMANSEVAL initiative. The main goal of the initiative was to set up a framework for evaluation of Word Sense Disambiguation systems (WSDS) through the comparative analysis of their performance on the same type of data. In this experiment there are two aspects which are of relevance: first, the preparation of the reference annotated corpus, and, second, the evaluation of the systems against it. In both aspects we are mainly interested here in the analysis of the linguistic side which can lead to a better understanding of the problem of semantic annotation of a corpus, be it manual or automatic annotation. In particular, we will investigate, firstly, the reasons for disagreement between human annotators, secondly, some linguistically relevant aspects of the performance of the Italian WSDS and, finally, the lessons learned from the present experiment.

Key words: semantic tagging, word sense disambiguation, WSDS evaluation, inter-annotator agreement, Italian corpus annotation

1. Introduction

One of the most important aspects of the SENSEVAL/ROMANSEVAL initiative was the objective of setting up a comparative framework for evaluating WSDS in a **multilingual** environment, with two Romance languages – French and Italian – in addition to English. An innovative side was the selection of the corpus material for French and Italian and the definition of a common annotation methodology in order to allow cross-lingual comparison and evaluation of data and results.

The experiment on semantic tagging implied different phases:

1) selection of the material, i.e., a corpus and a reference dictionary;
2) selection of a list of lemmas and extraction of a subset of their corpus occurrences;
3) semantic tagging performed in different sites, consisting of the assignment of the dictionary reading numbers to the corpus occurrences;
4) comparison and evaluation of the results;
5) running of the WSDS;
6) evaluation and comparison of the WSDS' results;
7) evaluation of the experiment in view of future extensions.

A further step, consisting of a cross-lingual comparison of French and Italian, can be performed in cooperation between the University of Aix-en-Provence (Laboratoire Parole et Langage), Rank Xerox Research Centre of Grenoble and the Institute of Computational Linguistics (ILC) of Pisa.

In this introductory section we provide an overview of the selected text corpus, lemmas, dictionary and defined rules for manual annotation.

The selected corpus was a **parallel multilingual corpus** of approximately 1.1 million words per language, consisting of extracts from the *Journal of the European Commission, Written Questions* (1993).[1]

The dictionary selected was a medium-sized printed Italian dictionary of about 65,000 lemmas (Garzanti, 1995), with no hierarchical structure within entries and not corpus and frequency based. This choice was determined by the fact that presently no large coverage computational semantic lexicon exists for Italian, even though it is obviously of less interest in view of automatic tagging for Language Engineering (LE) applications. Moreover, a medium-sized dictionary was preferred to a more fine-grained and larger dictionary since an extended set of reading numbers – not necessarily and always well differentiated – would make not only the automatic WSD task too complex, but also the evaluation task much more difficult, since annotators would tend to disagree or to assign multiple tags, thus augmenting the disagreement rate.

As to the **selection of the words** to be tagged, it was based on three criteria: (i) their being translations of words chosen for French, in order to allow comparative evaluations of the results, (ii) their polysemy, (iii) the number of occurrences in the corpus (at least 50). Twenty nouns, 20 verbs and 18 adjectives were selected and their corpus occurrences were extracted. Of these, 40 words were translations of words selected for French. Not all translated lemmas were kept as some were not polysemous in Italian. The number of corpus occurrences to be tagged was 2701 (954 nouns, 857 verbs and 890 adjectives).

The semantic annotation was performed – for each word – by two human annotators. Three sites were involved in tagging (Pisa: ILC; Roma: University of Tor Vergata; Torino: CELI). The result is a list of occurrences with two reading numbers (assigned by the two annotators) taken from the definitions of the dictionary. A few conventional tags were defined to cover some particular cases, i.e.: (i) a question mark (?) when the meaning of the occurrence was missing in the paper dictionary, or more generally when semantic annotation was quite problematic, (ii) reading numbers separated by a slash when more than one dictionary meaning could be assigned to the same corpus occurrence, (iii) a star (*) to mark cases in which a different POS was wrongly selected among the occurrences of a given syntactic category.

The main issues on which we report in the following sections are: (i) the level of agreement between human annotators, (ii) the evaluation of the main reasons for disagreement focusing on the linguistic aspects, (iii) some general observations

concerning the performance of the Italian WSDS, (vi) lessons learned from the present experiment in view of future evaluation tasks.

2. Manual Annotation: Agreement vs. Disagreement Rate

A single reading number was assigned by the annotators 91% of the time; in a much smaller number of cases two or more reading numbers (4.8%) or a question mark (1.9%) was given.[2]

Therefore, in 6.7% (4.8% + 1.9%) of the cases, the paper dictionary turned out to be somehow not sufficiently representative of the language attested in the text corpus. The specificity of the corpus partially explains that, but this crucial point will be further examined with illustrative examples in the following section.

We mainly focus here on the comparison of the semantic tagging by the different human annotators. The level of agreement among annotators was computed according to two criteria:

- **full agreement**, when there is complete agreement on all senses proposed for a given wordform;
- **partial agreement**, when there is agreement on at least one of the senses proposed for a given wordform (this can be obtained e.g., between senses 1 and 1/2).

The following table displays the results in terms of partial vs. full agreement for each POS:

PoS	Occurr.	Part.Agr.	Full.Agr.
N	954	863 (90.4%)	814 (85.3%)
V	857	716 (83.5%)	681 (79.4%)
A	890	677 (76%)	552 (62%)
Tot.	2701	2256 (83.5%)	2047 (75.7%)

We can notice a rather broad convergence between annotators, probably due also to a dictionary with not too fine-grained distinctions. The highest level of agreement was reached on nouns, while the other two syntactic categories, especially adjectives, show more divergence. It is evident that by allowing the assignment of multiple tags to the same wordform, and accepting only partial agreement (e.g., between 1/2 and 2), the opportunities of agreement between annotators are sensibly increased. On the other hand, considering the results in terms of full agreement, the distance between nouns and verbs slightly decreases, but the distance between verbs and adjectives becomes much higher (adjectives seem more difficult to agree on).

If we take into account now the tags assigned to each occurrence by the two annotators, we obtain three types of possible combinations: (i) the two tags are

identical; (ii) the two tags are only partially equivalent (e.g., 1/2 and 2); (iii) the two tags are completely different. Almost all identical answers are single sense tags, while multiple tags are rarely exactly the same (only 6 cases). On the other hand, complete divergences are mainly due to different single reading numbers, but also to the fact that in a high number of cases at least one annotator judged a given word meaning missing from the dictionary.

Equiv.Tags	**N**	**V**	**A**
I (e.g., 1 and 1)	812 (85.1%)	661 (77.1%)	514 (57.7%)
II (* and *)		6 (0.7%)	33 (3.7%)
III (? and ?)		14 (1.6%)	1 (0.1%)
IV (e.g., 1/2 and 1/2)	2 (0.2%)		4 (0.4%)
Tot.	814 (85.3%)	681 (79.4%)	552 (62%)

Part.Equiv.Tags	**N**	**V**	**A**
I (e.g., 1/2 and 1)	49 (5.1%)	35 (4%)	117 (13.1%)
II (e.g., 1/2 and 1/5)			8 (0.8%)
Tot.	49 (5.1%)	35 (4%)	125 (14%)

Divergent.Tags	**N**	**V**	**A**
I (e.g., 1 and 2)	71 (7.4%)	92 (10.7%)	154 (17.3%)
II (e.g., 1 and ?)	17 (1.7%)	37 (4.3%)	22 (2.4%)
III (e.g., 1 and *)	1 (0.1%)	2 (0.2%)	17 (1.9%)
IV (e.g., 1 and 4/5)	2 (0.2%)	8 (0.9%)	17 (1.9%)
V (e.g., 1/2 and 4/5)			3 (0.3%)
VI (* and ?)		1 (0.1%)	
VII (e.g., 1/2 and ?)		1 (0.1%)	
Tot.	91 (9.5%)	141 (16.4%)	213 (23.9%)

Finally, it is worth noting that the agreement between annotators depends also on the individual words. Upon closer analysis, it turns out that two verbs which have two senses in the dictionary (*arrestare* (*to arrest; to stop*), *comprendere* (*to understand; to include*)) and three nouns (*agente* (*agent*) (3 senses), *compagnia* (*company*; *group*) (6 senses), *lancio* (*throwing*; *launching*) (3 senses)) were annotated in exactly the same way. In terms of partial agreement, also the verb *rendere* (*to render*; *to return*) (6 senses), the noun *corso* (*course; stream; current use; circulation*) (8 senses) and the adjective *stretto* (*narrow; tight; close*) (5 senses) were treated in the same way.

It is worth noting that there is an apparent absence of correlation between the polysemy of a lemma and the agreement vs. disagreement rate. Indeed, highly

polysemous words such as *passare* (*to pass*) (16 readings) and *corso* (8 readings) do not have the highest disagreement rate (16% and 1.9%), while lemmas such as *biologico* (*biological*) (3 readings) and *popolare* (*popular*) (4 readings) show a remarkable disagreement between annotators (73.6% and 75%). However, this is mainly due to the fact that only 4 senses of *passare* and 2 of *corso* are attested in the selected corpus. In fact, because of the specificity of the corpus at hand, only some senses of the most polysemous words are attested. For instance *compagnia* has six senses in the dictionary but only three of them occur in the corpus according to all annotators. For the same reason the verb *importare* occurs only with the meaning *to import* and never with the meaning *to matter*.

Indeed, the degree of attested and actual polysemy in the corpus seems more important than the more 'abstract' or potential degree of polysemy displayed in the dictionary.

3. Major Reasons for Disagreement between Annotators

In this section we discuss the most frequent and regular types of disagreement between annotators and illustrate their causes. We examined in detail the cases where the annotators had disagreed, and classified them according to the scheme below. Generally speaking, divergences of judgement seem to be due to all the elements involved in the experiment, namely, the dictionary (88.3%), the human annotators (7.9%), and the corpus (2.3%). The weight of the first element with respect to the other ones is striking.

We mainly focus here on the problems related to the dictionary and the corpus, which can be subclassified as shown in the table below.

Causes of Divergence	**N**	**V**	**A**	**Tot.**
Dictionary Problems				
Ambiguity of Dict. Read.	107 (76.4%)	103 (58.5%)	285 (84.3%)	495 (75.6%)
Missing Reading	11 (7.8%)	34 (19.3%)	15 (4.4%)	60 (9.1%)
Multiword Expression	4 (2.8%)	3 (1.7%)	17 (5%)	24 (3.6%)
Metaphorical usage		7 (3.9%)		7 (1%)
Corpus Problems				
Too short context	1 (0.7%)	4 (2.2%)	8 (2.3%)	13 (1.9%)
Type of text		3 (1.7%)		3 (0.4%)
Human Errors and Others	17 (12.1%)	22 (12.5%)	13 (3.8%)	52 (7.9%)
Tot.	140	176	338	654

The ambiguity of dictionary readings is the most important cause of divergence for all POS and especially for nouns and adjectives. On the other hand, many verbal occurrences were tagged differently because their sense in the corpus was considered missing from the dictionary by one annotator. The other reasons for divergence between annotators seem to be far less important. Nevertheless their

relevance has to be measured with respect to the type of selected corpus. For instance, multiword expressions (from now on MWEs) do not seem to be numerous in the text corpus under scrutiny.

3.1. AMBIGUITY OF DICTIONARY READINGS

By ambiguity of dictionary readings we mainly refer to three different problematic aspects of dictionary definitions that will be examined one by one in this section: vagueness, excessive granularity, inconsistency. In a high number of cases the disagreement between annotators about the interpretation, and therefore assignment, of two or more readings is due to the above mentioned problems. For instance, for the word *soluzione* (*solution*), in 31 cases out of 51, one annotator chose reading No. 2, the other, reading No. 3, thus showing the difficulties raised by the choice between the 'event' interpretation of reading No. 2 (to solve, to be solved) and the 'result' interpretation of reading No. 3 (solution, agreement). Another example is *alto* which, in 24 cases out of 51, receives reading No. 4 and 8 by different annotators. In this case the problem was to select between *alto* as *big, tall* of reading No. 4 and *important, elevated* of reading No. 8. There are many cases of such 'regular disagreement', and the most striking cases of this kind are listed below:

PoS	Lemma	Dic.Readings	Number of Disagr.	N.Occ.
N	soluzione	2 and 3	31	51
N	ordine	1 and 2	14	51
N	esercizio	1 and 3	14	51
N	diritto	3 and 5	11	51
V	mantenere	1 and 2	14	51
V	chiedere	1 and 2	11	51
V	rispondere	5 and 6	10	51
A	stretto	2 and 4	35	51
A	utile	1 and 2	24	51
A	alto	4 and 8	24	51
A	civile	1 and 2	19	51
A	particolare	1 and 2	13	51
A	biologico	2 and 3	11	38
A	sicuro	1 and 4	10	43

Let us examine the problems of dictionary interpretation more in detail by providing illustrative examples.

3.1.1. *Vagueness*

The borderline between slightly different meanings is not always clearly stated in dictionary definitions, and neither the examples nor the synonyms provided for

each meaning allow a better differentiation. For example, *mantenere* (*to maintain/ to keep*) – which means in the dictionary both *1. tenere, far durare in modo che non venga meno* (*i contatti*) (*to keep contacts*) and *2. tenere saldo, difendere* (*un primato*) (*to hold the supremacy/ a position*) – occurs, among others, in the following 'ambiguous' contexts (i.e., where both readings can apply):

- *le Nazioni Unite dispongono di forze armate proprie per* **mantenere** *la pace.* (*United Nations have their own army to maintain peace.*)
- *Potranno essi ad esempio* **mantenere** *la loro condizione di neutralità?* (*Will they be able to hold, for instance, their position of neutrality?*)
- *Mentre taluni donatori sono disposti a* **mantenere** *l'attuale livello dei loro stanziamenti di aiuto* (*While some donors are ready to maintain their level of financial help*)

In 14 cases, reading No. 1 was chosen by one annotator while the other one assigned reading No. 2 (ten cases) and 1/2 (four cases) for the same corpus occurrences.

The vagueness of some sense distinctions in the dictionary is definitely the most important cause of disagreement.

3.1.2. *Excessive granularity: need for under-specification*

In a number of occurrences, the sense in the corpus context is under-specified with respect to the distinctions in the dictionary, which are, by the way, good and necessary in other contexts. This is a consequence of the lexicographer's need to classify in disjoint classes what frequently appears – in the actual usage – as a 'continuum' resistent to clear-cut disjunctions. For instance, *conoscere* (*to know*) is defined both as *1. sapere, avere esperienza* (*to know, to have experience*) and as *2. avere notizia, cognizione di qualcosa* (*to be informed*). This distinction is in some ways too fine-grained and cannot be easily applied to all contexts. For example:

- *La Commissione* **conosce** *i gravi problemi che la siccità pone all'agricoltura portoghese.* (*The Commission is aware of the big problems that drought causes to the Portuguese agriculture.*)
- *La Commissione* **conosce** *perfettamente l'insoddisfacente situazione fiscale in cui si trovano le persone soggette all'imposta sul reddito.* (*The Commission is fully aware of the unsatisfactory fiscal situation of people who have to pay tax on their income.*)

In five cases one annotator chose reading No. 1 and the other reading No. 2, while in two cases the choice was respectively reading No. 2 and 1/2. For these contexts it would be necessary, in reality, to have a reading which is underspecified with respect to the source of the knowledge.

3.1.3. *Inconsistency*

The same linguistic phenomenon is sometimes treated in different ways in the dictionary. This lack of a coherent theoretical approach behind dictionary definitions forces the annotators to decide individually about the treatment of particular cases. In this sense dictionary inconsistency is indirectly responsible for the disagreement between different annotators.

An interesting example is provided by deverbal nouns which often have a 'process/event' and a 'result' interpretation. The dictionary is rather incoherent with respect to this property, since it provides this distinction for lexical items such as *acquisto* (*buying*), *produzione* (*production*), etc., but not, for instance, for *comunicazione* (*communication*), etc. which are defined only as event nominal. Indeed, the disambiguation of these two senses is perhaps translationally and syntactically irrelevant and quite problematic in most of the contexts, e.g., in the following:

– *In una* **comunicazione** *al Consiglio e al Parlamento europeo, del 30 aprile 1992 (1), la Commissione ha illustrato le sue riflessioni sulle future relazioni tra la Comunità europea e il Magreb. (In a communication to the Council and to the European Parliament on 30 April 1992, the Commission illustrated its observations about the future contacts between the European Community and Maghreb.)*

so that it seems unrealistic to expect both readings to be present and distinct in the dictionary but, at least, both word senses should be mentioned together in the definition. However some contexts clearly select one or the other meaning, as in the following example where the lemma *comunicazione* has only a 'result' interpretation:

– *La Commissione continuerà pertanto ad esaminare la* **comunicazione** *della Commissione del 18 gennaio 1990 dal titolo 'Un grande mercato interno dell'automobile'.* (*The Commission will continue therefore to examine the communication of the Commission on 18 January 1990, entitled 'A big internal car market'.*)

3.2. MISSING READINGS

A surprising regularity of treatment is found for occurrences which receive the question mark by one annotator (judging that the meaning is missing from the dictionary) and one reading number (which looks at a closer analysis as the most general sense) by the other. These cases reveal the presence of a real problem of interpretation of the context that the dictionary does not help to solve. For example, *coprire* (*to cover*) combined with the contexts: *settori* (*areas*), *zone rurali* (*rural areas*), *foreste lontane* (*far forests*), *i casi* (*cases*), *un divario* (*a gap*), *tutte*

le regioni (*all the departments*), *il fabbisogno* (*needs*), *le esigenze* (*requirements*) receives reading No. 4 by one annotator (*No. 4: proteggere, difendere dall'offensiva del nemico o dell'avversario: – la ritirata – nel linguaggio bancario e delle assicurazioni, garantire: – un rischio – le spese, recuperare le spese sostenute* (*to protect, to defend from the enemy's attack – in the banking, insurance domain: to guarantee risks, expenses, to get one's money back*) while the other considers these occurrences as senses missing from the dictionary.

Another example is *perseguire* (*to pursue*). In 7 corpus contexts, it has a juridical meaning which is not explicitly mentioned in the dictionary (1. *cercare di raggiungere, ottenere* (*un obiettivo*) (*to pursue an aim*); 2. *perseguitare* (*to prosecute*; *to indict*)) as in **perseguire** *i responsabili di gravi violazioni dei diritti internazionali/ perseguire le violazioni commesse dagli Stati membri . . .* (*to prosecute those who are responsible for violation of international rights*). One annotator judged this meaning as missing from the dictionary, while the other assigned reading No. 2 which seems to be the closest to the meaning of these corpus occurrences. Once more, the dictionary turns out to be unsatisfactory at least when confronted with this corpus.

3.3. MULTIWORDS AND METAPHORICAL USAGES

One of the problems of semantic tagging is the treatment of MWEs, even though their frequency depends on the type of selected text corpus and lemmas. For example, *breve* (*short*) in the corpus at hand occurs only in MWEs, as well as most of the occurrences of *capo* (*head*). Examples of MWE are *fare* **capo** *alla direzione generale* (*to link up with the administrative department*); *in* **ordine** *al prelievo parafiscale* (*as to the fiscal system*); **libero** *arbitrio*; *ribadire a* **chiare** *lettere* (*to say clearly*) etc.

The semantic tagging of the words in bold raises the following problem: how should we annotate MWEs, i.e., should they be annotated as (i) a set of single elements, or as (ii) non–compositional units? In the first case, which semantic tag (reading number) should be assigned to each element?

These questions are strictly related to the way traditional dictionaries provide and structure lexical information. Indeed, (a) only a restricted number of MWEs is provided, and (b) they are usually more or less arbitrarily assigned to one or another reading of the lemma. For instance, figurative expressions such as *aprire gli occhi a qualcuno* (*to make someone aware of something*), *aprire l'animo a qualcuno* (*to open one's heart*) are considered equivalent to *aprire una bottiglia* (*to open a bottle*) and included in the first reading of the verb *aprire* (*to open*) which is *dischiudere, disserrare* (*to disclose*). In this case, should the reading No. 1 be assigned to the verb as a single word?

The previous questions are also connected to the semantic and syntactic peculiarity of MWEs, i.e., to their 'non-compositionality' (Corazzari, 1992). Indeed, the semantic annotation of their single components does not allow us to access

all the semantic – and indirectly morpho-syntactic – properties of the sequences as a whole. For instance, if we consider the example *aprire la strada a qualcuno/qualcosa* (*lit. trans.: to open the road to someone/something*) we may say that:

- although this expression is structurally complex, it behaves semantically, as well as syntactically, like a single predicate;
- the global meaning of the MWE cannot be derived from the meaning of its components;
- the selectional restrictions as well as the argument structure of the verb *aprire* are not the same as those of the expression *aprire la strada*: the first one selects a Subject and an Object, while *aprire la strada* requires a Subject (either 'human' or 'non-human') and an obligatory Indirect Object.

Also, much simpler MWEs show the same properties. For instance **in ordine al** *problema economico* (*as far as the economical problem is concerned*) is a combination of two prepositions and a noun, but has a prepositional function as a whole. The non-compositionality of this MWE is particularly evident at the translational level where *in order to* (literal translation) has a totally different meaning.

We have just outlined some obvious and well-known reasons in favour of an annotation of MWEs as non-compositional units.

Another phenomenon – somehow connected to MWEs – is an important cause of disagreement between annotators, i.e., the metaphorical usage of a lemma. The borderline between MWEs and metaphorical expressions is sometimes quite fuzzy, even though the latter are potentially unlimited and unpredictable depending only on the writer/speaker's imagination. Indeed, only the most commonly used metaphorical usage of lemmas are included in the dictionary, under the label 'figurative meanings'.

A specific annotation strategy should be set up for handling coherently the metaphorical usage of lemmas, reminding us that they could never be exhaustively listed in the dictionary.

3.4. PROBLEMS RELATED TO THE CORPUS

The annotation problems related to the corpus concern, on one hand, the type of text and, on the other hand, the size of the context of the word occurrences. Dealing with a multilingual corpus and therefore – as far as Italian is concerned – with a translated corpus, we find wrong or unusual Italian expressions which cannot be easily classified according to the dictionary definitions. For instance *non* **aprono** *nessun diritto particolare* (*lit. trans.: they do not open any particular right*) does not seem a correct Italian expression: indeed *aprire* is used improperly and therefore it is quite difficult to choose among the different dictionary reading numbers.

Other cases which were differently coded by the annotators for the same reason are:

- **condurre** *una riflessione* (*lit. trans.: to do an observation*)
- **condurre** *una politica di parità* (*to do a politic of equality*)

As to the second problem, context size, which was established as the sequence of variable length included between two carriage returns, turned out to be insufficient in some rare cases.

4. Some Observations about the Performance of the WSDS

Two systems participated in the evaluation for Italian: from Pisa (ILC) and Rome (Eulogos). The quantitative evaluation of their results is given in Veronis (1998). We provide here only a few observations concerning linguistic aspects related to their performance.

4.1. POLYSEMY AND PERFORMANCE

Also for WSDS there is no clear correlation between degree of polysemy and performance of the systems, i.e., correctness of their results. For instance the adjectives *alto* (8 senses) and *biologico* (3 senses) are wrongly tagged (by one system) in 29 of the occurrences, *legale* (*legal*) (2 senses) in 41, and *libero* (*free*) (8 senses) in 27. The same is true for nouns and verbs: e.g., *centro* (*centre*) (8 senses) is wrong in 6 of the occurrences, while *concentrazione* (*concentration*) (2 senses) in 39; *rendere* (6 senses) is tagged completely correct by one system (syntactic clues were very relevant for this particular verb), and *passare* with 16 senses receives just one incorrect tag. We must observe, however, that most words are used in the chosen corpus in just very few senses: e.g., *libero* in 2 of the 8 senses, *centro* also in 2 out of the 8, etc. This may have a strong impact on performance and may be more relevant than dictionary polysemy.

It is worth noting that sometimes wrong tags were assigned by a system exactly where the human annotators were in disagreement. This happens more often than expected by chance, and signals clear cases of not enough or not good information either in the corpus context or in the dictionary. In a few cases we also observed that a system produced a disjunction of tags exactly in those cases where annotators gave a multiple tag. This is a strong sign of real ambiguity (or too great similarity) in the dictionary definitions.

4.2. DIFFERENCE IN PERFORMANCE BETWEEN THE SYSTEMS

The two Italian WSDS, even though similar in terms of a global quantitative evaluation (see Veronis, 1998), present very often quite different distributions of wrong

and correct tags, obviously due to the different techniques and approaches used. This is a sign of the **need for a qualitative analysis/evaluation** of the results accompanying the quantitative one, both for an interpretation of the reasons for success and failure, and for the evaluation task to be of real help in improving the system. We enumerate here some of the differences: (i) the use of multiple tags was much more frequent by one system, thus increasing the possibility of 'partial agreement' with the reference corpus, (ii) the '?' sign was much more frequently used by one system, to signal cases of inability to assign a tag, thus increasing precision, (iii) one system gave one and the same tag to **all** corpus occurrences for many words (8 verbs out of 20), thus hinting at the possible technique of choosing the most probable word sense (the disadvantage being that they may be all wrong, as happened with one word!).

4.3. USE OF MULTIPLE TAGS AND CASES OF DISAGREEMENT IN HUMAN ANNOTATION: THEIR EFFECTS ON THE EVALUATION

The use of multiple tags or – even worse – the cases of disagreement in the human annotated corpus largely increase the possibility of success for the WSDS calculated in terms of 'partial agreement'. Where human annotators disagree, the 'gold standard' includes all the tags that either of the annotators gave, so there is much more chance of a WSDS coinciding with at least one of the two (or more) tags. Therefore, the paradoxical situation arises that the most complex or difficult cases (where multiple tags are given or there is disagreement between annotators) are somehow the easiest for the systems if calculation of success is done in terms of partial agreement. This has to be weighted in the quantitative evaluation of WSDS.

4.4. SOME CONCLUSIONS WITH RESPECT TO THE ANALYSIS OF WSDs' RESULTS

The first important observation is that it is necessary to analyse qualitatively (not only quantitatively) the results, because the simple numbers can be misleading, e.g.:

- a specific text type may privilege one or two readings only, thus allowing an easier tuning of the system;
- a text with many recurrent MWEs may facilitate disambiguation.

It is therefore better to test systems with contexts taken from many different text types, so that a larger variety of readings is attested. We noticed in fact that actual polysemy in the text corpus is much more problematic than theoretical/potential polysemy in the dictionary.

In general, there is no correlation between multiple tags assigned by annotators and by the systems. However, the contexts with different tags given by different annotators present a quite different typology of cases, which must be carefully considered in order to better evaluate the quantitative results. The following cases require a different interpretation:

- at a better analysis, one tag is correct, the other is wrong: the 'partial agreement' evaluation with respect to one tag only (the incorrect one) may wrongly inflate success rate;
- a '?' tag, saying that a reading is missing, and a reading number are given: this is more difficult to match by the system than if two different reading numbers are given (one of the two may be more easily matched);
- the two tags are both applicable, because the context is actually ambiguous between the two and/or the dictionary readings are not differentiated enough (e.g., *chiedere* (*to ask*), between 1. in order to obtain and 2. in order to know, or *conoscere*, between 1. to experience and 2. to know): many contexts can express both senses at the same time (these are the cases for which an underspecified reading/tag would be useful).

This last type of disagreement, i.e., the cases of 'real' ambiguity, are common in the contexts examined. This is clear evidence of the gap existing very often between (i) a sort of 'theoretical language', used by linguists/lexicographers who have to classify the linguistic world in disjoint classes, and (ii) actual usage of the language, which is very often a 'continuum' resistant to clear-cut disjunctions, and needs to remain ambiguous with respect to imposed classifications. This is particularly true at the level of semantic analysis and annotation, where vagueness of language is a 'requirement' and not a 'problem' to be eliminated.

The problem is then how to individuate when this second type of 'only apparent disagreement' is present, thus pointing to a problem in the dictionary used: partial agreement by the system is here perfectly acceptable.

Again, figures must be carefully handled. Paradoxically, if there is disagreement between annotators or if there are multiple tags – as said above – it is much easier for a system to agree with at least one annotator: if both tags are possible (as for ambiguous contexts) there is no problem, but if only one tag is correct and the system agrees with the other tag, then a system may be evaluated highly while making mistakes. The same situation arises if it is the system which uses many multiple tags (at least one may more easily agree with an annotator). The conclusion in these cases is that 'the more difficult the easier' for a system. On the other side, to discard all cases of disagreement or multiple tags is obviously incorrect: they have different meanings in different situations – as said above. The conclusion is that more attention should be paid to the definition of the quantitative criteria for evaluation, to take care of these aspects.

5. Lessons Learned from the Present Experiment and Main Conclusions

Finally, we would like to draw some conclusions about the way the experiment was conducted in order to point out its limits and to contribute to improving future initiatives of this kind.

5.1. THE DICTIONARY: TOWARDS A COMPUTATIONAL LEXICON WITH SEMANTICS

The choice and interpretation of the dictionary turned out to be a critical issue. In particular, the printed dictionary proved to be not sufficiently representative of the language attested in the text corpus. In a next round, a computational lexicon could be used for Italian, e.g., the EuroWordNet (Alonge et al., 1998; Rodriguez et al., 1998) or SIMPLE (Ruimy et al., 1999) lexicons (with their extensions as provided in the Italian National Projects starting in '99). This will give more coherent and useful results from a LE viewpoint, with use of semantic types and hierarchical information enabling semantic generalisations.

In general, disagreement between annotators (and sometimes the use of multiple tags) is to be interpreted as a warning that there is something wrong in the dictionary used (or in its interpretation by the annotator, which frequently amounts to something not being clear in it). Some important requirements for a computational lexicon with semantics – as emerged from this analysis – are the following:

- need for under-specified readings in particular cases (maybe subsuming more granular distinctions, to be used only when disambiguation is feasible in a context): this implies paying careful attention to the phenomenon of regular under-specification/polysemy as occurring in texts;
- need for different readings to be well-differentiated, otherwise annotators and systems tend to disagree or to give multiple tags, thus inappropriately augmenting the chances of success in the evaluation;
- need for good dictionary coverage with respect to attested readings (to avoid the gap between current dictionaries' 'theoretical' language and 'actual' language as used in text corpora), possibly with indication of domain/text type differences;
- need for encoding/listing MWEs;
- need for encoding metaphorical usage.

A detailed analysis of representation and encoding of the last two aspects has to be done. It is worth noting that from a practical point of view a better encoding of MWEs could simplify automatic annotation, since they could be provided as a mere list to WSDS.

Crucial questions for a semantic computational lexicon are the following:

- Should/could a dictionary contain indication of clues for disambiguation associated with each reading (e.g., syntactic vs. semantic vs. lexical clues) when this is feasible?
- If so, could we profit from the task of manual semantic annotation of the so-called 'gold standard', and ask lexicographers to make explicit such clues where they can be individuated? It is well-known that this is not an easy task, because often different strategies – working at different levels of linguistic analysis – are at play in a disambiguation task. This is one of the aspects that makes semantic disambiguation such a difficult and challenging task.
- Do available dictionaries contain all that is needed for semantic classification/disambiguation? Or is there the need for other dimensions?

These are non-trivial aspects which deserve attention when planning and designing a computational lexicon.

5.2. THE NATURE OF 'MEANING'

Nevertheless, it is worth noting that one of the central questions is the nature of 'meaning' itself, that is rather a 'continuum' resistant to clear-cut distinctions – as the need for multiple tags for semantic annotation proves. Indeed, human intuition and sensibility still play a relevant role in word sense disambiguation, especially so when dictionary definitions are unsatisfactory and leave to the annotator the task of interpreting them. From this point of view it is interesting that, for instance, multiple tags are very rarely equivalent between annotators. Underspecification tries to partially tackle this aspect.

5.3. THE CORPUS: TOWARDS A SEMANTICALLY TAGGED CORPUS FOR LE

The phase of selection of the corpus material appears to be crucial for a correct performance of the experiment. In particular it seems advisable to select a 'balanced' reference corpus which reflects a variety of text types, genres and domains rather than a specific text corpus like the one that we chose for satisfying the multilingual requirement. (It is well known that only a narrow range of parallel corpora are available).

Indeed a specific text type may privilege only a subset of senses of a given lemma, thus simplifying the annotation task and increasing the chances of success, since a WSDS may be tuned ad hoc for choosing only among the most probable readings in that domain/text type/genre. At the same time, also a text with many recurrent MWEs may facilitate disambiguation, since WSDS can be provided with an ad hoc list. Conversely, a text corpus with a small number of MWEs provides a wrong view on the language, leading to the conclusion that MWEs are not an

important problem. Variety and representativity of (i) lemmas, (ii) MWEs, (iii) senses, and (iv) linguistic problems are only guaranteed by a well-balanced corpus, in the same way as correctness/reliability of the results is guaranteed by a well designed dictionary.

Again, in a next round a more balanced semantically tagged corpus produced within the Italian National Project will be used, similarly to what happened this time for English.

5.4. THE CHOICE OF THE LEMMAS

Considering now the selected lemmas, it is advisable to extract for each of them a reasonable amount of different word-forms for two main reasons: first, some specific senses are connected to a particular morpho-syntactic form, which implies that by excluding a certain word-form we exclude also some senses; secondly, a particular word-form can occur preferably in a given text type with only one sense providing a partial view on the different senses of the lemma and a wrong view on their frequency (this is why all the examined corpus occurrences of *breve* are the same, i.e., the MWE *in breve*). As we have already stressed, the context size of the occurrence is also relevant to a correct semantic annotation. It seems advisable to choose a more significant or extended window in order to allow better sense disambiguation, be it manual or automatic.

5.5. INTERACTION BETWEEN SEMANTICS AND SYNTAX

The aspect of the interaction between semantics and syntax is interesting from the perspective of automatic tagging, i.e., for WSDS. An analysis of the linguistic level at which to find the optimal clues for disambiguation (e.g., a particular subcategorised preposition, or a lexical collocation, or the co-occurrence with a specific subject, or even a particular morphological inflection, etc.) could lead to adding a very useful type of information to the different senses of an entry in a computational lexicon. The expensive phase of human semantic annotation, necessary to build a large and representative semantically tagged corpus, could aim also at getting this result, i.e., at individuating – when possible – the clues for disambiguation, for them to be encoded in a computational lexicon.

5.6. NEED FOR A COMMON ENCODING POLICY?

The present initiative was intended to prepare the ground for a future real task of semantic tagging/evaluation for LE applications. From this perspective one of the questions to be asked is the following:

– Can we define, and how, a ‘gold standard’ for evaluation (and training) of WSDS?

To answer this question in a way that is meaningful for LE applications implies not only an analysis of the state-of-the-art, and experiments like the present one, but also careful consideration of the needs of the community – also applicative/industrial requirements – before starting any large development initiative of corpus annotation which can fulfil NLP application requirements with respect to WSD. This aspect has not been really considered in the present initiative. The above question implies other questions:

– Can we agree on a common encoding policy? Is it feasible? Desirable? To what extent?

A few actions in this direction could be the following:

– to base semantic tagging on commonly accepted standards/guidelines (with implications for a future EAGLES initiative): up to which level this can be done is a matter of consideration;
– to involve the community and collect and analyse existing semantically tagged corpora used for different applications;
– before providing the necessary common platform of semantically tagged corpora, the different application requirements must be analysed;
– to build a core set of semantically tagged corpora, encoded in a harmonised way, for a number of languages.

A future EAGLES group could work on these tasks, building on and extending results of the current group on Lexicon Semantics (Sanfilippo A. et al., 1999), towards the objective of creating a large harmonised infrastructure for evaluation and training, as is so important in Europe where all the difficulties connected with the task of building language resources are multiplied by the multilingual factor.

Notes

[1] The corpus is part of the MLCC Corpus distributed by ELRA.
[2] The missing 2.3% concerns the semantic tags which are a star, not considered in the calculation because it is irrelevant to the present discussion.

References

Alonge, A., N. Calzolari, P. Vossen, L. Loksma, I. Castellon, M. A. Marti and W. Peters. “The Linguistic Design of the EuroWordNet Database”. Special issue on *EuroWordNet, Computers and the Humanities*, 32(2–3) (1998).

Busa, F., N. Calzolari, A. Lenci and J. Pustejovski. *Building a Lexicon: Structuring and Generating Concepts*. Computational Semantics Workshop, Tilburg, 1999.

Corazzari, O. *Phraseological Units*. NERC Working Paper, NERC-92-WP8-68, 1992.

Garzanti Editore. *Dizionario Garzanti di Italiano*. Garzanti Editore, Milano, 1995.

Rodriguez, H., S. Climent, P. Vossen, L. Loksma, W. Peters, A. Alonge, F. Bertagna, A. Roventini. "The Top-Down Strategy for building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology". Special Issue on *EuroWordNet, Computers and the Humanities*, 32(2–3) (1998).

Ruimy N. et al. *SIMPLE – Lexicon Documentation for Italian*. D.03.n.1, Pisa, 1999.

Sanfilippo A. et al. *Preliminary Recommendations on Semantic Encoding*. EAGLES LE3-4244, 1999.

Veronis J. *Presentation of SENSEVAL*. Workshop Proceedings, Herstmonceux, 1998.