Natural Language Engineering **8** (4): 375–390. © 2002 Cambridge University Press
DOI: 10.1017/S1351324902003017 Printed in the United Kingdom

Evaluating lexical resources using SENSEVAL*

NICOLETTA CALZOLARI, CLAUDIA SORIA
and FRANCESCA BERTAGNA
Istituto di Linguistica Computazionale-CNR, Via Moruzzi 1, 56100 Pisa, Italy
e-mail: {glottolo, claudia.soria, francesca.bertagna}@ilc.cnr.it

FRANCESCO BARSOTTI
Università di Pisa, Via S. Maria 36, 56100 Pisa, Italy
e-mail: francesco.barsotti@ilc.cnr.it

(*Received 12 November 2001; revised 13 July 2002*)

Abstract

The aim of our paper is twofold: to introduce some general reflections on the task of lexical semantic annotation and the adequacy of existing lexical-semantic reference resources, while giving an overall description of the Italian lexical sample task for the SENSEVAL-2 experiment. We suggest how the SENSEVAL exercise (and comparison between the two editions of the experiment) can be employed to evaluate the lexical reference resources used for annotation. We conclude with a few general remarks on the gap between the lexicon, a partially decontextualised object, and the corpus, where context plays a significant role.

1 Introduction

Starting from the assumption that no well established methods are yet available for evaluating lexical resources, the main goal of this paper is to explore the extent to which the SENSEVAL task can be exploited for this purpose. A comparison is made between the Italian tasks in SENSEVAL-1 and SENSEVAL-2, focusing in particular on an evaluation of the impact of the different lexical resources used in the two editions of the competition. SENSEVAL-2 was organized following the very same principles as SENSEVAL-1. The main difference between the two editions is that in SENSEVAL-2 the majority of the tasks used computational lexicons and in particular three of them used a WordNet-like database, while in SENSEVAL-1 the reference lexicons were based on traditional dictionaries. We focus in particular on an analysis of the differences for the two editions of the Italian task between the use of a computational lexicon and a traditional paper dictionary. SENSEVAL hence provides us with an ideal observational scenario to evaluate the use of a WordNet-like database as reference resource for manual and automatic sense annotation.

*We would like to thank Adam Kilgarriff for all his help, two anonymous referees for their comments and also Paolo Allegrini and Roldano Cattoni for their assistance.

In the first section we describe the preparation of the Italian task, explaining the methods and modalities used to set up the reference annotated corpus and the sense inventory. Then, we tackle the issue of manual sense annotation of a corpus, comparing, on the one hand, annotations performed by different human annotators and, on the other hand, the agreement/disagreement rate of the first and second editions of the experiment. We also try to analyze some results of the automatic annotation performed by the systems in SENSEVAL-2. In the last section we introduce some observations about what we have learnt from these experiences and what we think should be the focus of future research.

2 Preparing the corpus and the sense inventory for SENSEVAL-2

The corpus and the sense inventory used for the SENSEVAL-2 Italian lexical sample task were provided by two resources developed in the framework of the SI-TAL project.[1] The data were not adapted in order to be used for the competition, apart from the necessary format conversions. A common encoding format (XML) was exploited to facilitate reuse and sharing of the data.

2.1 Corpus preparation

The Italian lexical sample corpus consisted of about 3900 instances for 83 lexical entries (46 nouns, 21 verbs, and 16 adjectives), with an average of 47 contexts per entry.

The lexical samples were taken from the SI-TAL Italian Syntactic-Semantic Treebank (ISST[2]), which was in its completion phase when the SENSEVAL task was organized. For each instance, the context corresponded to the sentence containing the target word.

The ISST consists of two subcomponents: a generic and a domain-specific (financial) corpus, of about 215,000 and 90,000 tokens, respectively. The annotated material comprises instances of newspaper articles, representing everyday journalistic Italian language. As far as annotation is concerned, the ISST has a three-level structure: two levels of syntactic annotation (a constituency-based and a functional-based annotation level) and a lexical-semantic level of annotation. ISST is supposed to be used in different types of applications, ranging from training of grammars and sense disambiguation systems, to the evaluation of language technology systems.

Even if a system could be conceived in such a way as to make use of syntactic information, only the semantic annotation of the ISST was considered for its use in the SENSEVAL-2 task.

[1] SI-TAL ('Integrated System for the Automatic Treatment of Language') is a National Project, coordinated by Antonio Zampolli at the 'Consorzio Pisa Ricerche', involving several research and industrial centers in Italy, aiming at developing large linguistic resources and software tools for the Italian written and spoken language processing.

[2] See Montemagni *et al*. (2000a, 2000b). ISST is a joint effort among the Consorzio Pisa Ricerche (Pisa, Italy), Certia (Rome, Italy), Consorzio Venezia Ricerche (Venice, Italy) and IRST (Istituto per la Ricerca Scientifica e Tecnologica), Trento, Italy.

In ISST, sense annotation was performed manually using the ItalWordNet (IWN) lexicon as a reference resource (see section 2.2).

Semantic annotation was performed assigning a given sense number to each full word or sequence of words corresponding to a single unit of sense (such as compounds, idioms, etc.). Sense numbers, referring to specific synsets, were taken from IWN. Specific features not belonging to IWN were created for the annotation task to account for idioms, compounds and multiwords, figurative uses, evaluative suffixation, foreign words, proper nouns and titles, etc. However, in order to comply with the SENSEVAL-2 lexical sample format, the only semantic information used was the sense number of ISST, corresponding to the sense number of IWN synset variants, while the supplementary features had to be discarded.[3] Although the original ISST contained multiwords expressions, none of these were included in the SENSEVAL lexical sample.

The lemmas included in the SENSEVAL-2 corpus were selected on the basis of the following criteria:

- polysemy in the reference lexicon;
- polysemy attested in the corpus;
- frequency.

Average polysemy was of five senses per word (five for the nouns subset, six for the verbs and three for the adjectives). The average frequency turned out to be quite low, since the Italian Treebank from which the lexical sample was extracted was still incomplete, and we had to select the most frequent words with at least two senses in the lexicon and used at least in two of their senses in the annotated corpus. This led to choosing mainly words with a medium-high level of complexity, often with quite a high polysemy and rather generic senses. For instance, only 12 of the 46 nouns also had a concrete sense.

More importantly, since we had a rather low number of occurrences, no training data were provided for the Italian task. This makes the results for the Italian task hardly to compare with those which used similarly structured data, such as the Spanish, Swedish, Basque and Korean tasks, which all had training data available.

2.2 Lexicon preparation

As stated before, the occurrences provided for the WSD lexical sample task were annotated according to the lexical-semantic database ItalWordNet (see Roventini, Alonge, Calzolari, Magnini and Bertagna 2000), developed within the framework of the SI-TAL Project.[4]

ItalWordNet is an extension of the Italian WordNet built during the EuroWordNet project (Vossen 1999). The ItalWordNet database consists of:

[3] This fact obviously resulted in a loss of the overall semantic information available. For instance, the semantic annotation gave no information about the specific domain or about possible metaphoric senses: this will have to be changed in future SENSEVALS.

[4] ItalWordNet is a joint effort between Consorzio Pisa Ricerche (Pisa, Italy) and IRST (Istituto per la Ricerca Scientifica e Tecnologica), Trento, Italy.

1. a generic wordnet containing about 64,000 word senses corresponding to about 49,000 synsets;
2. a (generic) Interlingual-Index (ILI), which is an unstructured version of WordNet 1.5, also used in EuroWordNet (EWN) to link wordnets of different languages;
3. a terminological wordnet, containing about 5000 synsets from the economic-financial domain;
4. a terminological ILI, to which the terminological wordnet is linked;
5. the Top Ontology, a hierarchy of language-independent concepts, built within EWN and partially modified in IWN to account for adjectives (Alonge, Bertagna, Calzolari, Roventini and Zampolli 2000). Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the Top Ontology;
6. the Domain Ontology, containing a set of domain labels. Via the ILIs, all the concepts in the generic and specific wordnets are directly or indirectly linked to the Domain Ontology.

For the 83 lexical entries we provided the competitors with a hierarchical basic data structure: all the senses of the lemma organized in groups of synonyms (synsets), as well as their direct hyperonyms and a brief Italian definition. We also provided a set of semantic relations belonging to the set of Euro(/Ital)WordNet relations, such as hyponymy, role/involved, holo/meronymy, derivational relations etc. We did not provide the target entries of those relations (and all their semantic and ontological information), since only a portion of the whole wordnet[5] was made available. All entries were supplied with equivalence relations to at least one record of the EuroWordNet Interlingual Index and with the link to the EuroWordNet Top Concepts.

The entries have been used as they were in IWN, without making any specific adjustment for the SENSEVAL task. Although domain information, so useful in a WSD task, is included in the model (even if with only few labels), it was not available for any of the entries, as it had not been systematically codified in IWN and also because the majority of the entries were quite generic.

We are now evaluating whether a link between ItalWordNet and the semantic Italian SIMPLE[6] lexicon would be feasible, allowing ItalWordNet to inherit, among others, the rich domain information available in the SIMPLE database.

We did not consider POS-tagging as part of the task and we provided as corpus instances only those with the same POS as the previously selected lexical items, i.e. we eliminated occurrences of homographs belonging to different parts of speech.

3 The SENSEVAL-2 Italian annotation task

In the remaining part of this paper we overview the results of manual and automatic annotation in SENSEVAL-2, trying also to perform, wherever possible,

[5] The whole of the new version of IWN can be obtained through ELRA.
[6] See Lenci *et al*. (2000).

a comparison between the SENSEVAL-1 and the SENSEVAL-2 experiences. The main difference between SENSEVAL-1 and SENSEVAL-2, as far as organization of the Italian task is concerned, is that two different types of lexical resources were used in the two editions of the competition: in SENSEVAL-1, the Italian corpus was annotated according to a traditional printed dictionary (for a detailed description of the Italian SENSEVAL-1 task and results, see Calzolari and Corazzari 2000); while in SENSEVAL-2, the ItalWordNet computational lexicon was used.

We thus have the opportunity to compare the SENSEVAL-1 and SENSEVAL-2 annotation experiences from the point of view of the impact of a different type of lexical resource on the annotation task. The recommendation that a computational lexicon be used was exactly one of the outcomes of the SENSEVAL-1 evaluation.

It will not be a comparison 'given the same conditions', since in the two editions two completely different sets of lemmas were considered, with no overlap. Thus, the results of the comparison cannot be seen as a formal evaluation, which would have been possible only having at our disposal the same set of words, the same human annotators and also the same systems running on the data in the two editions. Nevertheless, it will be a chance to verify whether the same types of problem arise when we change one of the most important factors, i.e. the lexical resource providing the sense inventory.

3.1 Manual annotation

Manual annotation of the corpus was performed independently by two different annotators: the first annotated the data during the phase of ISST building, while the second performed the annotation when the SENSEVAL-2 subset was extracted. Annotators used two different tools in parallel: a tool for browsing the data in the semantic net[7] and a tool especially tailored to semantic annotation.[8] This latter tool provides a framework to display the corpus sentences containing a given lemma and to annotate the occurrences on the basis of the IWN sense inventory available for that lemma (displayed in a separate window).

The two versions of manual annotation provide the basis for calculating agreement rates and highlighting problematic cases.

Table 1 displays the results in terms of full agreement for each part of speech,[9] showing an overall high agreement between human annotators, with a decrease from nouns to adjectives. This pattern is consistent with the SENSEVAL-1 results of (Calzolari and Corazzari 2000) and those of (Fellbaum, Palmer, Dang, Delfs and Wolf 2001).

SENSEVAL-2 results, however, display a considerable *quantitative* overall improvement in terms of raw agreement rate, as illustrated in Table 2.

[7] The ItalWordNet tool was designed and implemented at IRST, Trento, Italy.
[8] The ISST tool (SemTAS) tool was designed and implemented by Certia, Roma, Italy.
[9] Annotators agreement is here expressed as raw percentages in order to ensure better comparability with the results of SENSEVAL-1, for which no statistic such as Kappa was used. Later in this paper we refine the analysis of the results by introducing kappa values.

Table 1. *Full agreement rate for each PoS*

	Nouns	Verbs	Adjectives	Total
Occurrences	2222	889	773	3884
Full agreement	2102	802	675	3579
% Agreement	94.6	90.2	87.3	92.1

Table 2. *Agreement rate in SENSEVAL-1 and SENSEVAL-2*

	Senseval-1	Senseval-2
Nouns	85.3	94.6
Verbs	79.4	90.2
Adjectives	62	87.3

This improvement can be explained both in terms of better coverage of the IWN resource than the traditional printed dictionary used for Senseval-1, and of the particular method adopted for the creation of the IWN resource.

Unlike a traditional lexicographic resource, IWN has been built partially taking into consideration the annotators' needs and feedback. This is due to the fact that the IWN database was still under construction while the ISST was being built. A protocol regulating the interaction between the IWN coders and the ISST annotators was established, and it has been possible to create or to adjust a sense (or a lemma) when needed by the annotation task. Therefore, while a traditional lexicographic resource is usually created independently of an annotation task, the construction phase of IWN benefited from the information derived from the annotation of the ISST.

Under the *qualitative* point of view, some considerations have to be made. In their analysis of Senseval-1 results, (Calzolari and Corazzari 2000) observed that no apparent correlation could be established between agreement rate and polysemy of the lemmas. However, they do find a relationship between annotators' agreement and actual polysemy, concluding that actual polysemy, namely polysemy attested in the corpus, seems more important than the potential degree of polysemy attested in the reference lexicon.

We thus performed a similar analysis of Senseval-2 manual annotation: we considered, for each part of speech, those lemmas that proved to be more problematic for annotators, and related them to their polysemy.

The Kappa statistic (Siegel and Castellan 1988; Carletta 1996) was used for this analysis, using as the basis for the random score the attested polisemy. Tables 3–5 show the most problematic lemmas for each part of speech.

The results highlight no apparent correlation between the polysemy of a lemma and annotators' agreement. Indeed, highly polysemous words such as the verb *coprire* (*to cover*), or the noun *mondo* (*world*), with 14 and 9 senses respectively, scored a

Table 3. *Annotators agreement for verbs*

Lemma	Overall senses	Attested senses	Kappa
comprendere – *to understand*	3	2	0.435
colpire – *to hit*	4	4	0.462
scoprire – *to discover*	8	6	0.56
lasciare – *to leave*	9	6	0.629
capire – *to understand*	5	3	0.722
coprire – *to cover*	14	7	0.736
entrare – *to enter*	5	5	0.833
trovare – *to find*	8	7	0.87

Table 4. *Annotators agreement for nouns*

Lemma	Overall senses	Attested senses	Kappa
rischio – *risk*	2	2	0.47
ora – *hour/time*	3	3	0.517
posto – *place*	3	3	0.579
senso – *sense*	6	4	0.633
controllo – *control*	6	5	0.661
forza – *strength*	3	2	0.685
colpo – *blow*	7	5	0.749
mondo – *world*	9	3	0.798
opera – *work*	8	5	0.805
politica – *politics*	4	3	0.808
lavoro – *job*	7	6	0.832

Table 5. *Annotators agreement for adjectives*

Lemma	Overall sense	Attested sense	Kappa
solo – *alone*	3	3	0.558
possibile – *possible*	2	2	0.626
nuovo – *new*	3	3	0.645
pronto – *ready*	4	4	0.649
lungo – *long*	3	2	0.688
piccolo – *small*	6	4	0.734
vero – *true*	3	3	0.799
grande – *big*	6	5	0.828
generale – *general*	3	2	0.878

high kappa value (0.736 and 0.798). On the other hand, the lemmas that appeared more difficult to disambiguate were in fact among those with few senses, such as the verb *comprendere* (*to comprise/understand*, three senses, k = 0.435), the noun *rischio* (*risk/danger*, two senses, k = 0.47), or the adjective *solo* (*alone*, 3 senses, k = 0.558).

Bearing in mind that the available data (a few dozens of lemmas) only allow tentative conclusions, it seems that the reason for low agreement between annotators must be found elsewhere than in the polysemy of the lemmas. (Calzolari and Corazzari 2000) argued that disagreement between annotators was mainly due to some intrinsic features of the dictionary, and in particular to the ambiguity of the dictionary reading interpretation, especially vagueness and excessive granularity of sense distinctions. A closer analysis of the SENSEVAL-2 manually annotated data allows us to conclude that the same range of dictionary problems explains the lack of agreement between annotators. In section 4 we give a response to these findings and also consider some observations coming from the analysis of the results of the automatic annotation, discussed in the next section.

3.2 Automatic annotation

In SENSEVAL-2, the system developed at ITC-IRST, Italy and the one developed at Johns Hopkins University, USA, participated in the evaluation for the Italian task; the quantitative evaluation of their performance is given in the SENSEVAL-2 proceedings (SENSEVAL-2, forthcoming).

Although our aim here is to make observations concerning linguistic aspects of their performance, a few technical details are required. The participating systems must assign an answer (i.e. a sense number) to each occurrence of a given lemma in the corpus. Three scoring policies are adopted in SENSEVAL-2: fine-grained scoring implies a one-to-one mapping between the gold-standard tags and the guess; coarse-grained scoring presupposes the availability of a sense subsumption table: the answers of both systems answers and gold-standard tags are mapped onto coarse-grained senses and then compared; finally, if a sense subsumption hierarchy is available, then the mixed-grained scoring gives some credit to choosing a more coarse-grained sense than the gold standard tag, but not full credit.

The results for fine, mixed and coarse-grained WSD tasks are illustrated in Tables 6–8. For some entries, low performance of the systems is related to subtlety in sense distinctions, since better results were obtained with coarse-grained scoring.

For the analysis of the data we used the same method employed for the analysis of manual annotation: analysing the results for fine-grained scoring, we considered a

Table 6. *Fine-grained scoring*

System	Precision	Recall	Attempted
IRST-ita-sample	0.406	0.389	95.783%
JHU_Italian	0.353	0.353	100%

Table 7. *Mixed-grained scoring*

System	Precision	Recall	Attempted
irst-ita-sample	0.482	0.461	95.783%
JHU_Italian	0.421	0.421	100%

Table 8. *Coarse-grained scoring*

System	Precision	Recall	Attempted
irst-ita-sample	0.483	0.463	95.783%
JHU_Italian	0.423	0.423	100%

failure of the system as an instance of disagreement, and then we verified whether any correlation between the most difficult lemmas and polysemy could be established. As for manual annotation, both potential polysemy and actual polysemy were considered.

For automatic annotation, our analysis confirms the tendency we found for manual annotation in the distribution of agreement over the lemmas. There was no correlation between the degree of polysemy (both actual and potential) and agreement rate. A highly polysemous word such as the verb *vedere* (*to see*, 7 senses attested out of 17) or the noun *lavoro* (*work*, 6 senses attested out of 7), had a high agreement rate (53% and 48%, respectively). *Anno* and *fine* (*year* and *end*, both three senses attested out of four lexicon senses) scored a 2% and 9% agreement rate, respectively[10]. The analysis of automatic annotation confirms the pattern found for manual annotation, namely that a failure in sense disambiguation is not be due to the number of senses of a lemma.

The observations overlap with those advanced by Calzolari and Corazzari (2000) for SENSEVAL-1 but a more precise comparison is not possible, since the systems that ran on the data of the two editions of SENSEVAL were different.

In the discussion of manual annotation, we indicated how the reason for annotators' disagreement should be found elsewhere than in the number of senses, a more significant factor being the way in which the different senses are distinguished.

In SENSEVAL-2, coarse-grained scoring gave us the opportunity to evaluate the impact of the complexity of sense distinction on sense disambiguation. We thus isolated some lemmas whose senses appeared to be poorly distinguishable on the basis of their definitions in the lexicon, and the impact of sense clustering on WSD systems was evaluated (see section 3.2.2). The results for some of the verbs and nouns are shown in Figure 1.

[10] These are the worst results and cannot be considered an everage of the two systems' performance (see Tables 6–8).

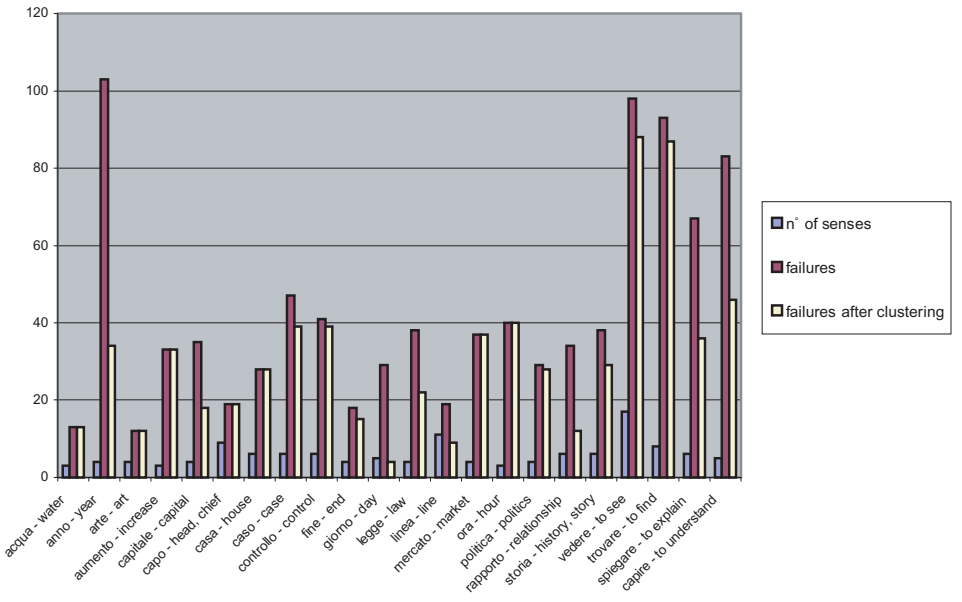

Fig. 1. Comparison between performance on normal and clustered senses.

4 General remarks about the experiment

A few useful observations can be made from the comparative analysis of manual and automatic annotation in SENSEVAL-1 and SENSEVAL-2.

The evaluation of specific cases highlights the persistence of the same typologies of problems encountered using a traditional printed dictionary, although to a lesser extent.

4.1 Sense distinction

As we have noted before, the main reasons for the disagreement between annotators are vagueness, ambiguity and excessive granularity of sense distinctions. These aspects of sense distinctions are spotted when one is trying to use them in a corpus annotation task. Some senses that would seem easily distinguishable and perfectly legitimate turn out to be hardly applicable in most corpus occurrences.

A typical example is given by the adjective *possibile* (*possible*):

possibile 1 – che può esistere, che può essere fatto (*that can exist, that can be done*)

possibile 2, probabile, verosimile – detto di ciò che è verosimile e può pertanto realizzarsi; che è simile al vero e come tale può essere creduto (*said about something that is possible and therefore can take place; that is similar to the truth and can be believed as such*)

This distinction follows the one present in the Garzanti Italian dictionary. The first sense refers to a more 'practical' nuance of possibility and is linked to the {realizzabile, attuabile, effettuabile, praticabile, eseguibile} (*feasible*, etc..) synset. The

second sense refers to a more 'philosophical' nuance of the concept of possibility in terms of likelihood.

The adjective *possibile* turned out to be one of the hardest to disambiguate, i.e. one with the lowest agreement rate. In what follows we exemplify some contexts of occurrence of the word:

Dall'incrocio dei dati in possesso delle Camere di commercio con quelli in possesso di grandi archivi telematici, come Inps, Enel, Inail, sarà infatti ***possibile*** *in futuro ottenere il controllo in tempo reale dell'intero sistema delle imprese italiane che attualmente conta quasi 4 milioni di posizioni.* (sense 1)
(Combining data owned by Chambers of Commerce with data owned by telematic archives, such as Inps, Enel, Inail, it will be **possible** in future to keep control of the whole system of Italian firms which actually adds up to 4 million positions.)

Forse non è il ***possibile*** *arrivo di Schumacher a demoralizzare i piloti ma la china discendente che la Ferrari sembra aver imboccato.* (sense 2)
(Maybe it's not Schumacher's **possible** arrival that depresses drivers but Ferrari's descending slope.)

E' necessario infatti stimolare tutti i ***possibili*** *recuperi di produttività;* (sense 1 or 2?)
(It is necessary to stimulate all **possible** recovery of productivity;)

Even if perfectly plausible, this distinction is probably too subtle and difficult to be used in corpus annotation and the disagreement highlights a 'grey zone' in which the two senses overlap.

Another interesting case is provided by instances of regular polysemy found for example in the very common distinction – typical of lexicographic practice – between *act* and *effect*.

The following are senses 3 and 4 of the *controllo* (*control*) lemma:

controllo 3 – atto del controllare e dirigere la correttezza di qualche evento (*the act of checking and verifying the correctness of an event*) (Top Concepts: Dynamic, Cause)

controllo 4 – stato che esiste quando una persona o un gruppo di persone ha potere sopra un altra persona (*situation that exists when a person or a group of persons have influence over another person*) (Top Concepts: Property, Static)

This distinction is very hard to make in the practice of corpus annotation. The examples below illustrate two occurrences of the 'controllo' lemma where the two senses could well be interchangeable.

Misure per il ***controllo*** *della spesa.* (sense 3)
(Measures for the expense **control**.)

Che dovrebbe fare, allora, Berlusconi? Se il disegno di legge approvato dal Senato sarà ratificato senza sostanziali variazioni dalla Camera, dalla prossima legislatura Berlusconi dovrà scegliere davvero se fare l'uomo di governo o mantenere il ***controllo*** *delle sue tv.* (sense 4)
(What should Berlusconi do? If the bill passed by the Senate is approved without substantial modifications, by the Chamber, starting from the next legislation Berlusconi will really have to decide whether to be an administration man or to maintain control of his tv.)

This type of regular polysemy, although quite normal in lexicographic practice, is in many cases very problematic to apply. Another example of too fine-grained

distinctions is the case of 'aumento' (*increase*). The first sense of 'aumento' refers to the 'change' event:

aumento 1, rialzo 3, crescita 1, incremento 1 – il crescere; accrescimento (*increase, growth*) – (Top Concepts: Quantity, Cause)

Sense number 2 refers to the quantity that contributes to the change:

aumento 2, ingrandimento 4, crescita 2, accrescimento 2, incremento 2 – ciò che accresce la quantità iniziale (*increase, rise*) – (Top Concepts: Part)

In what follows we can see how these senses are used in the annotation task.
Use of 'aumento 1':

In altri termini, per esempio, nel caso di ***aumento*** *dei debiti verso fornitori aggiungendo all' utile netto tale aumento si rettificano gli acquisti per riflettere solo quelli pagati nell' esercizio.*
(In other words, for instance, in the case of the increase of debts to suppliers adding to the net profit it is possible to modify the purchases to reflect only those paid in the exercise.)

Use of 'aumento 2':

Le aziende che lavorano su commessa – si legge nel comunicato CsC-hanno registrato nell' acquisizione dei nuovi ordini un ***aumento*** *su base annua del 5,8%; in aprile era stato del 3,4% e in marzo del 3,1 per cento.*
(All companies working on order – as you can read in the CsC comunication – have recorded in the acquisition of new orders a yearly increase of 5,8%; in April it had been of 3,4% and in March of 3,1%.)

We think that the difficulty of these examples is due to the over-specificity of the senses. In many cases it is too difficult for a system, or even for a human tagger, to distinguish between the change act/event and the quantity that changes the dimension/size/amount of something, because the two aspects are strongly intertwined, since no 'change event' exists without a changing quantity. Nonetheless, sometimes this distinction can be recognized in the corpus, when we find occurrences like 'misurare l'aumento di qualcosa' (*measuring the increase of something*) and it is important to note that the same distinction for the word 'increase' is also present in WordNet1.6.

A possible solution to this kind of problem would consist in adding an under-specified sense to the list of available readings which covered both nuances of the word.

4.2 Sense clustering

Sense clustering is obviously related to an improvement of the performance of automatic annotation and the results show that a rethinking of the quality/types of sense distinction is needed.

The effects of clustering are especially clear in the case of the lemma *anno* (*year*), whose sense distinctions are as follows:

anno 1 – tempo necessario alla Terra per compiere il suo giro intorno al Sole (*the time employed by the Earth to turn around the Sun*)

anno 2 – periodo di dodici mesi in genere (*a generic period of twelve months*)

anno 3 – periodo di tempo non determinato, di cui si sottolinea la lunghezza (*an undetermined period of time, usually very long*)

anno 4, annata – periodo di tempo (esp. nell'agricoltura); arco di tempo in cui si svolge un ciclo di attività (anno accademico, anno liturgico) (*a period of time, e.g. in agriculture; the span of an activity cicle*)

After senses 1 and 2 had been merged, one of the systems improved its performance going from 103 to 34 wrong answers (cf. Figure 1).

The same sense distinction between senses 1 and 2 can be found in Italian printed dictionaries and also in WordNet 1.5. It refers to a distinction between an astronomical-scientific sense of *year* and a more general, everyday sense of 'time measure'.

This distinction was not problematic for the human annotators (who always used sense 2 with the exception of the occurrences of *anno solare*, 'solar year'), whereas it was problematic for the automatic systems, which only applied sense 1.

The case of the lemma 'anno' raises another issue concerning the possibility in the WordNet model to discriminate among the different senses: all the four senses of 'anno' are correctly related to the same hyperonym ({tempo, periodo}) (*time, period*), and the same Top Concepts (Time and Quantity).

It is difficult to see how an automatic system could distinguish between the different senses when given the same hierarchical information without resorting to other means for capturing the differences.

While human disambiguation can be performed on the basis of the mere definitions, computational resources are useful only to the extent to which they provide a way to encode multi-dimensional semantic information, not only limited to taxonomic information; the model should provide the highest expressiveness in terms of sense discriminating power.

5 Conclusions

On the basis of the SENSEVAL-2 experience, we would like to conclude with a few general remarks, both about the adequacy of available lexical-semantic reference resources for WSD tasks and about the overall task of lexical-semantic annotation.

During the last years, many researchers have noted that it is misleading to reproduce in lexical resources what Fillmore calls the 'checklist theories of meaning' (Fillmore 1975). Kilgarriff (1997) and Hanks (2000), quoting Sue Atkins' well-known sentence 'I don't believe in word senses', expressed their skepticism about the possibility to capture through sense enumeration the overlaps, vagueness, and interplay of the different uses of a word in a language. Yet, these uses are exactly what contribute to giving the language its extraordinary dynamism and expressive power.

The information available in the ItalWordNet lexicon (and the same probably holds for the state of the art of computational lexicons and printed dictionaries in general) fails to account for the contextual aspects tied to word usage.

Fellbaum *et al.* (2001) similarly argue that what they call the 'dictionary model of word representation' does not allow a satisfactory representation of linguistic behavior as far as meaning is concerned.

However, it is precisely this model that has heavily influenced the practical realization of our currently available lexical resources: 'WordNet's entries resemble those of a traditional dictionary, though its organization is not alphabetical but that of a semantic network' (Fellbaum *et al.* 2001: 4).

This raises the more general issue of the relationship between (i) a lexical resource where senses (as well as other lexical information at other levels of linguistic analysis) are by necessity somehow 'decontextualized' (necessary if one is to capture generalizations), and (ii) a corpus sense annotation task, where, on the contrary, contextualization plays a predominant role and raises a range of pragmatic issues. In addition to this, the use of WordNet or WordNet-like resources significantly correlates with some worsening in the performance of WSD systems compared with previous results obtained using traditional dictionaries, as the overall SENSEVAL-2 results show (see Edmonds and Kilgarriff, forthcoming). This calls for a careful qualitative evaluation of the cases of divergence between lexicon encoding and corpus annotation requirements. Such an evaluation could shed light on recurrent types of mismatches, such as the regular polysemy cases discussed above. These could be dealt with in the lexicon by means of virtual underspecified senses semi-automatically generated.

We think that an important challenge in our field would be the transformation of theories able to deal with the extreme flexibility and multidimensionality of meaning into real, large-scale and exploitable resources.

The 'dictionary model' of word meaning representation, with its enumerative sense distinctions, can be changed in the near future with a new paradigm of representation, the generative one (Pustejovsky, 1995), in which senses related by systematic polysemy can be generated using rules that capture regularities in sense creation.

One of the models we have kept in mind during the construction of resources like ItalWordNet and SIMPLE is the 'repository' one, the 'store' from which it is possible to draw different pieces of information useful for various and specific applications. The problem is that it is impossible to hold a view of word meaning as a 'piece of information' provided with an autonomous status independent of its use. The criteria for sense distinction seem to be very application-dependent.

For the time being, it seems useless to abstractly ask ourselves how many senses a lexical entry should have. It might be more useful to capture just the core, basic distinctions in a core lexicon, trying to orient a resource towards different kinds of LE applications, in order to meet the different requirements of the different tasks.

Machine translation certainly needs a very fine-grained representation of meanings in order to deal with the many idiosyncrasies of bilingual/multilingual transfer, while coarse-grained sense distinctions may be sufficient for information retrieval applications.

In the near future, we would like to investigate how to make ItalWordNet a more flexible resource, able to provide different sense clusterings for different uses.

Moreover, there seem to be areas of meaning that cannot be easily encoded at the lexical-semantic level of annotation: sense interpretation may require appeal to, for example, extra-linguistic (world) knowledge which cannot be encoded/captured at the lexical-semantic level of description. We refer here to metaphors extending to entire word sequences, and not limited to the single word; to words acquiring a specific sense, strictly dependent on the context, that cannot be encoded at the lexical-semantic level; or to the complexity and variety of nuances implied e.g. by a verb, according to the type of co-occurring direct object. Not all these shifts of meaning can be captured through lexical-semantic annotation and at the level of the lexical entry.

We should start rethinking the complex relationships between the lexicon and the corpus in order to design a new model of the lexicon which does not suffer from the limitations of currently available static computational lexicons. We should move towards a more flexible model of a lexicon, i.e. a dynamic lexicon which extends the expressiveness of the core static lexicon by adapting to the requirements of language in use as attested in corpora, without a proliferation of senses, and the recent work on generative lexicons can be interpreted as a first step in this direction.

References

Alonge, A., Bertagna, F., Calzolari, N., Roventini, A. and Zampolli, A. (2000) Encoding information on adjectives in a lexical-semantic net for computational applications. *Proceedings 1st NAACL Meeting*, pp. 42–49. Seattle, WA.

Bertagna, F., Soria, C. and Calzolari, N. (forthcoming) The Italian lexical sample task. *SENSEVAL-2: Proceedings of SENSEVAL-2, Second International Workshop on Evaluating WSD Systems*, Toulouse, France. Association for Computational Linguistics.

Calzolari, N. and Corazzari, O. (2000) Senseval/Romanseval: The framework for Italian. In: Kilgarriff, A. and Palmer, M. (eds.), Special Issue on Senseval. *Comput. and the Humanities* **34**(1–2): 61–84.

Carletta, J. (1996) Assessing agreement on classification tasks: The Kappa statistics. *Computational Linguistics* **23**(1): 13–32.

SENSEVAL-2: Proceedings of SENSEVAL-2, Second International Workshop on Evaluating WSD Systems, Toulouse, France. Association for Computational Linguistics.

Fellbaum, C., Palmer, M., Dang, H. T., Delfs, L. and Wolf, S. (2001) Manual and automatic semantic annotation with WordNet. *Proceedings NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Pittsburgh, USA.

Fillmore, C. J. (1975) An alternative to checklist theories of meaning. *1st Annual Meeting of the Berkeley Linguistic Society*, pp. 123–132.

Hanks, P. (2000) Do word meanings exist? In: Kilgarriff, A. and Palmer, M. (eds.), Special Issue on Senseval. *Comput. and the Humanities* **34**(1–2): 205–215.

Kilgarriff, A. (1997) I don't believe in word senses. *Comput. and the Humanities* **31**(2): 91–113.

Kilgarriff, A. and Palmer, M. (eds.) (2000) Special Issue on Senseval. *Comput. and the Humanities* **34**(1–2): 61–84.

Lenci, A., Bel, N., Busa, F., Calzolari, N., Gola, E., Monachini, M., Ogonowsky, A., Peters, I., Peters, W., Ruimy, N., Villegas, M. and Zampolli, A. (2000) SIMPLE: A general framework for the development of multilingual lexicons. *Int. J. Lexicography* **13**(4): 249–263.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F. M., Mana, N., Pianesi, F. and Delmonte, R. (2000) The Italian syntactic-semantic treebank:

architecture, annotation, tools and evaluation. *Proceedings Workshop on Linguistically Interpreted Corpora,* Luxembourg.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M. T., Saracino, D., Zanzotto, F. M., Mana, N., Pianesi, F. and Delmonte, R. (2000) Building the Italian syntactic-semantic treebank. In: Abeillé, A. (ed.), *Building and Using Syntactically Annotated Corpora.* Language and Speech Series, Dordrecht: Kluwer.

Pustejovsky, J. (1995) *The Generative Lexicon.* MIT Press, Cambridge, MA.

Roventini, A., Alonge, A., Calzolari, N., Magnini, B. and Bertagna, F. (2000) ItalWordNet: a large semantic database for Italian. *Proceedings 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Siegel, S. and Castellan, Jr., N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences* (2nd Ed.). New York: McGraw-Hill.

Vossen, P. (ed.) (1999) EuroWordNet General Document. http://www.hum.uva.nl/~ewn.