Polysemy and Sense Proximity in the Senseval-2 Test Suite.

Irina Chugur *irina@lsi.uned.es*
Julio Gonzalo *julio@lsi.uned.es*
Felisa Verdejo *felisa@lsi.uned.es*
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
Madrid, Spain

Abstract

We report on an empirical study of sense relations in the Senseval-2 test suite. We apply and extend the method described in (Resnik and Yarowsky, 1999), estimating proximity of sense pairs from the evidence collected from native-speaker translations of 508 contexts across 4 Indoeuropean languages representing 3 language families. A control set composed of 65 contexts has also been annotated in 12 languages (including 2 non-Indoeuropean languages) in order to estimate the correlation between parallel polysemy and language family distance. A new parameter, sense stability, is introduced to assess the homogeneity of each individual sense definition. Finally, we combine the sense proximity estimation with a classification of semantic relations between senses.

1 Introduction

Our goal is to characterize sense inventories, both qualitatively and quantitatively, so that the following questions can be answered:

- Given a pair of senses of the same word, are they related? If so, in what way and how closely?
- How well are individual senses defined? For each sense, how homogeneous are its examples of use? How coarse is its definition? Should it be split into subsenses?
- How do these issues affect the evaluation of automatic Word Sense Disambiguation (WSD) systems using the sense inventory? What penalty should be assigned to a WSD system that confuses two senses, i.e. how much should it be penalized according to how close these senses are? Can the sense inventory be improved for evaluation purposes, for instance, splitting senses into finer-grained distinctions or collapsing close senses into coarser clusters?

In particular, we are interested in characterizing WordNet 1.7 as sense inventory for the Senseval-2 WSD comparative evaluation. Unlike conventional dictionaries, WordNet does not group senses of the same word in a hierarchical structure; every sense belongs to a synset, and can only be related to other senses via conceptual relations (rather than sense relations). Conceptual relations can be used to define measures of semantic distance (Sussna, 1993; Agirre and Rigau, 1996; Resnik, 1995), but topic relatedness is not well captured by wordnet relations, and this is a fundamental parameter to estimate sense similarity in many NLP applications (Gonzalo et al., 2000).

The issue of estimating semantic distance between senses of a polysemous word has been previously addressed in (Resnik and Yarowsky, 1997; Resnik and Yarowsky, 1999). They propose a measure of semantic distance based on the likelihood of the sense distinction being lexicalized in some target language. The measure was tested using statistics collected from native-speaker translations of 222 polysemous contexts across 12 languages. The results obtained showed that monolingual sense dis-

Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 32-39. Association for Computational Linguistics.

tinctions at most levels of granularity can be effectively captured by translations into some set of second languages, especially as language family distance increases. The distance matrices obtained reproduced faithfully the hierarchical arrangement of senses provided by the Hector database used in Senseval-1 (Kilgarriff and Palmer, 2000).

In order to characterize the Senseval-2 Wordnet 1.7 subset, we have adopted such methodology, extending it to capture also individual sense homogeneity, and comparing both quantitative measures with a coarse, qualitative classification of semantic relations between senses of a word.

In Section 2 we introduce the quantitative measures of *sense relatedness* (as defined by Resnik & Yarowsky) and *sense stability* (as an extension to it). In Section 3, we describe the qualitative classification that will be confronted to such measures. In Sections 4 and 5, we describe the experiment design and discuss the results obtained. Finally, we draw some conclusions.

2 Estimating sense relatedness and sense stability

In order to characterize a sense repository and evaluate the quality and nature of its sense distinctions, two aspects of sense granularity should be addressed:

- Are there sense distinctions that are too close to be useful in WSD applications, or even to be clearly distinguished by humans? In general, what is the semantic distance between senses of a given word?
- Are there sense definitions that are too coarse-grained, vague or confusing? If so, should they be split into finer-grain senses?

Our goal is to give a quantitative characterization of both aspects for the Senseval-2 test suite. Such measures would enable a finer scoring of WSD systems, and would provide new criteria to compare this test suite with data in forthcoming Senseval evaluations.

The first question can be answered with a quantitative estimate of sense proximity. We will use the cross-linguistic measure of sense distance proposed in (Resnik and Yarowsky, 1999), where sense relatedness between two meanings of a given word, w_i and w_j, is estimated as the average probability of receiving the same translation across a set of instances and a set of languages:

$$P_L(\text{same lexicalization}|w_i, w_j) \equiv$$

$$\frac{1}{|w_i||w_j|} \sum_{\substack{\text{x} \in \{\text{w}_i \text{ examples}\} \\ \text{y} \in \{\text{w}_j \text{ examples}\}}} tr_L(x) = tr_L(y)$$

$$proximity(w_i, w_j) \equiv$$

$$\frac{1}{|languages|} \sum_{L \in \text{languages}} P_L(\text{same lexicalization}|w_i, w_j)$$

where $tr_L(x), tr_L(y)$ are the translations of instances x, y into language L.

The second question can be addressed by estimating sense *stability*. Extending the assumption in (Resnik and Yarowsky, 1999), we propose a sense stability score based also on cross-lingual evidence: stability will be estimated with the likelihood for a pair of occurrences of a word sense w_i of receiving the same translation for a language L, averaged for as many language (and language families) as possible:

$$stability(w_i) \equiv$$

$$\frac{1}{|languages||w_i|^2} \sum_{\substack{L \in \{\text{languages}\} \\ x, y \in \{\text{w}_i \text{ examples}\}}} tr_L(x), tr_L(y)$$

This value depends on various factors. Too coarse-grained sense definition should lead to a lower stability, since different contexts may highlight subsenses differently lexicalized across the selected languages. The stability index also reflects the adequacy of the selected instances and how well they have been understood by annotators. A sense with three instances translated into a target language always by the same word form, will receive the maximum stability score (1). On the contrary, if all translations are different, the stability index will be minimal (0).

3 Typology of polysemic relations

According to (Resnik and Yarowsky, 1999), the cross-lingual estimate of sense proximity introduced

above is highly consistent with the sense groupings of the Hector database, as used in the Senseval-1 evaluation. However, in our opinion, the hierarchical structure of senses in Hector (and dictionaries in general) does not necessarily reflect sense proximity. Metaphorical sense extensions of a word meaning are a good example: while they are closely related (in such hierarchical arrangement of senses) to the source meaning, the metaphorical sense usually belongs to a different semantic field. If the cross-lingual measure of sense proximity is also high for such metaphors, that would mean that they are highly generalized across languages, but not that the meanings are related.

In addition, WordNet 1.7, which replaces Hector as sense inventory in Senseval-2, does not provide such an explicit arrangement of word senses. Thus, we decided to classify sense pairs according to a simple typology of sense extensions (including homonymy as absence of relation) and to verify that the proximity measure is in agreement with such classification.

We have considered three types of semantic relation, previously introduced in (Gonzalo et al., 2000):

- **metonymy** (semantic contiguity), for example, *yew-tree* and *yew-wood* or *post-letters* and *post-system*.
- **metaphor** (similarity), for example, *child-kid* and *child-immature*.
- **specialization/generalization** (based on extending or reducing the scope of the original sense), for example, *fine-greeting* and *fine-ok*.
- **homonymy** (no relation). For example, *bar-law* and *bar-unit_of_pressure*.

4 Experiment design

Following (Resnik and Yarowsky, 1999), we carried out an experiment based on free annotations with the first preferred translation of Senseval-2 nouns and adjectives in hand-disambiguated contexts.

11 native or bilingual speakers of 4 languages[1] with a level of English proficiency from medium to high were asked to translate marked words in context into their mother tongue.

As working material, we used part of the Senseval-2 data. Whenever possible, we selected three contexts (with the highest inter annotator agreement) for each of the 182 noun and adjective senses in the Senseval-2 English test suite. However, for 16 senses there was only one instance in the corpus and 6 senses had only two occurrences. The final data set was composed of 508 short contexts for 182 senses of 44 words of the Senseval-2 test suite. These instances, randomly ordered, were presented to annotators without sense tag, so that each tagger had to deduce the sense of the marked word from the context and type its first preferred translation into his language in the answer line. This is an example of the input for annotators:

> **fine 40129**
> Mountains on the other side of the valley rose from the mist like islands, and here and there flecks of cloud, as pale and <tag>**fine**</tag> as sea-spray, trailed across their sombre, wooded slopes.
> ANSWER: * *

The collected data was used to compute proximity for every sense pair w_i, w_j in the sample. Stability was computed using the same data, for all senses in the sample except for 16 cases which had one instance in the corpus.

In order to evaluate how using a more extensive and more varied set of languages can affect the results of the experiment, we selected a smaller control subset of 65 instances of 23 senses corresponding to 3 nouns and 2 adjectives. Annotations for this subset were collected from 29 speakers of 12 languages [2] covering 5 language families.

5 Results and discussion

Distribution of proximity and stability indexes in the main set (the whole set of senses tagged in 4 languages) is shown in Figure 1 and Figure 2.

As can be seen, few sense pairs in Senseval-2 data have been assigned a high proximity index. This means that most of the senses considered in Senseval-2 are adequate distinctions to be used in a WSD evaluation. The average proximity (0.28) is

[1] This main set of languages includes Bulgarian, Russian, Spanish and Urdu.

[2] This set of languages included Bulgarian, Danish, Dutch, Finnish, German, Hungarian, Italian, Portuguese, Rumanian, Russian, Spanish and Urdu.

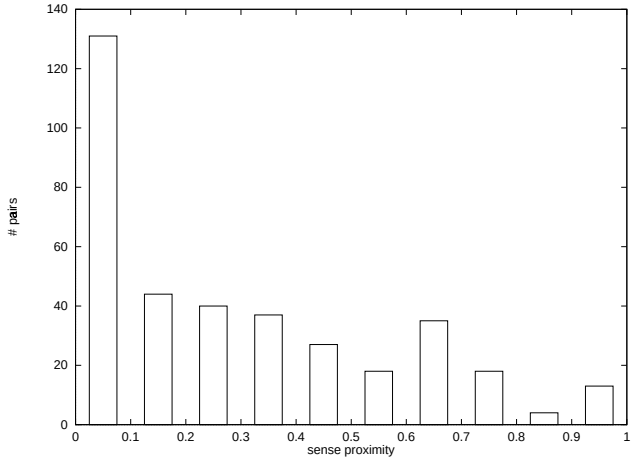

Figure 1: Global distribution of sense proximity

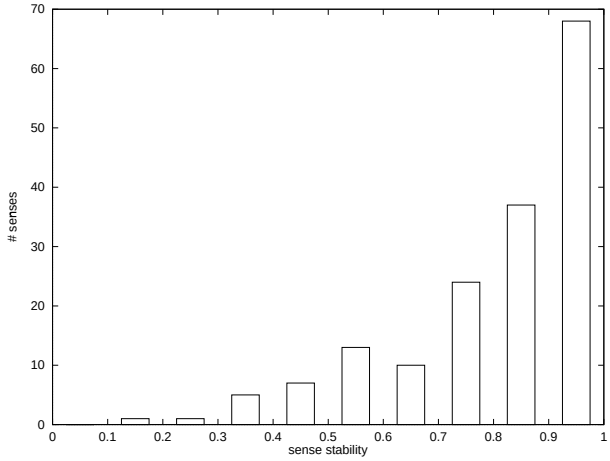

Figure 2: Distribution of sense stability

close to Resnik and Yarowsky's result for a sample of Senseval-1 data.

Global stability is very high, which is also a positive indication (both for sense granularity and for the quality of the annotated instances). The average is 0.80, and Figure 2 shows that the distribution is highly skewed towards the high end of the graph. The discussion of the few cases with low stability is included in section 5.4 below.

5.1 Language family distance

We compared the results for the main set (all annotations in four languages), with the results for the control set (a subset of the annotations in 12 languages).

Figure 4 shows the average semantic proximity obtained for the whole Senseval-2 test suite annotated in 4 languages, and for the subset of 23 senses (65 instances) annotated in 12 languages. The average difference is large (0.29 vs 0.48); however, a direct comparison only for the senses annotated in both samples gives a very similar figure (0.49 vs 0.48). Therefore, it seems that the effect of adding more languages is not critical, at least for this sense inventory.

5.2 Proximity matrices

Stability and proximity indexes have been integrated into proximity matrices as shown in the example in Figure 3. Stability is shown in the diagonal of the matrices, and proximity in the cells above the diagonal.

mouth	cave	lips	oral
cave	0.96	0.13	0.13
lips		1.00	1.00
oral			1.00

Figure 3: Semantic proximity matrix for *mouth*

On the basis of the translation criterion, all the senses of *mouth* in the example have high stability. Proximity indexes point at two close senses: *mouth-lips* and *mouth-oral cavity*, whereas *mouth* as *opening that resembles a mouth (as of a cave)* appears to be rather distant from the other two, confirming our intuitions.

These matrices (especially the non-diagonal proximity values) can be used to re-score Senseval-2 systems applying the measures proposed in (Resnik and Yarowsky, 1997; Resnik and Yarowsky, 1999).

5.3 Similarity and semantic relations between senses

As mentioned in Section 3, all the sense pairs have been manually classified according to the adopted typology. Figures 5, 6, 7, 8 show the distribution of semantic proximity according to this classification of sense relations holding between senses of the same word.

5.3.1 Homonyms

As expected, most homonym pairs displayed low proximity. Only few very specific cases were an exception, such as the pair formed by *bar* in the sense

Main set				Control set			
language	*prox.*	*family*	*# taggers*	*language*	*prox.*	*family*	*# taggers*
Bulgarian	0.30	Slavonic	1	Bulgarian	0.54	Slavonic	1
Russian	0.25	Slavonic	1	Russian	0.39	Slavonic	1
Spanish	0.31	Romance	8	Spanish	0.53	Romance	8
Urdu	0.28	Indo-Iranian	1	Urdu	0.47	Indo-Iranian	1
				Hungarian	0.56	Fino-Hungarian	1
				Italian	0.76	Romance	6
				Portuguese	0.44	Romance	2
				Rumanian	0.59	Romance	1
				Danish	0.48	Germanic	3
				Dutch	0.40	Germanic	1
				Finnish	0.26	Fino-Hungarian	2
				German	0.38	Germanic	1
Average	**0.29**			**Average**	**0.48**		

Figure 4: Average sense proximity

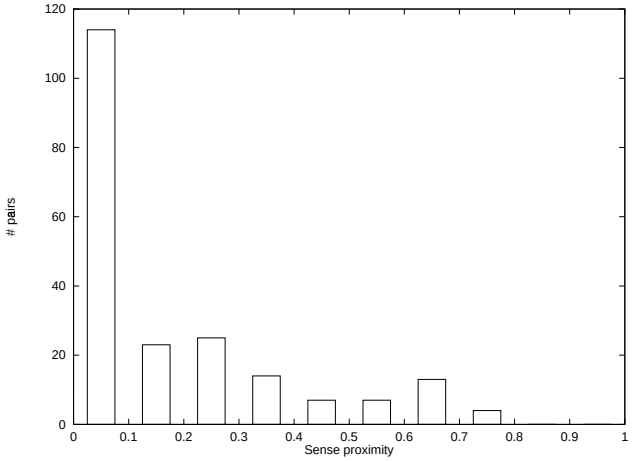

Figure 5: Distribution of homonyms

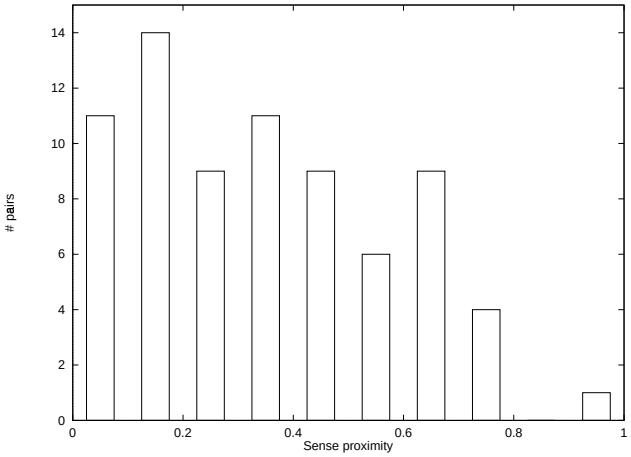

Figure 7: Distribution of metaphors

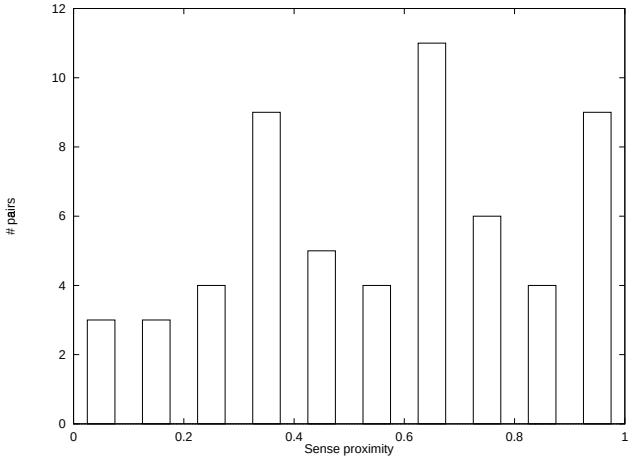

Figure 6: Distribution of metonymy

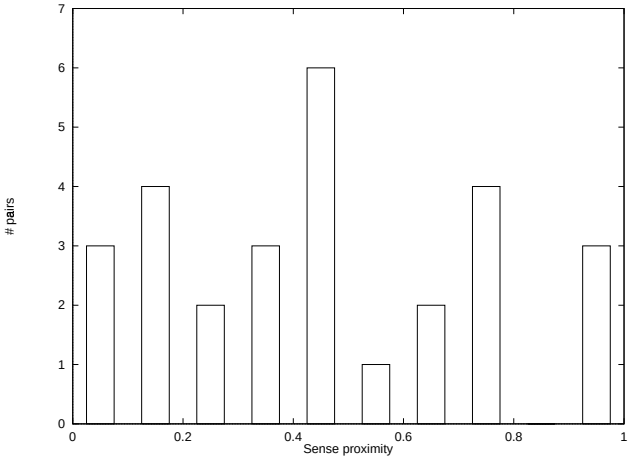

Figure 8: Distribution of specializations

of *establishment* and *bar* in the sense of *unit of pressure*. These two senses have been loaned by several languages from English. Therefore, in spite of being unrelated, these senses yielded a proximity of 0.69. Except for exceptional cases of this kind, the method based on multilingual translations proved to be valid for capturing homonyms. However, it is clear that an explicit account of homonymy in the sense inventory (not available in wordnet) would prevent such erroneous assignments.

5.3.2 Metaphors

One of the features of metaphors is their capability of linking remote and unrelated semantic domains. Then, sense pairs classified as metaphorically related should have low proximity indexes. Unexpectedly, we found that this was not always true. Proximity of 27% of metaphoric sense pairs was equal or greater than 0.50. Since all of them were examples of very extended (if not universal) metaphoric patterns like *blind-sight* and *blind-irrational* or *cool-cold* and *cool-colour*, it seems that calculating sense distance using only the multilingual translation method does not always yield good estimations.

5.3.3 Specialization/Generalization

The sense pairs tagged as instances of specialization/generalization, in general, behaved as we expected, although there also appeared few cases that contradicted our predictions (medium or high proximity indexes). The exceptions involved several senses of *fine* (*fine-superior to the average*, *fine-being satisfactory or in satisfactory condition*, *fine-all right, being in good health* and *fine-of weather*). Besides technical issues (discussed in section 5.4), we believe there are problems of overlapping sense definitions in WordNet 1.7. Compare for instance:

> fine 1, good – (superior to the average; ”in fine spirits”; ”a fine student”; ”made good grades”; ”morale was good”; ”had good weather for the parade”)

with

> fine 9 – ((of weather) pleasant; not raining, perhaps with the sun shining; ”a fine summer evening”)

and

> fine 2, all right(predicate), all-right(prenominal), ok, o.k., okay, hunky-dory – ((informal) being satisfactory or in satisfactory condition; ”an all-right movie”; ”the passengers were shaken up but are all right”; ”is everything all right?”; ”everything’s fine”; ”things are okay”; ”dinner and the movies had been fine”; ”another minute I’d have been fine”)

with

> fine 5, all right, fine – (being in good health; ”he’s feeling all right again”; ”I’m fine, how are you?”)fine-all right, being in good health

Indeed, *fine 2*, for example, is one of the senses with lowest stability (0.33) in the sample.

5.3.4 Metonymy

As for metonymy, the distribution of proximity indexes indicates that this kind of relation seems to include different subpatterns (Gonzalo et al., 2000). A further study of metonymically related senses is needed in order to correctly interpret sense proximity in these cases.

Overall, the evidence provided by the classification of the results suggests that calculating sense similarity using multilingual translations is a good first approximation that should be combined with additional criteria based on a qualitative consideration of sense relations.

5.4 Consistency of the data

The analysis of the results revealed aspects of the experiment design worth mentioning. Free use of synonyms seems to be one of the factors affecting both sense proximity and stability. As the condition of being coherent and using the same translation for all instances of the same sense was not imposed in the experiment instructions, it is no surprise that some taggers opted for variability.

Obviously, the inter-annotator agreement for our data is rather low (54%)[3] due to the extensive (and free) use of synonyms. Even if intra-annotator variations would have been prevented, there seems to be no feasible way of guaranteeing that different speakers of the same language choose the same term among several synonyms.

A closer look at sense pairs with unexpectedly low proximity indexes showed that, besides the syn-

[3]The inter-annotator agreement index has been calculated using Spanish translations provided by 8 annotators. We found that two taggers gave the same answer in 54% of cases.

onyms issue, there are other factors that should have been considered in the experiment design:

• Different syntactic realizations:

1. `N N` $\longrightarrow$ `Adj N` | `N Adj` | `N Prep N`

An English noun modifying another noun becomes an adjective or a preposition phrase in languages such as Russian, Bulgarian or Spanish. In the example below, the marked word is translated with a noun and with an adjective. Both translations have the same root, but they were computed as different translations by the exact-match criterion:

> 1. Charles put on a low dynamic vice which rose in crescendo to an order. "Listen, Platoon. Every man who can beat me to the windmill is excused all <tag>**fatigues**</tag>
> ANSWER (Russian): *__rabota__*
> 2. Though in <tag>**fatigue**</tag> uniform, they were almost as smart and well turned out as Charles had seen such soldiers in peace time behind the railings of Wellington barracks.
> ANSWER (Russian): *__rabochy__*

2. `Adj` $\longrightarrow$ `Adv`

English adjectives in predicative position become adverbs in other languages. For example, these two instances of *fine*:

> 1. Young Duffy was in <tag> **fine** </tag> form when he defeated B. Valdimarsson of Iceland 4–0
> 2. Mr. Frank told Mr. Pilson that Olympics officials wanted $290 million or more for TV rights to the 1994 Winter Games in Norway. The CBS official said that price sounded <tag>**fine**</tag>

have a very close meaning. However, in languages such as Spanish or Russian the first one is translated by an adjective and the second one by an adverb. This caused that adverb translations were computed negatively, as they did not match exactly with the form of the corresponding adjective.

• Collocations

Collocations constituted another problem, that should have been foreseen when selecting representative contexts for senses. Some of the instances turned out to be part of complex expressions and could not be naturally translated separated from the rest of the collocation. Some examples are:

> 1. As the waves crashed round the hilltops the wizards' palaces **broke** <tag>**free**</tag> and floated on the surface of the waves.
> ANSWER (Spanish): *__librarse__*
> 2. In these circumstances, executives **feel** <tag>**free**</tag> to commit corporate crimes.
> ANSWER (Spanish): *__no cortarse__*
> 3. Suddenly they come on cos they don't give as much notice, they are a pain in the <tag>**bum**</tag>
> ANSWER (Spanish): *__pesado__*

• Quality of Senseval-2 annotations

Finally, some erroneous manual annotations of the Senseval-2 corpus were highlighted by unexpected proximity or stability values. For instance, this sense of fine:

> fine - ((metallurgy); free or impurities; having a high or specified degree of purity; "gold 21 carats fine").

was used to tag (incorrectly) the following instances:

> **fine 40089**
> An NTA regional commendation was awarded to Hickson & Welch, <tag>**fine**</tag>chemical specialist, for its training initiative to develop multi-skilled employees and "world-class" standards in safety and efficiency.
> **fine 40144**
> There are many custom and <tag>**fine**</tag>chemical manufacturers but few, if any, have EFC's long experience in multi-step, complex, organic synthesis – knowledge gained from years of experience in the manufacture of photochemicals, dyes and pharmaceuticals.
> **fine 40162**
> The manufacturer's safety data sheet warns of a potentially hazardous reaction between sodium borohydride and <tag>**fine**</tag>dispersed heavy metals but this reaction with charcoal and solid sodium borohydride are stored and handled in such a way that the risk of contact between them is avoided.

Probably the extensive use of Senseval-2 data will permit pruning such kind of errors in a near future.

6 Conclusions

We have provided a qualitative and quantitative characterization of the subset of WordNet 1.7 used as sense inventory in Senseval-2. Individual senses are given a *stability* measure that indicates their degree of homogeneity and whether they should be revised for further division into sub-senses. Sense pairs are classified as homonyms, generalization/specialization, metonymic or metaphorical extensions. In addition, a *proximity* measure gives a numerical account of their similarity.

The experiment conducted for the Senseval-2 test suite supports the validity of the proposals in (Resnik and Yarowsky, 1999). Multilingual translations collected for a set of monolingual senses provide a helpful measure of semantic proximity. In addition, the same data can also be used to measure sense stability successfully. The matrices obtained in the experiment are of a practical interest: they can be used to re-score Senseval-2 systems taking semantic distance into account.

Our global results indicate that WordNet 1.7 is a reliable sense inventory for developing and testing WSD systems. Cases of very close sense pairs or too coarse sense definitions are marginal according to our data.

We have also provided some evidence that cross-lingual estimation of sense proximity should however be combined with some additional criteria related to the nature of sense relations. In particular, an explicit account for homonyms and metaphors in WordNet would help to correct too high estimations of the translation criterion.

The full set of data obtained in the experiments, including proximity matrices for all nouns and adjectives in the Senseval-2 test suite, can be downloaded from:

http://sensei.lsi.uned.es/senseval2

Acknowledgments

This work has been funded by a Spanish government grant, project Hermes (CICyT TIC2000-0335-C03-01), by a UNED PhD. grant. We would also like to thank Adam Kilgarriff, whom we owe the idea of this study, and the ITRI (University of Brighton) for their help and support. Our acknowledgments go as well to the volunteers who annotated the corpus and without whom this study would not have been possible.

References

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*.

Julio Gonzalo, Irina Chugur, and Felisa Verdejo. 2000. Sense clustering for information retrieval: evidence from Semcor and the EWN InterLingual Index. In *Proceedings of the ACL'00 Workshop on Word Senses and Multilinguality*.

A. Kilgarriff and M. Palmer. 2000. Special issue on Senseval. *Computers and the Humanities*, 34(1-2).

P. Resnik and D. Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *Proc. ACL SIGLEX Workshop on tagging text with lexical semantics: why, what and how?*

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5.

P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*.

M. Sussna. 1993. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Base Management, CIKM'93*.