Natural Language Engineering **12** (1): 1–27. © 2006 Cambridge University Press
doi:10.1017/S135132490500402X Printed in the United Kingdom

Making fine-grained and coarse-grained sense distinctions, both manually and automatically

MARTHA PALMER
Department of Linguistics, University of Colorado, CO, USA
e-mail: mpalmer@colorado.edu

HOA TRANG DANG
National Institute of Standards and Technology, Gaithersburg, MD, USA
e-mail: hoa.dang@nist.gov

CHRISTIANE FELLBAUM
Princeton University, NJ, USA
e-mail: fellbaum@clarity.princeton.edu

(*Received 12 August 2002; revised 12 January 2005*)

Abstract

In this paper we discuss a persistent problem arising from polysemy: namely the difficulty of finding consistent criteria for making fine-grained sense distinctions, either manually or automatically. We investigate sources of human annotator disagreements stemming from the tagging for the English Verb Lexical Sample Task in the Senseval-2 exercise in automatic Word Sense Disambiguation. We also examine errors made by a high-performing maximum entropy Word Sense Disambiguation system we developed. Both sets of errors are at least partially reconciled by a more coarse-grained view of the senses, and we present the groupings we use for quantitative coarse-grained evaluation as well as the process by which they were created. We compare the system's performance with our human annotator performance in light of both fine-grained and coarse-grained sense distinctions and show that well-defined sense groups can be of value in improving word sense disambiguation by both humans and machines.

1 Introduction

Highly ambiguous words pose continuing problems for Natural Language Processing (NLP) applications. They can lead to irrelevant document retrieval in IR systems, and inaccurate translations in Machine Translation systems (Palmer, Han, Xia, Egedi and Rosenzweig 2000). Several efforts have been made to develop automatic Word Sense Disambiguation (WSD) systems that are capable of addressing these problems (Ide and Véronis 1998; Palmer and Light 1999). While homonyms[1] like *bank* are fairly tractable, polysemous words like *run*, with related but subtly distinct meanings, present the greatest hurdle for WSD. The most polysemous words are not

[1] word forms with multiple unrelated meanings.

only the most frequently occurring ones, but many of their senses are also domain-independent, making the WSD problem ubiquitous. The Senseval (Kilgarriff and Palmer 2000; Edmonds and Cotton 2001) exercises for evaluating automatic WSD systems attempt to create corpora annotated with sense tags to enable the training and testing of supervised WSD systems, and in the process raise questions about how to define the senses in the first place.

In this paper, the central question we ask is, *Which senses CAN be distinguished?*, and we examine both manual and automatic tagging results in our quest for an answer. There is a separate but related question which is equally important, *Which senses NEED to be distinguished?* The answer to this, however, is more contextually dependent, in that it can vary from application to application. We will touch on it briefly, with some illustrative examples, at the end of the paper.

In section 2 we review the Senseval-1 and Senseval-2 exercises and the impact the choice of sense inventory (Hector vs. WordNet) had on them. Section 3 discusses general criteria for sense distinctions with examples from both inventories. Next, sections 4 and 5 present our semantic groupings of related WordNet senses and the quantitative evaluation of their effect on both manual and automatic tagging results. We also present our preliminary set of criteria for creating the groupings. We found that the same sense distinctions often prove troublesome for both human taggers and automatic systems. Using our independently derived groupings as a more coarse-grained set of sense distinctions results in similar improvements for both manual and automatic tagging scores. Section 6 concludes with an informal discussion of the utility of our group distinctions for applications such as Machine Translation. We see the groupings as a promising avenue for achieving more accurate automatic word sense disambiguation systems.

2 Senseval tagging exercises

The methodology for applying supervised machine learning techniques to WSD involves a series of steps, beginning with the preparation of tagged data and a corresponding evaluation of its quality. The data are typically a large number of naturally occurring sentences containing a given word, each of which has been tagged with a pointer to a sense entry from a pre-existing sense inventory (a computational lexicon or machine-readable dictionary). A section of the tagged data is used for training, while another section is reserved for testing purposes. Unsupervised machine learning systems and rule-based systems can also be evaluated against the same test data, which is considered to be a Gold Standard. Where the sense inventory provides levels of granularity with respect to the entries the evaluation metric can provide both fine-grained and coarse-grained scores. Since the consistency of the systems cannot be expected to surpass that of humans, high interannotator agreement provides reassurance of the quality of the tagged data. This is, in turn, facilitated by a high quality sense inventory with clear sense distinctions. Unfortunately, sense inventories for a language can be discouragingly diverse, with significant differences with respect to entries for polysemous words (Atkins and Levin 1991), raising doubts about the utility of the tagged data. Since the first two Senseval evaluation exercises used different sense inventories, they provide an opportunity to study the impact of different sense inventories on system performance and inter-annotator agreement.

SENSEVAL-1 The first exercise in automatic WSD, SENSEVAL-1 (Kilgarriff and Palmer 2000), used a DARPA-style evaluation format where the participants were provided with hand-annotated training data and test data and a pre-defined metric for evaluation. The evaluation scheme provided a scoring method for exact matches to fine-grained senses as well as one for partial matches at a more coarse-grained level. ROMANSEVAL, an evaluation for French and Italian, was run in parallel (Véronis and Segonde 2000; Calzolari and Corazzari 2000).

The lexical inventory for SENSEVAL-1 was the Hector lexicon, developed jointly by DEC and Oxford University Press using a corpus-based approach and traditional hierarchical dictionary entries (Kilgarriff and Rosenzweig 2000).[2] After selecting the 34 target lexical items, professional lexicographers tagged sentences containing those items that had been extracted from the Hector corpus. By allowing for discussion and revision of confusing lexical entries before the final test data was tagged, inter-annotator agreement (ITA) of over 80% was eventually achieved. Replicability was also measured. Replicability is determined by having two different teams of taggers tag the same instances in parallel. After adjudication of each set, the agreement between the sense tags for the two sets is measured. Replicability for four SENSEVAL-1 words (*generous, onion, sack, shake*) was 95.5%. The initial ITA of each team was in the 80s. In general the 24 participating systems did surprisingly well, with several of the supervised systems getting precision and recall numbers in the high 70s and low 80s on a data set with an average polysemy of 10.7 (Kilgarriff and Rosenzweig 2000). The evaluation metric allowed for both fine-grained scores for exact matches and coarse-grained scores where the tag chosen was a daughter, parent or sibling of the correct tag, based on the entry's hierarchical structure as encoded in the sense inventory. The best scoring system achieved a fine-grained score of 77.1%[3] accuracy and a coarse-grained score of 81.4%. In general, the lower the system performance, the larger the gap between the fine-grained and coarse-grained scores. The highest fine-grained score on just the verbs, which had an average polysemy of 7.79, was 70.9%. See Table 1 for the complete results.

The SENSEVAL-1 workshop provided convincing evidence that automatic systems can perform WSD satisfactorily, given clear, consistent sense distinctions and suitable training data. However, the Hector lexicon was very small and under proprietary constraints, and the question remained whether it was possible to have a publicly available, broad-coverage lexical resource for English (or any other language) with the requisite clear, consistent sense distinctions.

[2] An example hierarchical entry from Hector: *bother*: 1. intransitive verb, (make an effort), after negation, usually with to infinitive; (of a person) to take the trouble or effort needed (to do something). Ex. "About 70 percent of the shareholders did not bother to vote at all." 1.1 (can't be bothered), idiomatic, be unwilling to make the effort needed (to do something), Ex. "The calculations needed are so tedious that theorists cannot be bothered to do them."

[3] The systems we discuss here all attempt every possible tagging, so there is no need to report separate precision and recall scores

Table 1. *Accuracy of the LESK-CORPUS baseline and the best, average and worst performing systems in* Senseval-1, *broken down by part of speech*

POS	Baseline	Best	Average	Worst
Verbs	0.700	0.709	0.610	0.421
Nouns	0.569	0.865	0.635	0.388
Adjs	0.717	0.777	0.615	0.377
All	0.719	0.787	0.544	0.161

Senseval-2 – *The English Verb Lexical Sample Task* Subsequently, the Senseval-2 (Edmonds and Cotton 2001) exercise was run, which included WSD tasks for 12 languages. A concerted effort was made to use existing WordNets as sense inventories because of their wide-spread popularity and availability. English WordNet is a large electronic database organized as a semantic network built on paradigmatic relations like synonymy, hyponymy, antonymy, and entailment (Miller *et al.* 1990; Miller and Fellbaum 1991; Fellbaum 1998b), and this approach has now been ported to several other languages. The English lexical sample task for Senseval-2 involved 73 lexical items (29 verbs, and the rest nouns and adjectives) taken from WordNet 1.7 and was the result of a collaboration between the authors, who provided training/test data for the verbs and the all-words task,[4] and Adam Kilgarriff, who provided the training/test data for the nouns and adjectives (Kilgarriff 2001; Palmer *et al.* 2001). Between 75 and 300 instances of each word in the lexical sample task were hand-tagged, depending on the number of senses for the word; the formula of 75 plus 15*n*, given *n* senses, was roughly adhered to in determining the number of instances to tag. Multi-word constructions in the corpus (e.g., "call attention to") were explicitly marked for head word ("call") and all satellites in the construction ("attention to"). The data came primarily from the Penn Treebank II Wall Street Journal corpus (Marcus *et al.* 1993), but was supplemented with data from the British National Corpus whenever there was an insufficient number of Treebank instances. The instances for each word were partitioned into training/test data using a ratio of 2:1.

Because the verbs were the most polysemous words in Senseval-2, they will remain the focus for the rest of the paper. The lexical sample verb task consisted of twenty-nine verbs, with an average polysemy of 16.28 senses using the pre-release version of WordNet 1.7. These were chosen from among the most polysemous verbs in the all-words task. Double blind annotation by two linguistically trained annotators was performed on corpus instances, with a third linguist adjudicating between inter-annotator differences to create the "Gold Standard." Most of the revisions of sense definitions relevant to the English tasks were done by the adjudicator prior to the bulk of the tagging, although there was much less discussion among the taggers of

[4] For the details of this task, which involved 5K words of running text consisting of three Penn TreeBank II articles, see Palmer, *et al.*, (Palmer *et al.* 2001). A simple baseline strategy which simply tags each head word with the first WordNet sense for the corresponding Treebank part-of-speech tag, has a score of 57%, as compared to the best system score of 69%. Complete results are at http://www.sle.sharp.co.uk/senseval2/

Table 2. *Accuracy of the LESK-CORPUS baseline and the best, average and worst performing systems in* Senseval-2*, broken down by part of speech*

POS	baseline	best	average	worst
Verbs	0.445	0.576	0.419	0.186
Nouns	0.547	0.695	0.540	0.244
Adjs	0.591	0.732	0.572	0.216
All	0.512	0.642	0.489	0.239

how senses were to be applied than there had been with the Senseval-1 taggers. The average inter-annotator agreement (ITA) rate achieved with these verb senses was 71% (see Table 7) which is comparable to the 73% agreement for all words for SemCor, a previous tagging project using WordNet 1.4. (Fellbaum *et al.* 1997; Fellbaum *et al.* 1998). The nouns and adjectives, which were less polysemous overall, have an ITA of 85% (see Table 2).

WordNet does not offer the same type of hierarchical entry that Hector does, so the verbs were also grouped by two or more people, with differences being reconciled, and the sense groups were used for coarse-grained scoring of the systems. Using these independently derived grouped senses the inter-annotator agreement figures rose to 82%. Section 3 contains a detailed discussion of the criteria for grouping and the impact of the groups on ITA.

For system comparisons we ran several simple baseline algorithms similar to the ones that had been used in Senseval-1, including COMMONEST, LESK (Lesk 1986), LESK-DEFINITION, and LESK-CORPUS (Kilgarriff and Rosenzweig 2000). In contrast to Senseval-1, in which none of the competing systems performed significantly better than the highest baseline, this time most of the systems performed comparably to the highest baseline (LESK-CORPUS, at 45.5%), with approximately half performing better, and the top system achieving 57.6% (Palmer *et al.* 2001) on the verbs alone, with 64.2% on the overall task (nouns, verbs and adjectives). Again the groupings which were used for coarse-grained scoring produced significantly better results for most systems for most verbs. Our own system, which was not officially entered in the exercise, performed well at 62.5% for verb senses and 71.7% for grouped senses (Dang 2004). For the entire lexical sample task (verbs, nouns and adjectives), the highest system scores (from Johns Hopkins University) were 64.2% fined-grained and 71.3% coarse-grained.[5] In general the nouns and adjectives had lower polysemy and higher scores (71.8% score, ITA 85%, polysemy 4.9) (Yarowsky, Florian, Cucerzan and Schafer 2001). See Table 2 for the complete results.

Senseval-3 The most recent Senseval, Senseval-3, was held in Barcelona, Spain in conjunction with ACL-04 (Senseval 2004). The scope of the evaluation expanded yet again, this time including 16 different tasks and literally hundreds of teams.

[5] They have since improved their overall fine-grained score to 66.3%.

One major difference was that the English lexical sample task tried to avoid the expensive overhead of a supervised manual tagging project and made use of the Open Mind Word Expert interface to collect tagged instances. This resulted in a fair amount of data being produced, although the ITA was somewhat lower than that attained by more traditional methods, 62.8% for single words (Mihalcea and Kilgarriff 2004). There were many new techniques described at the workshop, although on the whole system performance is still clearly tied to ITA, and when this is low, system performance follows suit. Given the lower ITA, we have not included this data in our current discussion.

Comparison of Tagging Exercises Prior to the SENSEVAL-2 exercise, there were concerns expressed about whether or not WordNet had the requisite clear, consistent sense distinctions. Both the inter-annotator agreement figures and the performances of the systems are lower for SENSEVAL-2 than for SENSEVAL-1, which seemingly substantiates these concerns (see Tables 1 and 2). However, in addition to the differences in sense inventories, one must also bear in mind the highly polysemous nature of the SENSEVAL-2 verbs which are on average twice as polysemous as the SENSEVAL-1 verbs, an average polysemy of 16.28 compared to 7.79.[6] High polysemy has a detrimental effect on both manual and automatic tagging, although it does not correlate negatively with system performance as well as entropy does (Palmer *et al.* 2001).

We can get a better comparison of the quality of the tagged data (and, indirectly, of the sense inventories) for SENSEVAL-1 and SENSEVAL-2 by comparing the performance of our automated system on similar subsets of data from the two exercises. Does the system perform comparably given data for verbs of similar polysemy? To test this, we first found the most polysemous verbs from SENSEVAL-1, *bury, float* and *seize*, with a polysemy in Hector of 15, 18 and 11, respectively.[7] Factoring out the verbs with lower polysemy and discarding the phrasal filter from our system[8] which was only applicable for SENSEVAL-2, we find very little difference in the system performance: 59.7% for SENSEVAL-1 versus 60.0% for SENSEVAL-2 with a baseline well below 45%. This small sample indicates that even with different sense inventories, when controlling for polysemy, SENSEVAL-2 data gives rise to very similar system performance as SENSEVAL-1 data. The simplest explanation of the lower system performance overall on SENSEVAL-2 is therefore the higher average polysemy of the verbs in the task. It is likely that, in spite of the lower inter-annotator agreement for SENSEVAL-2, the double blind annotation and adjudication

[6] Overall polysemy for SENSEVAL-1 is 10.7. The Hector sense inventory is more hierarchical and makes different sense distinctions, but on the whole has a total number of senses for individual words that is similar to WordNet's.

[7] Their WordNet 1.7 polysemy Figures are 6, 8, and 8, illustrating the variable nature of sense distinctions across different lexical resources. (Atkins and Levin 1991)

[8] In contrast to SENSEVAL-1, senses involving multi-word constructions in SENSEVAL-2 could be directly identified from the sense tags themselves (through the WordNet sense keys that were used as sense tags), and the head word and satellites of multi-word constructions were explicitly marked in the training and test data.

provided a reliable enough filter to ensure consistently tagged data with WordNet senses.

We are still faced with the challenge of improving ITA and system performance on highly polysemous lexical items, regardless of the sense inventory being used. As an additional check, we had our taggers re-tag a subset of the Senseval-1 data to see if we could replicate the higher ITA results. We tagged 35 words (at least 10 each of nouns, verbs, and adjectives). For nine of the words we tagged all of the Senseval-1 instances, several hundred for each word. Due to time constraints, for the remaining 26 words, we took fifty instance subsets for each word, ensuring the different senses were distributed as evenly as possible. For the 9 large data-set words the ITA was 81.1%; for the 50-instance words the ITA was 75.8%, with an overall average of 80.1%. This is in keeping with the reported ITA of over 80% for Senseval-1, and is certainly much higher than the 71% ITA for the Senseval-2 verbs. This is almost certainly due to the lower polysemy and not just the different sense inventories. It is not surprising that the larger data sets result in higher ITA. This may be be partly an artifact of practice, but is also because these larger sets contain a much higher proportion of the most frequent senses, so they have a higher baseline.

We also considered the effect of training set size on the performance of the systems. Ignoring outliers, there were on average half as many training samples for each verb in Senseval-2 as there were in Senseval-1. However, the smaller set of training examples did not seem to be a major factor in the performance of our system on verbs of similar polysemy in Senseval-1 and Senseval-2. Others have found that the accuracies of automatic WSD systems over all parts of speech (nouns, verbs, and adjectives) of Senseval-2 increased as training sizes increased (Yarowsky and Florian 2002). Although we also found additional data useful for Chinese sense tagging (Dang, Chia, Chiou and Palmer 2002), when we used 10-fold cross-validation and enlarged our Senseval-2 training set for verbs by using a partition of 9:1 instead of 2:1, we found a relative improvement in accuracy of only 2.0%. However, because these training set sizes were increased by only 35%, further experimentation is needed to determine whether or not significantly more training data would benefit high polysemy verbs.

Given the close correlation between lower polysemy and higher ITA, we feel this is an important angle to pursue. In the next section, we will first examine the nature of sense distinctions, and the sources of sense tagging disagreements. We then present our criteria for creating sense groups, and discuss the impact these groups have on the human tagger disagreements as well as automatic tagging errors for highly polysemous verbs. One can take comfort from the knowledge that the majority of lexical items do not exhibit the high polysemy of the verbs discussed here.

3 Sense groupings

The difficulty of achieving accurate data for sense tagging has been thoroughly attested to in the literature (Kilgarriff 1997; Hanks 2000). There is little optimism about finding criteria for making indisputable sense distinctions, with difficulties being found with truth-theoretical criteria, linguistic criteria and definitional criteria (Jones 1986; Geeraerts 1993). Several decades ago, Karen Sparck Jones proposed data-driven synonym sets as the only reliable means of characterizing a word's behavior, similar to the approach later adopted by WordNet. In spite of the proliferation of dictionaries, there is no current methodology by which two lexicographers working independently are certain to derive the same set of distinctions for a given word. Even given identical corpus-based examples there are still many fairly arbitrary judgements for the lexicographer to make, such as when to stretch an existing sense to encompass extended meanings, and when to create a new sense. The inherent complexity of objects ensures that references to them must often be multi-faceted (Cruse 1986; Asprejan 1974; Pustejovsky 1991). Events are at least equally complex and perhaps even more difficult to characterize (Talmy 1991). The inherent fluidity of language ensures that a definition of a word is a moving target; as long as it is in use its meaning could continue to expand. One of the greatest challenges for the creation of a static sense inventory lies in the complex and constantly changing nature of the vocabulary, bringing into question the feasibility of the sense tagging task.

The mapping between Hector and WordNet 1.6 that was made available for Senseval-1 provides striking evidence of the different choices lexicographers can make in determining sense distinctions. It is immediately apparent that Hector and WordNet often have different numbers of senses for the same lemma (see footnote 7). Closer examination of individual words such as *shake* reveals even more fundamental mismatches. Hector and WordNet entries for *shake* have the same number of main senses (8). However, there is variation in the verb-particle constructions they have chosen to include, with the result that Hector has 27 total senses while WordNet only has 15. At a more fundamental level, while Hector distinguishes between *shaking hands with someone*, *shaking one's fist* and *shaking one's head*, WordNet does not. Hector also distinguishes between the unaccusative TREMBLE sense, *My hands were shaking from the cold*, and the more active, transitive, causative MOVE sense, *He shook the bag violently*, where someone intentionally moves an object back and forth. WordNet collects these together, along with *She shook her cousin's hands*, as WN1, and instead makes distinctions with respect to the type of motion: WN2, gentle tremors; WN3, rapid vibrations; or WN4, swaying, which Hector does not. These distinctions can all be seen as justifiable choices, but they carve the semantic space up in very different ways.

As we demonstrate below, coarser-grained sense distinctions can sometimes alleviate the difficulties involved in mapping between sense inventories, as well as reconcile inter-annotator disagreements. We begin by introducing the criteria for creating the groups which led to significant revisions of pre-existing WordNet groups, and discuss the factors behind their positive impact on performance. There are situations where, rather than trying to force an exact match with a fine-grained sense, it may be more prudent to equivocate by choosing a less-specific cluster of senses.

3.1 Using Levin classes to group shake

Our interest in grouping was initially sparked by the mismatched Hector and WN entries described above, and our attempts to reconcile them. *Shake* proved especially amenable to grouping, and most of the *shake* differences mentioned above were resolved by the groups. Our *shake* groups were inspired by our study of Levin classes, where verbs are grouped together based on their ability to appear in similar sets of syntactic frames which are assumed to reflect underlying semantic similarities (Levin 1993; Dang *et al.* 1998; Dang *et al.* 2000; Kipper *et al.* 2000). These frames often correspond to syntactic alternations such as indefinite object drop, [*We ate fish and chips./We ate at noon.*]; cognate object realization, [*They danced a wild dance./They danced.*]; and causative/inchoative [*He chilled the soup./The soup chilled.*].

Several basic senses of *shake* appear in different Levin classes from which we derived five major, coarse-grained sense divisions, each one of which can be subdivided further. The 27 Hector *shake* senses and the 15 WordNet *shake* senses can all be partitioned into these five divisions (although WN1:**move** still gets mapped to more than one division), with idioms being listed separately (Palmer, Dang and Rosenzweig 2000).

The basic sense, Sense 1 (Levin Class 47.3), is the externally controlled *shaking* motion which results when a person or an earthquake or some other major force causes an object to move back and forth. This same motion can be amplified with directional information indicating a result such as *off, down, up, out* or *away* (Classes 26.5, 22.3). If a path prepositional phrase is specified, such as *shook the apples out of the tree* or *shook water from the umbrella*, then a change of location (CH-LOC) occurs, Sense 2 (Class 9.3). The same back and forth motion can occur during Body-Internal states such as *shaking from cold or fear*, i.e. TREMBLING, which gives us Sense 3 (Class 40.6). If a particular BODY-PART is shaken in a conventionalized gesture, such as *shaking hands, fists or fingers*, then a communicative act takes place, Sense 4 (Class 40.3.2). Finally non-physical usages are all classified as Sense 5 (Class 31.1, Psych verbs), such as *shaken by the news/the attack/his father's death.*

3.2 Criteria for WordNet sense grouping

The success we had in using these coarse-grained partitions to reconcile Hector and WordNet led us to re-examine the use of previous groupings in WordNet 1.6. One of the main differences between WordNet and a standard dictionary is the lack of an hierarchical organization for the distinct senses of an entry. They are all simply listed sequentially. WordNet, after all, supplies a wealth of inheritance information via hypernyms and synonym sets. However, these do not lend themselves readily to forming natural sense hierarchies, and have not been especially beneficial for automatic WSD systems (Lin 1998; Mihalcea and Moldovan 2001). The variation in hypernyms that occurs in most of the groups listed below provides evidence for why automatic grouping by hypernyms has not been more successful.

We decided to substantially revise and augment the existing WordNet 1.6 groupings for WordNet 1.7. In this section we discuss the rationale behind our new groupings and the methodology used to produce them. Coarse-grained sense distinctions are only slightly easier to define than fine-grained ones, and there are often cases where a sense appropriately belongs to more than one group. We chose the simplest possible style of grouping with no overlaps, acknowledging that this would sometimes be less than satisfactory.

WordNet 1.6 Groups WordNet typically has distinct entries for different syntactic forms of a verb. The result is that the syntactic alternations, such as causative/inchoative, that are grouped together by Levin into one class are often treated as distinct senses by WordNet. This design choice also forces sense distinctions for a given verb based on argument alternations (Levin 1993) to appear in different hypernym hierarchies (Fellbaum 1998a). So *He chilled the soup* is WN2, CAUSE TO CHANGE, and *The soup chilled* is WN3, UNDERGO A CHANGE. These approaches are not as contradictory as they seem. The Levin classes focus on the commonality of meaning that is preserved across different syntactic frames, allowing for subtle differences, while the WordNet senses capture the subtle shift in meaning occasioned by each different syntactic frame, without excluding possible commonalties.

WordNet 1.6 groupings were limited to linking together certain pairs of syntactic alternations, such as causative/inchoative. These links only affected 3.5% of the senses of our SENSEVAL-2 verbs, and had no impact on system performance or reconciliation of inter-annotator agreements.

WordNet 1.7 Groups We decided to do a much more comprehensive grouping of the SENSEVAL-2 senses, following the lead of the *shake* example, and attempting to provide specific criteria for the human groupers, as described below.[9] The groupings were made without reference to any corpus instances, although most of the groupers were also taggers. Each set of senses was grouped independently by two separate taggers. Discrepancies in the groupings were discussed and then adjudicated by a third tagger (Fellbaum *et al.* 2001). In contrast with hierarchical dictionary entries, this approach has a distinctly bottom-up, self-organizing flavor, and varies quite a bit from verb to verb. This is not really arbitrary, since different words require different criteria. A pure *change-of-state* verb like *break* will be especially sensitive to the type of object undergoing the change of state. Members of the same semantic class might cluster together quite naturally, as in *breaking vases/windows/glasses* vs. *breaking arms/legs/heads*. For activity verbs such as *shake* or *wipe*, the distinctions might become much more fine-grained, as in *shaking a fist* vs *shaking hands* (Hanks 1996), or less fine-grained *wiping one's face/the table*. The necessity of varying the criteria for sense distinctions on a word by word basis is also borne out by several machine learning experiments, where optimal performance is only achieved by allowing each word to choose its own parameter values (Veenstra *et al.* 2000; Yarowsky *et al.* 2001).

[9] Many of these criteria stemmed from a set of lectures given at Penn by Patrick Hanks in the fall of 2000.

Table 3. *Play senses, WordNet 1.7*

Sense No.	Description	Example	Hypernym
WN3	play (music) on an instrument	"The band played on..."	PERFORM
WN6	play a melody	"Play it again, Sam"	RECREATE
WN7	perform music on (a musical instrument)	"play the flute"	SOUND

Syntactic criteria Syntactic structure performed two distinct functions in our groupings. Recognizable alternations with similar subcategorization frames were often a factor in choosing to group senses together, as in the Levin classes, whereas distinct subcategorization frames were also often a factor in putting senses in separate groups as discussed below.

Syntax is often considered a mirror of the underlying semantics. Major differences in subcategorization frames for the same verb can reflect correspondingly major differences in meaning, e.g. *John left the room* (one object) vs. *Mary left her daughter-in-law her pearls in her will* (double object). When this is the case, applying a coarse syntactic filter to a verb's usages can be the simplest way of quickly capturing the underlying sense distinction. In other cases, where the subcategorization frames correspond to the types of alternations discussed above, the changes in meaning can be very slight. For example, in WordNet 1.7 *play* has the three separate senses given in Table 3, all of which can refer to the same event of playing music on a musical instrument.

These different senses are clearly related (in fact, 6 and 7 have the same syntax and differ only in the type of direct object), but these relations are not reflected in their hypernyms which emphasize the differences in what is being highlighted by each sense, rather than the similarities. Some lexicographers might argue that these slightly different usages should not be considered separate senses at all, but in the event that they are distinguished there can be little controversy in creating a group for them. The sense group also clearly corresponds to a broader, more underspecified sense which is not explicitly listed and which does not participate in any of the WordNet semantic relations. The groupings determined by this criteria had the most overlap with the previous groupings from WordNet 1.6. The complete grouping of *play* is given in Table 9.

We used Levin class membership as much as possible in assessing syntactic criteria, and while it was always useful, it was rarely as comprehensive as it had been for *shake*. The Levin classes were never intended to provide complete coverage of all senses, and they often only include one or two major senses of a word. An expanded version of these classes that included more senses could be very helpful. VerbNet, a computational lexicon based on the Levin classes is being developed, and the classes are currently being extended (Dang *et al.* 1998; Kipper *et al.* 2000; Dang *et al.* 2000; Kipper *et al.* 2004). Future research will be aimed at investigating the relevance of these expanded classes to sense distinctions.

Semantic criteria Clear semantic criteria for groupings are even more variable (Hanks 2000). Senses were grouped together if they were more specialized versions of a general sense. Our criteria for grouping senses separately included:

- differences in semantic classes of arguments (abstract versus concrete, animal versus human, animacy versus inanimacy, different instrument types...),
- differences in entailments (a change of state of an existing entity versus the creation of a new entity),
- differences in the type of event (abstract, concrete, mental, emotional...),
- whether there is a specialized subject domain, etc.

Note that many of our criteria, such as semantic class, subject domain and underlying predicate-argument structure, represent a shallow level of semantic representation that is becoming increasingly accessible to automatic processing techniques.

3.3 Applying the criteria to a specific verb

In the grouping of an actual verb entry, reference is made to both syntactic and semantic criteria, in whatever way is most suitable for that verb. WordNet often has separate entries for the intransitive form of a transitive verb, but that is not always the case when the transitive usage predominates. The same semantic criterion, such as the semantic class of the syntactic subject or object, is sometimes used to separate entries into two different sense groups and sometimes not. The more explicit we can be about the criteria we use for each verb for both grouping senses and distinguishing them, the more consistent we can be in tagging and in the categorizations of new usages, so we have attempted to label our sense groups with the criteria for distinguishing them. The sets of criteria are by necessity verb specific, and a slightly different set of criteria, applied in a unique order, is used for each verb, as illustrated below by our two example verbs. As discussed above, it's not surprising that different criteria would apply to different verbs. *Change-of-state* verbs like *break* or *melt* are much more affected by the type of object they are applied to than *action* verbs such as *sweep* or *wipe*. One of our goals for future work is looking for generalizations about sense distinctions that can be made with respect to classes of verbs rather than individual verbs.

Call The case of major subcategorization frame differences corresponding to clear sense distinctions is illustrated by *call*. Our final grouping of *call* includes several groups with distinctive predicate-argument structures, as in the sentential complements of Group 1: [1,3,19,22], *X call Y Z: ascribe an attribute Z to Y* (see Table 4). This is in contrast with most of the other *call* groups, which shared a straightforward transitive frame associated with a binary predicate-argument structure, as in Group 2: [2,3], *X called Y (on the phone)* or Group 3: [4,7,9], *X called a meeting/X was called for jury duty*. These groups are distinguished from each other semantically rather than syntactically. Group 2 involves a specific type of instrument and Group 3 is a summons to a particular type of activity (Fellbaum *et al.* 2001).

Table 4. *Call senses, Group I, WordNet 1.7*

Sense No.	Description	Example	Hypernym
WN1	name, call	"They named their son David"	LABEL
WN3	call, give a quality	"She called her children lazy and ungrateful"	LABEL
WN19	call, consider	"I would not call her beautiful"	SEE
WN22	address, call	"Call me Mister"	ADDRESS

Develop In contrast, syntactic structure did not play as primary a role in our grouping of *develop*, as shown in Table 5. Two entire groups (1 and 4) are separated from two other groups (2 and 3) simply on the basis of whether or not the *development* process had to be instigated by an outside causal agent. Group 4 is distinguished from Group 1 because it involves the improvement of an existing entity rather than the creation of a new entity. The outside agent usages are more likely to be transitive, whereas the internally controlled ones are more likely to be intransitive, but alternations do occur. These major distinctions affect 17 of the senses, while the remaining 5 are each associated with specialized subject domains, such as CHESS, FILM or MATH.

4 Impact of groupings on manual taggings

Our inter-annotator agreement figures, given in Table 7, were not as high as we had hoped, prompting a closer look. For several of the verbs we retagged a subset of 50 sentences distributed as evenly as possible among the different possible senses. This time the overall figures were higher, with agreement against the Gold Standard going up to over 90% using the groups. In general the inter-annotator agreement rose between 10% and 20% when measured against the grouped senses.[10] In this section we discuss one of the verbs in detail, *develop*, to illustrate the types of disagreements that occurred across the board. Measured against the Gold Standard, the fine-grained score on *develop* was 66% (33 correct tags) while the coarse-grained score rose to 90%. There are at least four different sources of annotator errors: *sense subsumption, missing or insufficient entries, vague usages,* and *world knowledge* (Fellbaum *et al.* 2001). The twelve *develop* tagger disagreements reconciled by the groups can be categorized into these four types, with 5 accounted for by sense subsumption or missing entries and 7 due to vague usages or inadequate world knowledge. These are discussed in more detail below, with reference to the effectiveness of the groupings in reconciling these differences.

Sense subsumption There were several disagreements on *develop* which stemmed from the choice between a more general or a more specific entry, well-known

[10] Agreement with the Gold Standard is generally higher than ITA, since only one set of errors is involved.

Table 5. *Develop senses, grouped, WordNet 1.7*

Group	Type	Sense No.	Description-Example	Hypernym
1-agent	new, abstract			
		WN1	products, or mental or artistic creations	CREATE
		WN 2	mental creations - "new theory of evolution"	CREATE
4-agent	improve item			
		WN6	resources - "natural resources"	IMPROVE
		WN7	ideas - "ideas in your thesis"	THEORIZE
		WN8	train animate beings - "violinists"	TEACH
		WN11	civilize - "countries are developing"	CHANGE
		WN12	make grow, ex. plants- "climate develops the grain"	CHANGE
		WN13	business, grow the market - "develop more customers"	GENERATE
		WN19	music, make more elaborate - "develop the melody"	COMPLICATE
2	new, property			
		WN3	personal attribute - "a passion for painting"	CHANGE
		WN4	physical characteristic - "beard"	CHANGE
3	new, self			
		WN5	originate - "new religious movement"	BECOME
		WN9	gradually unfold - "the plot developed slowly"	OCCUR
		WN10	grow - "a flower developed on the branch"	GROW
		WN14	mature - "The child developed beautifully… "	CHANGE
		WN20	happen - "Report the news as it develops"	OCCUR
5	chess			
		WN17	strengthening the position - "Spassky developed.."	PLAY
		WN18	a better position for a piece - "develop the rook"	PLAY
6	film			
		WN15	make visible - "develop the film"	CHANGE
7	math			
		WN16	3D mapped onto 2D, as in geometry	SUPERIMPOSE
		WN21	expand a series - "develop the function"	EXPAND

among lexicographers as "lumping" versus "splitting" (Fellbaum *et al.* 2002). Two easily confused *develop* senses involve the creation of new entities, characterized as either "products, or mental or artistic creations: CREATE (Sense 1)" or "a new theory of evolution: CREATE BY MENTAL ACT (Sense 2)." Three of the *develop* disagreements involved determining which of these two senses should be applied to phrases like *develop a better way to introduce crystallography techniques*. Either definition could fit; it's merely a question of determining among the taggers whether *ways* should be treated as things or theories. Since Sense 1 specifically mentions *mental creations* in addition to other types of creations, it can be seen as a more general definition which could *subsume* Sense 2. These more general senses, when present, provide the requisite flexibility for encompassing new usages.

In this case the discrepancies involved two different members of the same sense group, so the more coarse-grained evaluation reconciles them. Similar examples have been reported for *live* and *use* (Fellbaum, Palmer, Dang, Delfs and Wolf 2001). We have described the sense groups themselves as constituting a broader, more underspecified sense. When there is not an explicit entry in a group that is more general than the others it would be helpful to use the group itself as a tag.

Missing or Insufficient Dictionary Entries Other disagreements are introduced because the sense inventory against which a tagger is annotating may have gaps or redundancies; the glosses may also have ambiguous wordings or contradictory examples. Even if the annotator is working with an extremely clear, extensive entry, it may not cover novel or unusual usages, or domain-specific ones. For instance, WN 1.7 did not have a domain-specific sense for *develop* to handle the real-estate sense of *developing land.* The taggers agreed on the meaning of these verb tokens when they appeared, but used different strategies to stretch the pre-existing sense inventory to fit this usage, hesitating between Sense 6 and Sense 13, both in Group 4 (see Table 5.)[11] Two of the other *develop* disagreements involved deciding whether or not *understanding* as in *develop a much better understanding of* … constituted an attribute (Sense 3) or a physical characteristic (Sense 4), which was also finessed by the groupings (Group 2). In this case neither of the pre-existing senses is general enough to subsume the other.

Vague Contexts There are sentences where an author intentionally invokes a rich representation of a word that includes two or more related senses. For instance, *onion* (Senseval-1) typically has a food sense and a plant sense, and in a phrase such as *planting, harvesting and marketing onions* both are invoked (Krishnamurthy and Nicholls 2000). An instance of *play* (Senseval-2) said only, *he played superbly.* It was clear from the context that music was being played, but did the author intend to praise the playing of the instrument (Sense 3) or the melody (Sense 6) or both? Probably both. The grouping of these senses, as given in section 3.2, softens the penalty for failing to read the author's mind.

World knowledge Perhaps the most intractable tagging issues arise when the meaning of a word in a particular context depends not only on its syntactic use or the semantics of its arguments, but on world knowledge. For instance, the final seven of the *develop* disagreements all pertained to Group 4. Three of the sentences involved the development of *cancer tumors.* Do cancer tumors originate spontaneously, as

[11] Sometimes, in fact, one tagger would double-tag a particular instance while the second tagger chose a single sense that matched one of the two selected by the first annotator. This happened twice in the fifty *develop* sentences that were tagged, but they were not counted as disagreements.

would a religious movement (Sense 5), or are they more like a flower, a product of natural growth and evolution (Sense 10)? This choice involves a depth of medical knowledge which few doctors would claim, and in such a case tagging with a more coarse-grained sense that subsumes both 5 and 10 offers a more judicious option.

Discussion Differences in annotator choices often involve subtle semantic distinctions between senses where one sense might be slightly more specific or more applicable (in the case of a gap) than the other. Extremely high inter-annotator agreement with highly polysemous words is an unrealistic goal, given the inherent difficulty in attaining a consensus on word meaning and the changeable nature of language. Since a semantic grouping of senses with similar meanings puts the most easily confused senses in the same group, the annotator disagreements can often be reconciled by evaluating with the groups instead of the more fine-grained senses: Senseval-2 verb disagreements were reduced by more than a third, from 27% to 18%. Equally valuable is the opportunity to treat the group as a more underspecified sense in itself, for new usages that do not exactly fit a pre-existing sense. The ITA based on the grouped senses is 82%, much more in line with the Senseval-1 ITA, and leads to correspondingly higher system scores, again more in line with Senseval-1, as discussed in the next section. In a more recent experiment where the annotators actually tagged with the grouped senses, which had first been augmented with more explicit syntactic and semantic criteria for distinguishing them, the ITA for grouped senses rose to 86% (Palmer, Babko-Malaya and Dang 2004).

There are also reassuring parallels with even more coarse-grained distinctions based on subcategorization frames. In a separate project for the semantic annotation of predicate-argument structure, PropBank (Palmer, Gildea and Kingsbury 2005), very coarse-grained sense distinctions were made for the 700 most polysemous verbs in the Penn TreeBank (Kingsbury and Palmer 2002). These distinctions are based primarily on different subcategorization frames that require different argument label annotations, and all of the verbs were sense-tagged accordingly. Since the same data was tagged independently for both PropBank and Senseval-2, it has been possible to evaluate how well the Senseval-2 groupings correspond to the PropBank sense distinctions. They are surprisingly compatible; 95% of our groups map directly onto a single PropBank sense (Palmer *et al.* 2004), with a single PropBank sense typically corresponding to 2 or more groups.

The groupings have an intuitive appeal; a reader can readily appreciate the semantic coherence of the senses. However, if too much information is being lost by failing to make the more fine-grained distinctions, the groups will avail us little. We begin to address this question in Section 6, but first present the impact the groups had on our automatic system's performance.

5 Impact of groupings on automatic tagging

In this section we compare the performance of our automatic WSD system with both fine-grained senses and grouped senses. We analyze the tagging errors made by the system in the same way we analyzed the inter-annotator disagreements. Our system is based on a maximum entropy framework which combines linguistic contextual features from corpus instances of each verb to be tagged (Dang, Chia, Chiou and Palmer 2002). It performs comparably to the best performing systems in Senseval-1 and Senseval-2 (Dang and Palmer 2002), providing a reliable benchmark for comparing different data sets and sense inventories.

5.1 System description

Under the maximum entropy framework (Berger, Della Pietra and Della Pietra 1996), evidence from different features can be combined with no assumptions of feature independence. The automatic tagger estimates the conditional probability that a word has sense x given that it occurs in context y, where y is a conjunction of features. The estimated probability is derived from feature weights which are determined automatically from training data so as to produce a probability distribution that has maximum entropy, under the constraint that it is consistent with observed evidence.

In order to extract the linguistic features necessary for the model, all sentences were first automatically part-of-speech-tagged using a maximum entropy tagger (Ratnaparkhi 1998) and parsed using the Collins parser (Collins 1997). In addition, an automatic named entity tagger (Bikel, Miller, Schwartz and Weischedel 1997) was run on the sentences to map proper nouns to a small set of semantic classes. Following work by Chodorow, Leacock and Miller (Chodorow *et al.* 2000), we divided the possible model features into topical and local contextual features. Topical features looked for the presence of keywords occurring *anywhere* in the sentence and any surrounding sentences provided as context (usually one or two sentences). The set of 200–300 keywords is specific to each lemma to be disambiguated, and is determined automatically from training data so as to minimize the entropy of the probability of the senses conditioned on the keyword.

The local features for a verb w in a particular sentence, given in Table 6, tend to look only within the smallest clause containing w. This set of local features relies on access to syntactic structure as well as semantic class information, and represents our move towards using richer syntactic and semantic knowledge sources to model human performance. When we incorporated WordNet semantic class information, the noun complements were not disambiguated in any way, and all possible synsets and hypernyms for the noun were included. No separate disambiguation of noun complements was done because, given enough data, the maximum entropy model should assign high weights to the correct semantic classes of the correct noun sense if they represent defining selectional restrictions.

5.2 Sources of automatic tagger disagreements

In this section we describe the system performance on the verbs from Senseval-2. We compare the coarse-grained scores using our new groupings versus random groupings and the previous WordNet 1.6 groupings, and substantiate the greater coherence of the new groupings. In addition, we experiment with training the system

Table 6. *Local features used by the Maxent WSD system*

lexical and part-of-speech unigrams, bigrams, and trigrams of words within a window of 2 words to the left and right of *w*
whether or not the sentence is passive
whether there is a sentential complement, subject, direct object, or indirect object
the words (if any) in the syntactic positions above and particle, prepositional complement (and its object)
a Named Entity tag (PERSON, ORG, LOC) for proper nouns appearing in the positions above
WordNet synsets and hypernyms for the nouns appearing in the positions above

on the coarse-grained senses, to see if the larger amounts of training data per sense result in higher performance.

We tested the WSD system on the verbs from the English lexical sample task for Senseval-2. The annotation for multi-word constructions made them straightforward to identify and consequently made it much easier for our system to incorporate information about the satellites, typically verb-particles, without having to look at the dictionary (whose format may vary from one task to another). All the best-performing systems (including our own) on the English verb lexical sample task filtered out possible senses based on the marked satellites, and this improved performance significantly. For example, the particle in *call on* was marked by the annotation as *call* having an *on* satellite, distinguishing it from *call* by itself.

Our system achieved 62.5% and 71.7% accuracy using fine-grained and coarse-grained scoring, respectively. This is in comparison to the next best-performing system, which had fine- and coarse-grained scores of 57.6% and 67.2% (Palmer *et al.* 2001). Here we see the benefit of including a filter that only considered phrasal senses whenever there were satellites of multi-word constructions marked in the test data; had we not included this filter, our fine- and coarse-grained scores would have been only 60.0% and 69.1%.

Table 7 shows a breakdown of the number of senses and groups for each verb, as well as human inter-tagger agreement on fine-grained (ITA-fine) and coarse-grained (ITA-coarse) senses.[12] Overall, coarse-grained evaluation using the groups improved the system's score by about 10%. This is consistent with the improvement we found

[12] We do not include kappa figures because the standard formulation of kappa doesn't address our situation where multiple tags are allowed for each instance. Although there were relatively few multiply tagged instances in the Gold Standard, 84 out of over 5000 instances, in the raw human annotator data there are substantially more. We also find that inter-tagger agreement is sufficient for the comparisons that we wish to make between system and human performance, and between Senseval-1 and Senseval-2.

Table 7. *Number of WN 1.7 (corpus) senses for each verb, not including multi-word expressions; number of WN 1.7 (corpus) groups for each verb, not including multi-word expressions; inter-tagger agreement for fine-grained senses and sense groups; accuracy of maximum entropy system under fine- and coarse-grained scoring; accuracy of system trained using groups and scored with coarse-grained method. *No inter-tagger agreement figures were available for "play" and "work"*

Verb	WN (cor) Sen	WN (cor) Grp	ITA-f	ITA-c	MX-f	MX-c	MX-g
begin	10 (9)	10 (9)	0.812	0.814	0.893	0.893	0.893
call	28 (14)	11 (7)	0.693	0.892	0.545	0.697	0.697
carry	39 (22)	16 (11)	0.607	0.753	0.394	0.485	0.515
collaborate	2 (2)	2 (2)	0.750	0.750	0.900	0.900	0.900
develop	21 (16)	9 (6)	0.678	0.852	0.580	0.725	0.739
draw	35 (21)	15 (9)	0.767	0.825	0.317	0.463	0.439
dress	15 (8)	7 (4)	0.865	1.000	0.729	0.915	0.898
drift	10 (7)	6 (4)	0.500	0.500	0.406	0.406	0.469
drive	21 (9)	8 (4)	0.588	0.717	0.595	0.833	0.810
face	9 (7)	4 (4)	0.786	0.974	0.839	0.903	0.903
ferret	3 (0)	3 (0)	1.000	1.000	1.000	1.000	1.000
find	16 (15)	8 (7)	0.443	0.569	0.368	0.500	0.441
keep	22 (15)	11 (10)	0.791	0.801	0.612	0.627	0.597
leave	14 (12)	7 (7)	0.672	0.805	0.606	0.636	0.591
live	7 (6)	4 (4)	0.797	0.872	0.701	0.731	0.731
match	9 (8)	5 (4)	0.565	0.826	0.500	0.714	0.690
play	35 (21)	17 (12)	*	*	0.530	0.530	0.591
pull	18 (10)	10 (5)	0.681	0.722	0.500	0.667	0.683
replace	4 (4)	2 (2)	0.659	1.000	0.600	0.933	0.933
see	24 (19)	12 (10)	0.709	0.755	0.391	0.464	0.507
serve	15 (12)	7 (6)	0.908	0.932	0.745	0.824	0.843
strike	20 (16)	12 (10)	0.762	0.905	0.389	0.519	0.444
train	11 (9)	6 (4)	0.288	0.550	0.635	0.730	0.714
treat	8 (6)	6 (5)	0.969	0.975	0.500	0.614	0.591
turn	26 (16)	14 (9)	0.742	0.894	0.493	0.627	0.612
use	6 (6)	3 (3)	0.743	0.894	0.711	0.842	0.829
wander	5 (5)	3 (3)	0.650	0.900	0.800	0.900	0.900
wash	12 (6)	6 (3)	0.875	0.906	0.667	0.667	0.750
work	27 (13)	10 (7)	*	*	0.450	0.583	0.633
Total	16.28 (10.83)	8.07 (5.90)	0.713	0.820	0.625	0.717	0.715

in inter-tagger agreement for groups over fine-grained senses (82% instead of 71%). In addition to the fine- (MX-f) and coarse-grained (MX-c) scores for our system, we report the coarse-grained score (MX-g) for a variant of the system that was trained on sense groups instead of the fine-grained senses. Because the training corpus was so small, we expected that training on groups would mitigate the sparse data problem.

Call We found that the grouped senses for *call* improved performance over evaluating with respect to fine-grained senses; the system achieved 67.9% accuracy with coarse-grained scoring using the groups, as compared to 54.5% accuracy with fine-grained scoring. When evaluated against the fine-grained senses, the system got 30 instances wrong, but 10 of the "incorrect" instances were tagged with senses that were actually in the same group as the correct sense. Almost all the confusion (9 instances) involved senses from Group 1 (see Table 4). This group of senses differs from others in the ability to take a sentential complement, which is explicitly modeled as a feature in our system. For each of these senses, the maximum entropy model assigned very high weights to the feature "has sentential complement." Here we see that the system benefits from using syntactic features that are linguistically richer than the features that have been used in the past. Furthermore, we would not expect to be able to differentiate Senses 3 and 19 from the other two senses in the group without looking deeper into the structure of the sentential complement to identify whether the small clause has a predicative noun or adjective.

Develop The system's performance on *develop* also improved significantly using the groups, confirming the apparent close correspondence between ITA and system performance. Eight of the 29 errors made by the tagger were due to confusing Sense 1 and Sense 2 of *develop*, which are in the same group. Instances of Sense 1 that were tagged as Sense 2 by the system included: ... *they have developed a genetic engineering technique for creating hybrid plants ... ; William Gates and Paul Allen in 1975 developed an early language-housekeeper system for PCs.* Conversely, the following instances of Sense 2 were tagged as Sense 1 by the tagger: *A ... team ... hopes to develop ways to magnetically induce cardiac muscle contractions; Kobe Steel Ltd. adopted Soviet casting technology ... until it developed its own system.* Based on the direct object of *develop*, the automatic tagger was hard-pressed to differentiate between developing a *technique* or *system* (Sense 1) and developing a *way* or *system* (Sense 2). These instances that were difficult for the automatic WSD system, are also difficult for human annotators to differentiate consistently, as discussed in Section 4.

Training on Groups Training on groups did not significantly change the overall coarse-grained scores of our system. It increased the coarse-grained score by at least 5% for some verbs (*drift, play, wash, work*) but decreased the score for others (*find, strike*). Defining more precisely the cases when groups do and do not mitigate the sparse data problem will be a subject for future investigation.

Comparing other groupings As a base-line, to ensure that the improvement did not come simply from the lower number of tag choices for each verb, we created random groupings in which each verb had the same number of groups, but with the senses distributed randomly. We found that these random groups provided almost no benefit to the inter-annotator agreement figures (74% instead of 71%), confirming the greater coherence of the manual groupings. The WordNet 1.6 groups reduced the polysemy of the same verbs from 14[13] to 13.5, and had even less effect on performance.

[13] some senses were added for WN 1.7.

Table 8. *Portuguese, German, Chinese and Korean translations of develop*

Groups	Senses (obj)	Portuguese	German	Chinese	Korean
G4	WN13 markets	desenvolver	entwicklen	kai1-fa1	hyengsengha-ta
G1	WN1 products WN2 ways	desenvolver	entwickeln	kai1-fa1	kaypalha-ta
G1	WN2 theory	desenvolver	entwickeln	fa1-zhan3	palcensikhi-ta
G2	WN3 understanding	desenvolver	entwickeln	pei2-yang3-chu1	palcensikhi-ta
G2	WN3 character	desenvolver	bilden	pei2-yang3	yangsengha-ta
G4	WN8 musicians	desenvolver	ausbilden	pei2-yang3	yangsengha-ta
G3	WN10 bacteria	desenvolver-se	sich bilden	fa1-yu4	paltalha-ta
G3	WN5 movements	desenvolver-se	bilden	xing2-cheng2	hyengsengtoy-ta

6 Suitability of groups for machine translation

We have demonstrated that semantically coherent groupings of senses can reconcile subtle disagreements between annotators. However, this is only a benefit if these subtle distinctions have little or no impact in NLP applications. To explore this issue we examined the translations of several *develop* sentences into Portuguese, German, Chinese and Korean, as summarized in Table 8. This is in the same spirit as recent work that examines cross-language variation in how senses are translated as a means of automatically determining sense distinctions (Resnik and Yarowsky 1999; Ide 2000). The examples here can be seen as the converse of that approach, in that our goal is determining how effective the given sense distinctions are in predicting translation differences. The translations were created by native speakers of each language who were given example sentences from the sense definitions and asked to translate just the verb in question into the most suitable expression in their language.

As might have been expected, the Portuguese translations are the most similar to English, and with the exception of an explicit reflexive marking for our "internally caused" senses, they are all the same (see Table 8). The grouped senses would certainly be sufficient for translation into Portuguese; in fact they are unnecessarily fine-grained. For German there are only two main choices, *entwickeln* (Groups 1, 2, 4) or a variant of *bilden (bilden, ausbilden, sich bilden)* (Groups 3, 2, 4) but the choices do not match the group boundaries of Groups 2 and 4 very precisely. It is even less clear that the groups would be helpful for translating into Korean or Chinese, where there are almost as many translation possibilities as there are

Table 9. *30 of 35 Play senses, grouped, WordNet 1.7*

Group	Type	Sense No.	Description-Example	Hypernym
1-trans	sports	WN1	games – “we played hockey”	COMPETE
		WN34	an opponent – “Princeton played Yale”	MEET
2-trans	with LOC-PP	WN15	put a piece into play	DEPLOY
		WN31	play a good backhand	HIT
		WN32	play a piece	USE
		WN33	play a player	USE
3-trans	have an effect	WN2	contribute to – “play a part in his decision”	ACT
		WN17	behave as “play it safe, play fair”	ACT
4-trans	music	WN3	play (music) on an instrument	PERFORM
		WN6	play a melody	RE-CREATE
		WN7	perform music on (a musical instrument)	SOUND
5-trans	theatre	WN4	play a role – “Gielgud played Hamlet.”	RE-CREATE
		WN14	location – “play Carnegie Hall”	PERFORM
		WN25	perform – He played in “Julius Caesar”	PERFORM
		WN26	location – a show is playing at Carnegie	?
6-trans	musical device	WN13	mechanical device – “the tape was playing”	SOUND
		WN18	causative/animate – “please play my record”	?
7-trans	pretend	WN12	pretend – “let’s play Indians”	SIMULATE
		WN8	play deaf	ACT
8-trans	abstract	WN16	play the stockmarket	ACT
		WN20	play on someone’s emotions	EXPLOIT
		WN21	play with the thought of	CONSIDER
		WN23	flirt	ACT
		WN19	fiddle with – play with the idea of	MANIPULATE
9-trans	bet	WN29	make bets – “He plays the casinos in Trouville”	BET
		WN30	wager – “I bet $100 on that new horse”	GAMBLE
10-ditrans	bet	WN10	wager – “He played $20 on a horse”	GAMBLE
11-intrans	motion	WN24	move freely – this wheel has a lot of play in it	MOVE
12-intrans	recreation	WN5	be at play, “the kids played all day”	ACT
		WN11	recreate – “on weekends I play”	?

WordNet senses (Ng *et al.* 2003). We have similar translation results for *hit, learn, live,* and *call.*

There are interesting common themes between all five languages, such as considering properties of the objects (“abstract” vs “animate”) and whether or not an external agent is required, (i.e., causative/inchoative). Some languages group

WN3 and WN8 together based on being applied to a "human" or a "human characteristic."[14]

In sum, when mapping from English to one other language, our preliminary investigation indicates that the groups can provide a substantial benefit in lessening the amount of sense distinctions that must be made. However, if the goal is translation into several languages eventually every distinction that can be made will be made, and there are few useful generalizations. This does not mean that the groups become completely ineffective. Even if they do not provide a single translation, providing a small set of translations that are semantically related can still be beneficial. It is often possible for the target language to choose from such a set based on target language selectional restrictions (Palmer *et al.* 2000; Palmer and Wu 1995).

7 Conclusion

This paper has discussed the data preparation and system performance for the English verb Lexical Sample Task of the Senseval-2 Word Sense Disambiguation exercise, and demonstrated that WordNet 1.7 provided a useful sense inventory for training automatic systems. The manual groupings of the WordNet 1.7 verb senses and the methodology for producing them were presented. These provide a more coarse-grained view of the sense distinctions which plays an essential role in evaluation as well as applications. In examining the instances that proved troublesome to both the human taggers and the automatic system, we found several categories of errors that were tied to subtle sense distinctions which were reconciled by backing off to independently derived coarse-grained sense groups. These categories include different perspectives on sense subsumption, insufficient sense entries, vague contexts or inadequate world knowledge. The annotators also reported that the groupings allowed them to study side-by-side the senses that are most likely to be confused, improving tagging consistency. For the automatic systems, a potential advantage of the groupings is that, because the members are so closely related, their combination can contribute towards providing a critical mass of coherent examples, clarifying the differences between two groups and in some instances alleviating the sparse data problem. These groupings have been made available to the community via the Senseval2 web site, as well as a first pass at groupings for all WordNet 1.7 verbs with more than three senses. An NSF-funded project is currently underway which will provide grouped entries for over 1500 WordNet verbs and corresponding Penn Treebank tagged data. The associated research will include attempts to create the sense groupings in a more principled fashion, looking for commonalities across classes of verbs, as well as improvements to the automatic WSD systems. Access to larger amounts of tagged data for more lexical items will allow us to explore unsupervised clustering, perhaps following Bill

[14] The persistence in treating WN13 similarly to WN1 and WN2 prompts us to consider moving WN13 from Group4 to Group 1.

Dolan's lead (Dolan 1994), and bearing in mind that automatic techniques do not have access to all the relevant information.

Lexicographers have long recognized that many natural occurrences of polysemous words are embedded in underspecified contexts and could correspond to more than one specific sense. There will also always be gaps in inventories and available world knowledge. In such cases both manual and automatic tagging discrepancies are inevitable. Annotators and automatic systems need the option of selecting, as an alternative to an explicit sense, either a group of specific senses or a single, broader sense, where specific meaning nuances are subsumed (Palmer 1990). Although present to some degree in the hierarchical entries of traditional dictionaries, these have previously played only a small role in WordNet. The verb groupings presented here represent an important step in the direction of making WordNet more effective in computational applications such as cross-lingual information retrieval (Lee *et al.* 2004) and machine translation. These sense groupings can be guided and enhanced by the analysis of inter-annotator disagreements and the development of more principled sense distinction criteria. Comparing the annotations of an automatic system with that of a human tagger provides further insights into the nature of the relationships between senses from a computational perspective, and brings into focus the similarity in criteria being used by both humans and machines. There is no claim that the groupings presented here are correct in any absolute sense, but simply that they are of practical use. There could be alternative groupings that might prove equally useful, or even more useful, for particular types of applications.

Acknowledgments

This work has been supported by National Science Foundation Grants NSF-9800658, VerbNet, NSF-9910603, ISLE, NSF-0415923, WED, the DTO-AQUAINT NBCHC040036 grant under the University of Illinois subcontract to University of Pennsylvania 2003-07911-01 and DARPA grant N66001-00-1-8915 at the University of Pennsylvania, as well as the CIA (contract number 2000*SO53100*000). We would also like to thank Scott Cotton for system infrastructure support, Joseph Rosenzweig for building the annotation tool, and Lauren Delfs and Susanne Wolff for the manual annotation for Senseval-2.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, DARPA, the CIA or the DTO.

References

Asprejan, J. D. (1974) Regular polysemy. *Linguistics*, **142**: 5–32.

Atkins, B. T. S. and Levin, B. (1991) Admitting impediments. In: U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 233–262. Lawrence Erlbaum.

Berger, A. L., Della Pietra, S. A. and Della Pietra, V. J. (1996) A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1).

Bikel, D. M., Miller, S., Schwartz, R. and Weischedel, R. (1997) Nymble: A high-performance learning name-finder. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington, DC.

Calzolari, N. and Corazzari, H. (2000) Romanseval: Framework and results for italian. *Computers and the Humanities*, **34**(1–2).

Chodorow, M., Leacock, C. and Miller, G. A. (2000) A topical/local classifier for word sense identification. *Computers and the Humanities*, **34**(1–2).

Collins, M. (1997) Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.

Cruse, D. A. (1986) *Lexical Semantics*. Cambridge University Press.

Dang, H. T. (2004) *Investigations into the Role of Lexical Semantics in Word Sense Disambiguation*. PhD thesis, University of Pennsylvania.

Dang, H. T. and Palmer, M. (2002) Combining contextual features for word sense disambiguation. *SIGLEX Workshop on Word Sense Disambiguation, in conjunction with the 40th Meeting of the Association for Computational Linguistics, (ACL-02)*, Philadelphia, PA.

Dang, H. T., Chia, C., Chiou, F. and Palmer, M. (2002) Simple features for chinese word sense disambiguation. *Proceedings of the 19th International Conference on Computational Linguistics, COLING-02*, Taipei, Taiwan.

Dang, H. T., Kipper, K. and Palmer, M. (2000) Integrating compositional semantics into a verb lexicon. *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Germany.

Dang, H. T., Kipper, K., Palmer, M. and Rosenzweig, J. (1998) Investigating regular sense extensions based on intersective levin classes. *Proceedings of Coling-ACL98*, Montreal, Canada.

Dolan, W. B. (1994) Word sense ambiguation: clustering related senses. *Proceedings of the 15th International Conference on Computational Linguistics, COLING'94*, Kyoto, Japan.

Edmonds, P. and Cotton, S. (2001) SENSEVAL-2: Overview. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France.

Fellbaum, C. (1998a) The organization of verbs and verb concepts in a semantic net. In: P. Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*. Dordrecht.

Fellbaum, C. (editor) (1998b) *Wordnet: An Electronic Lexical Database*. MIT Press.

Fellbaum, C., Grabowski, J. and Landes, S. (1997) Analysis of a hand-tagging task. *Proceedings of the ACL/Siglex Workshop*, Somerset, NJ.

Fellbaum, C., Grabowski, J. and Landes, S. (1998) Performance and confidence in a semantic annotation task. In: C. Fellbaum, editor, *WordNet*. MIT Press.

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

Fellbaum, C., Delfs, L., Wolf, S. and Palmer, M. (2005) Word meaning in dictionaries, corpora, and the speaker's mind. In: Barnbrook, G., Danielsson, P. and Mahlberg, M. (Eds.), *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Birmingham University Press.

Geeraerts, D. (19993) Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, **4**(3): 223–272.

Hanks, P. (1996) Contextual dependency and lexical sets. *Int. J. Corpus Linguistics*, **1**(1).

Hanks, P. (2000) Do word meanings exist? *Computers and the Humanities*, **34**(1–2): 171–177.

Ide, N. (2000) Cross-lingual sense determination: Can it work? *Computers and the Humanities*, **34**(1–2): 223–234.

Ide, N. and Véronis, I. (1998) Introduction to the special issue on word sense disambiguation. *Computational Linguistics J.* **24**(1).

Jones, K. S. (1986) *Synonymy and Semantic Classification*. Edinburgh University Press.

Kilgarriff, K. (1997) I don't believe in word senses. *Computers and the Humanities*, **31**(2).

Kilgarriff, A. (2001) English lexical sample task description. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France.

Kilgarriff, A. and Palmer, M. (2000) Introduction to the special issue on senseval. *Computers and the Humanities*, **34**(1–2): 1–13.

Kilgarriff, A. and Rosenzweig, J. (2000) Framework and results for English SENSEVAL. *Computers and the Humanities*, **34**(1–2).

Kingsbury, P. and Palmer, M. (2002) From TreeBank to PropBank, *Third International Conference on Language Resources and Evaluation, LREC-02*, Las Palmas, Canary Islands, Spain, May 28–June 3, 2002.

Kipper, K., Dang, H. T. and Palmer, M. (2000) Class-based construction of a verb lexicon. *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX.

Kipper, K., Snyder, B. and Palmer, M. (2004) Using prepositions to extend a verb lexicon. *Computational Lexical Semantics Workshop, held in conjunction with HLT/NAACL-04*, Boston, MA.

Krishnamurthy, R. and Nicholls, D. (2000) Peeling an onion: the lexicographer's experience of manual sense-tagging. *Computers and the Humanities*, **34**(1–2).

Lee, J.-H., Kang, I.-S. and Na, S.-H. (2004) Influence of disambiguation on cross-language information retrieval. *In the Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Hainan Island, China.

Lesk, M. (1986) Lexical disambiguation using simulated annealing. *Proceedings 1986 SIGDOC Conference.*

Levin, B. (1993) *English Verb Classes and Alternations A Preliminary Investigation*, University of Chicago Press, Chicago, IL.

Lin, D. (1998) Automatic retrieval and clustering of similar words. *Proceedings of Coling-ACL98*, Montreal, Canada.

Marcus, M., Santorini, B. and Marcinkiewicz, M. A. (1993) Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, **19**(2).

Mihalcea, R. and Moldovan, D. I. (2001) Automatic generation of a coarse grained wordnet. *Proceedings of NAACL 2001 Workshop on WordNet and Other Lexical Resources.*

Mihalcea, T. C. R. and Kilgarriff, A. (2004) The senseval-3 english lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, held in conjunction with ACL-04*, pages 25–28, Barcelona, Spain, July 2004.

Miller, G. A. and Fellbaum, C. (1991) Semantic networks of english. *Lexical and Conceptual Semantics, Cognition Special Issue*, pp. 197–229.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. (1990) Five papers on wordnet. Technical Report 43, Cognitive Science Laboratory, Princeton University.

Ng, H. T., Wang, B. and Chan, Y. S. (2003) Exploiting parallel texts for word sense disambiguation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Saporo, Japan.

Palmer, M. (1990) Customizing verb definitions for specific semantic domains. *Machine Translation*, **5**.

Palmer, M. and Light, M. (1999) Introduction to the special issue on semantic tagging. *Natural Language Engineering*, **5**(2).

Palmer, M. and Wu, Z. (1995) Verb semantics for english-chinese translation. *Machine Translation*, **10**: 59–92.

Palmer, M., Dang, H. T. and Rosenzweig, J. (2000) Sense Tagging the Penn Treebank. *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Palmer, M., Han, C., Xia, F., Egedi, D. and Rosenzweig, J. (2000) Constraining lexical selection across languages using tags. In: A. Abeille and O. Rambow, editors, *Tree Adjoining Grammars: Formal, computational and linguistic aspects.* CSLI.

Palmer, M., Fellbaum, C., Cotton, S., Delfs, L. and Dang, H. T. (2001) English tasks: All-words and verb lexical sample. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France.

Palmer, M., Babko-Malaya, O. and Dang, H. T. (2004) Different sense granularities for different applications. *Proceedings of the 2nd Workshop on Scalable Natural Language Understanding Systems, at HLT/NAACL-04*, Boston, MA.

Palmer, M., Gildea, D. and Kingsbury, P. (2005) The proposition bank: An annotioted corpus semantic roles. *Computational Linguistics Journal* **31**: 1.

Pustejovsky, J. (1991) The generative lexicon. *Computational Linguistics*, **17**(4).

Ratnaparkhi, A. (1998) *Maximum Entropy Models for Natural Language Ambiguity Resolution.* PhD thesis, University of Pennsylvania.

Resnik, P. and Yarowsky, D. (1991) Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, **5**(2): 113–133.

Senseval-3 (2004) The third international workshop on the evaluation of systems for the semantic analysis of text. *In the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Talmy, L. (19991) Path to realization: A typology of event conflation. *Proceedings of the Berkeley Linguistics Society Parasession*, Berkeley, CA.

Véronis, J. and Segonde, F. (2000) Romanseval: Framework and results for French. *Computers and the Humanities*, **34**(1–2).

Veenstra, J., van den Bosch, A., Buchholz, S., Daelemans, W. and Zavrel, J. (2000) Memory-based word sense disambiguation. *Computers and the Humanities*, **34**(1–2): 171–177.

Yarowsky, D. and Florian, R. (2002) Evaluating sense disambiguation performance across diverse parameter spaces. *Natural Language Engineering*, **8**(4): 293–310.

Yarowsky, D., Florian, R., Cucerzan, S. and Schafer, C. (2001) The Johns Hopkins Senseval-2 system description. *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, in conjunction with the 39th Meeting of the Association for Computational Linguistics, (ACL-02)*, Toulouse, France.