

Disambiguating Coordinations Using Word Distribution Information

Francis Chantree¹ Adam Kilgarriff² Anne de Roeck¹ Alistair Willis¹

¹The Open University, Milton Keynes, U.K.

²Lexical Computing Ltd, Brighton, U.K.

¹{F.J.Chantree,A.DeRoeck,A.G.Willis}@open.ac.uk; ²adam@lexmasterclass.com

Abstract

In this paper we present some heuristics for resolving coordination ambiguities. This type of ambiguity is one of the most pervasive and challenging. We test the hypothesis that the most likely reading of a coordination can be predicted using word distribution information from a generic corpus. The measures that we use are: the relative frequency of the coordination in the corpus, the distributional similarity of the coordinated words, and the collocation frequency between the coordinated words and their modifiers. The heuristics that we present based on these measures have varying but useful predictive power. They also take into account our view that many ambiguities cannot be effectively disambiguated, since human perceptions vary widely.

1 Introduction

Coordination ambiguity is a structural (i.e. syntactic) ambiguity. Compared with other structural ambiguities, e.g. prepositional phrase (PP) attachment ambiguity, it has received little attention in the literature. This is despite the fact that coordinations are known to be a “pernicious source of structural ambiguity in English” (Resnik 99). Our work is novel in that we use several types of word distribution information to disambiguate coordinations of any type of word, and in that we acknowledge that some ambiguities are too ambiguous to be judged reliably. This latter point is an important consideration, as providing readings for such ambiguities would be misleading and potentially dangerous.

We test the hypothesis that the preferred reading of a coordination can be predicted using word distribution information from a generic corpus. To do this we present three heuristics. These use the relative frequency of the coordination in the corpus, the distributional similarity of the coordinated words, and the collocation frequency between the coordinated words and a modifier. All the heuristics use information generated by the Sketch Engine¹ (Kilgarriff *et al.* 04) operating on

the British National Corpus² (BNC).

The examples that we investigate contain a single coordination which incorporates two phrases and a modifier, such as in the phrase:

old boots and shoes,

(where *old* is the modifier). Applying our heuristics to this phrase, we find firstly that *boots and shoes* appears relatively often in the corpus. Secondly, *boots* and *shoes* are shown to have strong distributional similarity, suggesting that *boots and shoes* is a syntactic unit. Both these factors suggest that coordination takes place before the modifier *old* takes scope. Thirdly, the collocation frequency of *old* and *boots* is not significantly greater than that of *old* and *shoes*, suggesting that it is not likely that only *boots* is modified by *old*. All three heuristics agree therefore that coordination takes place before the modifier takes scope.

We test our hypothesis for text drawn from requirements engineering. This is a very suitable domain as ambiguity is recognised as being a serious and potentially costly problem (Gause & Weinberg 89). For instance, a system might be built incorrectly due to a requirement being read in a way that was unintended.

We have built and tagged a corpus of requirements specification documents, from which we extract a collection of sentences and phrases containing coordination ambiguities. We identify preferred readings for each of these by means of ambiguity surveys, in which we obtain human judgements on each example. This forms our evaluation dataset. We then apply our heuristics to the dataset to see if they can automatically replicate the consensus human judgements.

In this paper we first discuss the coordination ambiguity problem and research related to our own, and then outline how we create our evaluation dataset. We then describe our empirical research, beginning with methodology that is

¹<http://www.sketchengine.co.uk>

²<http://natcorp.ox.ac.uk>

| Researchers | Recall (%) | Baseline Precision (%) | Precision (%) | Precision % points above baseline (%) | F-Measure $\beta = 0.25$ (%) | F-Measure % points above baseline (%) |
|--------------------------|------------|------------------------|---------------|---------------------------------------|------------------------------|---------------------------------------|
| (Agarwal & Boggess 92) | n/a | n/a | 82.3 | n/a | n/a | n/a |
| (Goldberg 99) | n/a | 64 | 72 | 18 | n/a | n/a |
| (Resnik 99) (unweighted) | 66.0 | 66.0 | 71.2 | 5.2 | 70.9 | 3.5 |
| (Resnik 99) (weighted) | 69.7 | 44.9 | 77.4 | 32.5 | 76.9 | 30.5 |

Table 1: Performances of other researchers

generic to all our heuristics, followed by a description of each heuristic, and ending with an evaluation of our results. Lastly, we offer our conclusions and present some ideas for future work.

2 Coordination Ambiguity

Coordination ambiguity can occur whenever coordinating conjunctions are used, and it is a pervasive problem in English as coordinating conjunctions are common. Together, *and* and *or* account for approximately 3% of the words in the BNC, and they account for the great majority of coordinating conjunctions. We confine our investigations to *and*, *or* and *and/or*. Words and phrases of all types can be coordinated (Okumura & Muraki 94). The external modifier can also be a word or phrase of almost any type, and it can appear before or after the coordination.

In an example from our dataset:

Assumptions and dependencies that are of importance

the external modifier *that are of importance* applies either to both the *assumptions* and the *dependencies* or to just the *dependencies*. Because of the order in which the words are connected, we refer to the former case as *coordination-first*, and to the latter as *coordination-last*³. We concentrate on coordinations of this type where two syntactic readings are possible.

3 Related Research

There has not been a large amount of research on coordination ambiguity in English in the NLP community, and what has been carried out has been quite diverse. The results of the researchers discussed below are summarised in Table 1.

³Other terminology can be used, e.g. *low attachment* and *high attachment*, depending on where the coordinated phrase furthest from the modifier attaches in the parse tree (Goldberg 99).

Agarwal and Boggess present an algorithm that attempts to identify which phrases are coordinated by coordinating conjunctions (Agarwal & Boggess 92). Using the machine-readable Merck Veterinary Manual as their dataset, they achieve an accuracy of 81.6% for the conjunctions *and* and *or*. Their method matches parts of speech and case labels of the head words of the coordinated phrases. Pre-conjunction phrases are popped off a stack until a match with the post-conjunction phrase is found. Their method is a straightforward and potentially useful way of matching candidate coordinated phrases, but it does not deal adequately with ambiguity arising from modifier attachment.

Goldberg uses unsupervised learning to determine the attachment of noun phrases in ambiguous coordinations (Goldberg 99). She simplifies the text using a chunker, and then extracts the headwords of the coordinated phrases. Her test data, which is unannotated, includes a lot of noise. Also, as her method is a simple re-implementation of a PP-attachment method (Ratnaparkhi 98), it does not model information, such as word similarity, that is useful for coordination disambiguation. Goldberg’s system correctly predicts with an accuracy of 72% the annotated attachments of her development set drawn from the Wall Street Journal.

Using an unweighted heuristic, Resnik investigates the role of semantic similarity in resolving coordination ambiguities involving nominal compounds of the form *noun1 and noun2 noun3* (Resnik 99). (Note that this is not the same as the distributional similarity which we use.) He looks up the nouns in WordNet and determines which of the classes that subsume them both has the highest information content. Without any back-off strategy, this procedure results in 71.2% precision and 66.0% recall of the correct human dis-

ambiguations in his dataset drawn from the Wall Street Journal.

Using a weighted heuristic, Resnik adds an evaluation of the selectional association between the nouns to his semantic similarity evaluation (Resnik 99). He also restricts his dataset to coordinations of the form *noun0 noun1 and noun2 noun3*. Improved precision of 77.4% and 69.7% recall is achieved. We believe that this heuristic’s high performance is in no small part due to the highly specific dataset being used, allowing for more measurements of similarity to be factored in. The results are interesting, but we feel that a useful disambiguation heuristic should be able to cope with less constrained data.

Some research on disambiguating uncoordinated noun compounds using corpus information bears similarity to our own. Lauer provides a synopsis of some approaches to the binary decision problem of disambiguating the bracketings in compounds of the form *noun1 noun2 noun3* (Lauer 95). He reports a maximum accuracy of 81%, above a baseline of 67%, using a hand-disambiguated dataset drawn from a popular encyclopedia. Lauer shares our opinion that some linguistic constructions are too ambiguous to be assigned a reading with confidence, and as a result he excludes 11% of sentences from his dataset.

4 Developing an Evaluation Dataset

4.1 Human Judgements

Ambiguity is context-, speaker- and listener-dependent, and so there are no absolute criteria for judging it. Therefore, we capture human judgements about the ambiguity of the sentences in our surveys in order to form our evaluation dataset. Rather than rely upon the judgement of one human reader, we take a consensus from multiple readers. Such an approach is known to be very effective albeit quite expensive (Berry *et al.* 03).

4.2 The Ambiguity Surveys

The sentences in our ambiguity surveys are drawn from our corpus of requirements specifications. Sentences — or non-sentential titles, bullet points etc — that contain coordinating conjunctions are identified. We do not use all the sentences containing a coordination that we find. Sentences containing coordinations which are syntactically unambiguous are identified, by hand, and dis-

| Head Word | % of Total | Example from Surveys |
|-----------|------------|--|
| Noun | 85.5 | <u>Communication</u> and <u>performance</u> requirements |
| Verb | 13.8 | Proceed to <u>enter</u> and <u>verify</u> the data |
| Adjective | 0.7 | It is very <u>common</u> and <u>ubiquitous</u> |

Table 2: Breakdown of sentences by head word type (head words are underlined)

| Modifier | % of Total | Example from Surveys |
|-----------------|------------|---|
| Noun | 46.4 | (It) targeted the project and election <u>managers</u> |
| Adjective | 23.2 | ... define <u>architectural</u> components and connectors |
| Prep | 15.9 | Facilitate the scheduling and performing <u>of works</u> |
| Verb | 5.8 | capacity and network resources <u>required</u> |
| Adverb | 4.4 | (It) might be <u>automatically</u> rejected or flagged |
| Relative Clause | 2.2 | Assumptions and dependencies <u>that are of importance</u> |
| Number | 0.7 | <u>zero</u> mean values and standard deviation |
| Other | 1.4 | increased by the <u>lack of</u> funding and local resources |

Table 3: Breakdown of sentences by modifier type (modifiers are underlined)

carded. A breakdown of the sentences we use by the part of speech of the head word of the coordinated phrases is given in Table 2. A breakdown by the part of speech of the external modifier is given in Table 3.

In total, we extracted 138 suitable coordination constructions and showed each one to 17 judges. They were asked to judge whether each coordination was to be read coordination first, coordination last or “ambiguous so that it might lead to misunderstanding”. In the last case, the coordination is then classed as an *acknowledged ambiguity* for that participant. Clearly, the dividing line between what would and what would not lead to misunderstandings is elusive. We take the view that, by using a sufficiently large number of judges, rogue interpretations are not accorded undue significance. Then we use ambiguity thresholds to account for, to whatever extent we desire, the varying differences in opinion that occur.

5 Empirical Study

5.1 Methodology

Here we introduce the metrics, ranking cut-offs and ambiguity thresholds that we use to get the most predictive and appropriate results from our data.

For each heuristic, the number of true positives is the number of coordinations for which the heuristic predicts the consensus result deter-

mined by the surveys, taking the ranking cut-off and ambiguity threshold into consideration. Precision for each heuristic is the number of true positives divided by the total number of positive results achieved by that heuristic. Recall for each heuristic is the number of true positives divided by the number of coordinations which that heuristic should have judged positively.

Precision is much more important to us than recall: we wish each heuristic to be a reliable indicator of how any given coordination should be read, rather than a catch-all technique. (Ultimately, we envisage using each heuristic as one of a large suite of techniques which will disambiguate many coordinations with good precision. Good recall may thereby be achieved if the heuristics have complementary coverage.) We use a weighted f-measure statistic, based on van Rijsbergen’s e-measure (vanRijsbergen 79), to combine precision and recall:

$$F\text{-Measure} = \frac{(1 + \beta) * Precision * Recall}{\beta^2 * Precision + Recall}$$

A weighting of $\beta = 0.5$ is commonly used to ensure that true positives are not obtained at the expense of also obtaining too many false positives. We use a weighting of $\beta = 0.25$, even more strongly in favour of precision. We aim to maximise the f-measure for all of our heuristics.

We employ 10-fold cross validation, which is an accurate and efficient way of ensuring that data is considered uniformly and that the resulting statistics are not biased (Weiss & Kulikowski 91). Our dataset is first randomly sorted to remove any bias caused by the order in which the sentences were collected. Then it is split into ten equal parts. Nine of the parts are concatenated and used for training to find the optimum ranking cut-off and ambiguity threshold for each heuristic. The heuristics are then run on the heldout tenth part using those cut-offs and ambiguity thresholds. This procedure is carried out for each heldout part, and the performances on all the heldout parts are then averaged to give the performances of the heuristics.

The results that we use from the Sketch Engine, for all three heuristics, are in the forms of rankings. We use rankings, rather than actual measures of frequency or similarity, as it is suggested that they are a more accurate measure for analysis based on word distribution — see for example (McLauchlan 04). For each heuristic, in or-

der to maximise its performance, a ranking cut-off is chosen, and rankings below that cut-off are not considered. The cut-off is found experimentally for each fold in the cross-validation exercise. For each of the three heuristics, the optimum cut-off is in fact found to be the same for all 10 folds.

We also determine different ambiguity thresholds for each heuristic in order to maximise its performance, (although a non-optimal threshold may in fact be preferred by a user). These are not always the same for each of the 10 folds of any heuristic. The ambiguity threshold is the minimum level of certainty that must be reflected by the consensus of survey judgements. Let us say that a particular coordination has been judged by 65% of the judges in the surveys to be coordination-first, and we are using a heuristic that predicts coordination-first readings. Then, if the ambiguity threshold is 60% the consensus judgement will be considered to be coordination-first, whereas it will not if the ambiguity threshold is 70%. It must be noted that this can significantly change the baseline — the percentage of true positives found if all coordinations are considered to be (in this case) coordination-first.

5.2 BNC and the Sketch Engine

All our heuristics use information generated by the Sketch Engine with the BNC as its data source. The BNC is a modern corpus containing over 100 million words of English. It is collated from a variety of sources, including some that share specialist terminology with our chosen domain.

The Sketch Engine accepts input of lemmatised verbs, nouns and adjectives. We use two of the key facilities offered by the Sketch Engine: a word sketch facility giving information about the frequency with which words are found collocated with each other, and a thesaurus giving distributional similarity between words.

The word sketch facility, rather than looking at an arbitrary window of text around a word, finds the correct collocations for the word by use of grammatical patterns (Kilgarriff *et al.* 04). Head words of coordinated phrases can therefore be found with some certainty. Parameters for minimum frequency, minimum salience and maximum number of matches can be entered. We use a minimum frequency of 1 and a minimum salience of 0 throughout, to ensure that we get results even for unusual words.

The Sketch Engine’s thesaurus is a distributional thesaurus in the tradition of (Sparck-Jones 86) and (Grefenstette 94); it measures similarity between any pair of words according to the number of corpus contexts they share. The corpus is parsed and all triples comprising a grammatical relation and two collocates, (eg $\langle object, drink, wine \rangle$ or $\langle modifier, wine, red \rangle$) are identified. Contexts are shared where the relation and one collocate remain the same, so $\langle object, drink, wine \rangle$ and $\langle object, drink, beer \rangle$ count towards the similarity between wine and beer. Shared collocates are weighted according to the product of their mutual information, and the similarity score is the sum of these weights across all shared collocates, as in (Lin 98). Distributional thesauruses are especially suitable for analysis of coordinations. For instance, words which have opposite meaning, such as *good* and *bad*, are often coordinated, and such words often have strong distributional similarity.

5.3 Coordination-Matches Heuristic

One approach to finding the most likely reading of a coordination, using a generic corpus, is to find out if that coordination occurs within that corpus. Our hypothesis here is that if a coordination in our dataset is found within the corpus, then that coordination is likely to be a syntactic unit and a coordination-first reading is the most likely.

Using the Sketch Engine, we search the BNC for each coordination in our dataset. This is done using the word sketch facility’s list of words that are conjoined with *and* or *or*. Each head word is looked up in turn. The ranking of the match of the second head word with the first head word may not be the same as the ranking of the match of the first head word with the second head word. This is because of the difference in overall frequency of the two words. We use the higher of the two rankings. We find that considering only the top 25 rankings is a suitable cut-off. An ambiguity threshold of 60% is found to be the optimum for all ten folds in the cross-validation exercise.

For the example from our dataset:

Security and Privacy Requirements,

the highest of the two rankings of *Security* and *Privacy* in the word sketch facility’s *and/or* lists is 9. This is in the top 25 rankings, and so the heuristic yields a positive result. Of

the 17 survey judges, 12 judged this ambiguity to be coordination-first — 1 judged it to be coordination-last and 4 judged it to be ambiguous — which is a certainty of $12/17 = 70.5\%$. This is over the ambiguity threshold of 60%, so the heuristic always yields a true positive result on this sentence.

Averaging for all the ten folds, the heuristic achieves 43.6% precision, 64.3% recall and 44.0% f-measure. However, the baselines are low, given the relatively high ambiguity threshold, giving 20.0 % points precision and 19.4 % points f-measure above the baselines.

5.4 Distributional-Similarity Heuristic

Our hypothesis here is that if two coordinated head words in our dataset display strong distributional similarity, then the coordinated phrases are likely to be a syntactic unit and a coordination-first reading is therefore the most likely. This is an idea suggested by Kilgarriff (Kilgarriff 03).

For each coordination, the lemmatised head words of both the coordinated phrases are looked up in the Sketch Engine’s thesaurus. The ranking of the match of the second head word with the first head word may not be the same as the ranking of the match of the first head word with the second head word. We use the higher of the two rankings. We find that considering only the top 10 matches is the best cut-off for our purposes. An ambiguity threshold of 50% produces optimal results for 7 of the folds, while 70% is optimal for the other 3.

For the example from our dataset:

processed and stored in database,

the verb *process* has the verb *store* as its second ranked match in the thesaurus, and vice versa. This is in the top 10 matches, so the heuristic yields a positive result. Of the 17 survey judges, only 1 judged the ambiguity to be coordination-first — 11 judged it to be coordination-last and 5 judged it to be ambiguous — which is a certainty of $1/17 = 5.9\%$. This is below both the ambiguity thresholds used by the folds, so the heuristic’s performance on this sentence always yields a false positive result.

Averaging for all the ten folds, the heuristic achieves 50.8% precision, 22.4% recall and 46.4% f-measure. Again the baselines are quite low, giving 11.5 % points precision and 5.8 % points f-measure above the baselines.

| Heuristic | Recall (%) | Baseline Precision (%) | Precision (%) | Precision above Base line (%) | F-Measure $\beta = 0.25$ (%) | F-Measure above Base line (%) |
|------------------------------|------------|------------------------|---------------|-------------------------------|------------------------------|-------------------------------|
| 1: Coordination-Matches | 64.3 | 23.6 | 43.6 | 20.0 | 44.0 | 19.4 |
| 2: Distributional-Similarity | 22.4 | 39.3 | 50.8 | 11.5 | 46.4 | 5.8 |
| 3: Collocation-Frequency | 35.3 | 22.1 | 40.0 | 17.9 | 37.3 | 14.1 |
| Combination of 1 & not 3 | 64.3 | 23.6 | 47.1 | 23.5 | 47.4 | 22.9 |

Table 4: Performance of our heuristics

5.5 Collocation-Frequency Heuristic

The third heuristic differs from the other two in that it predicts coordination-last readings, and in that it involves the modifiers of the coordinated phrases in our dataset. The hypothesis here is that if a modifier is shown to be collocated, in a corpus, much more frequently with the coordinated head word that it is nearest to than it is to the further head word, then it is more likely to form a syntactic unit with only the nearest head word. This implies that a coordination-last reading is the most likely.

Using the Sketch Engine’s word sketch facility’s collocation lists, we find the frequencies in the BNC with which the modifier in each sentence is collocated with the coordinated head words. There are lists for most relationships that a word can have with a modifier. We experimented with using as a cut-off the ratio of the collocation frequency with the nearest head word to the collocation frequency with the further head word. However, the optimal cut-off is found to be when there were no collocations between the modifier and the further head word, and any non-zero number of collocations between the modifier and the nearest head word. An ambiguity threshold of 40% produces optimum results for 8 of the folds, while 70% is optimal for the other 2.

For the example from our dataset:

project manager and designer,

project often modifies *manager* in the BNC but never *designer*. The heuristic therefore yields a positive result. Of the 17 survey judges, 8 judged this ambiguity to be a coordination-last reading — 4 judged it to be coordination-first and 5 judged it to be ambiguous — which is a certainty of $8/17 = 47.1\%$. This is over the ambiguity threshold of 40% but under the threshold of 70%. On this sentence, the heuristic therefore yields a true positive result for 8 of the folds but a false positive result for 2 of them.

Averaging for all the ten folds, the heuristic achieves 40.0% precision, 35.3% recall and 37.3% f-measure. The baselines are low, giving performances of 17.9 % points precision and 14.1 % points f-measure above the baselines.

5.6 Other Heuristics Considered

We experimented using heuristics based on the lengths of the coordinated phrases and the number agreement of coordinated nouns. The hypothesis was that disparities in either of these two factors would suggest that the coordination was not a syntactic unit and that a coordination-first reading was therefore not likely. We also tested a simple metric of semantic similarity, based on the closeness of the coordinated head words to their lowest common ancestor in hierarchies of WordNet hypernyms. However, these three heuristics demonstrated only very weak predictive power.

5.7 Evaluation

Table 4 summarises our results. These are not directly comparable with the results of the researchers presented in Table 1. This is because of the absence of some statistics in the published results of those researchers, and because we consider that highly ambiguous coordinations cannot be judged accurately and consistently by humans. On one hand, our use of ambiguity thresholds, which implement this consideration, makes the task easier by restricting the target set to relatively clear-cut examples. On the other hand the task is more difficult as there are fewer examples to find. The worth of the ambiguity thresholds is shown, however, in the improvements in performance that they give over the baselines. Our precision and f-measure in terms of percentage points over the baseline, except for the distributional-similarity heuristic, are encouraging.

We combine the two most successful heuristics, as shown in the last line of Table 4. These results are achieved by saying that a coordination-

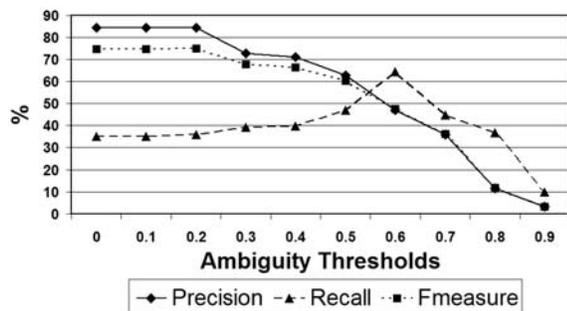


Figure 1: Heuristics 1 and 3 combined

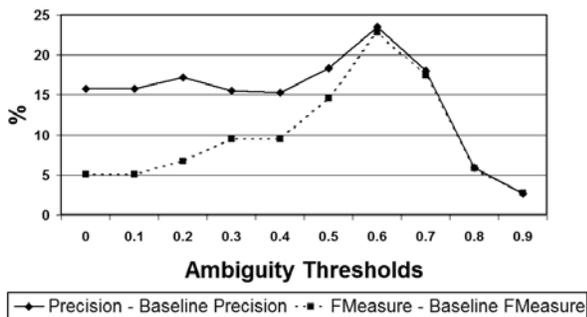


Figure 2: Heuristics 1 and 3 combined: percent-age points above baselines

first reading is predicted if the coordination-matches heuristic gives a positive result and the collocation-frequency heuristic gives a negative result. This gives the best performance of all. Figure 1 shows the precision, recall and f-measure for this combination of heuristics, at different ambiguity thresholds. As can be seen, high precision and f-measure can be achieved with low ambiguity thresholds. At these thresholds, even highly ambiguous coordinations are judged to be either coordination-first or -last. However, as can be seen in Figure 2, as percentage points above the baselines, these performances are relatively modest. The combination of heuristics performs best, relative to the baseline, when the ambiguity threshold is set at 0.6, aided by the high recall at this level.

Users of our technique can choose not to use the optimal ambiguity threshold. They choose whatever threshold they feel to be appropriate, considering the linguistic abilities of the people who will read the resulting documents and the importance that they give to ambiguity as a potential threat. Figure 3 shows the proportions of ambiguous and non-ambiguous interpretations at different ambiguity thresholds. It can be seen that none of the

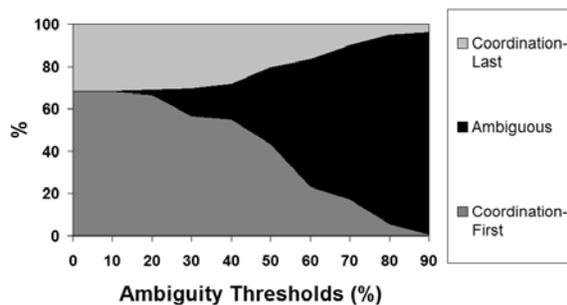


Figure 3: Proportions of ambiguous and non-ambiguous readings at different thresholds

coordinations are judged to be ambiguous with an ambiguity threshold of zero - a dangerous situation. At the other end of the spectrum, an ambiguity threshold of 90% results in almost everything being considered ambiguous - a situation which will waste users' time. From the former of these extremes to the other, the numbers of coordination-first and coordination-last readings are increasingly judged to be ambiguous at an approximately equal rate.

6 Conclusions

We conclude from our research that a surprising number of coordinations in a specialised corpus can also be found in a generic corpus. As a result, our heuristic for predicting that those former coordinations are to be read coordination first is the most effective and useful of the three which we present here.

We conclude that strong association between a modifier and the nearest coordinated head word — in comparison to the association with the further coordinated head word — indicates that those two words form a syntactic unit before the coordination takes place, and that a coordination-first reading is therefore less likely. We also conclude from the performance of our collocation frequency heuristic, that surprising numbers of those syntactic units which occur in our specialised corpus can also be found in a generic corpus.

We conclude from the performance of our distributional-similarity heuristic, that distributional similarity between head words of coordinated phrases is only a weak indicator that they form syntactic units leading to coordination-first interpretations. It might be concluded that this is due to the poor recall achieved by this heuristic, and it might still be the case

that this heuristic could be used in conjunction with other coordination-first predicting heuristics which have wider coverage. Currently, however, using this heuristic in conjunction with the other heuristics produces negligible improvements.

The improved performance obtained when we combined our two most successful heuristics shows that combining such predictors is beneficial. Overall, we conclude that word distribution information can be used effectively to indicate preferred readings of coordination ambiguities, particularly when they are not overly ambiguous. We have shown that this is achievable regardless of the type of words that are coordinated, and regardless of the type of word that modifies them.

We have found that people's judgements can vary quite widely. In addition to the acknowledged ambiguity that occurs when people judge a coordination to be ambiguous, there is also *unacknowledged ambiguity*. This occurs when various people have different interpretations of a sentence or phrase, but each of them thinks that theirs is the only possible interpretation of it. This is potentially more dangerous than acknowledged ambiguity: it is not noticed and it therefore doesn't get resolved. Unacknowledged ambiguity is measured as the number of judgements in favour of the minority non-ambiguous choice, over all the non-ambiguous judgements. The average unacknowledged ambiguity over all the examples in our dataset is 15.3%. Note that unacknowledged ambiguity is automatically included in the consensus judgement for each sentence.

7 Further Work

This paper is part of wider research into notifying users of ambiguities in text and informing them of how likely they are to be misunderstood by readers of the text. We intend to look at improving the heuristics that we have tested, and combining them with others in a manner which gives greater coverage and good precision. We will be testing heuristics based on morphology, typography and word sub-categorisation. Of interest to us in this further work is the analytical method of Okumura and Muraki, which incorporates three feature sets for analysing the parallelism of coordinated phrases (Okumura & Muraki 94).

At present in our dataset, although unacknowledged ambiguity generally occurs together with

acknowledged ambiguity, thereby reducing its danger, we consider that it may be interesting to investigate whether unacknowledged ambiguity has any particular characteristics.

References

- (Agarwal & Boggess 92) Rajeev Agarwal and Lois Boggess. A simple but useful approach to conjunct identification. In *Proceedings of the 30th conference on Association for Computational Linguistics*, pages 15–21. Association for Computational Linguistics, 1992.
- (Berry *et al.* 03) Daniel M. Berry, Erik Kamsties, and Michael M. Krieger. From contract drafting to software specification: Linguistic sources of ambiguity, 2003. A Handbook.
- (Gause & Weinberg 89) Donald C. Gause and Gerald M. Weinberg. *Exploring requirements: quality before design*. Dorset House, New York, 1989.
- (Goldberg 99) Miriam Goldberg. An unsupervised model for statistically determining coordinate phrase attachment. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 610–614. Association for Computational Linguistics, 1999.
- (Grefenstette 94) Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, 1994.
- (Kilgarrieff 03) Adam Kilgarrieff. Thesauruses for natural language processing. In *Proceedings of NLP-KE*, pages 5–13, Beijing, China, 2003.
- (Kilgarrieff *et al.* 04) Adam Kilgarrieff, Pavel Rychly, Pavel Smrz, and David Tugwell. The sketch engine. In *Proceedings of EURALEX 2004*, pages 105–116, 2004.
- (Lauer 95) Mark Lauer. Corpus statistics meet the noun compound: some empirical results. In *Proceedings of the 33rd conference on Association for Computational Linguistics*, pages 47–54, Morristown, NJ, USA, 1995. Association for Computational Linguistics.
- (Lin 98) Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774. Association for Computational Linguistics, 1998.
- (McLauchlan 04) Mark McLauchlan. Thesauruses for prepositional phrase attachment. In *Proceedings of CoNLL-2004*, pages 73–80. Boston, MA, USA, 2004.
- (Okumura & Muraki 94) Akitoshi Okumura and Kazunori Muraki. Symmetric pattern matching analysis for english coordinate structures. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 41–46. Association for Computational Linguistics, 1994.
- (Ratnaparkhi 98) Adwait Ratnaparkhi. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 1079–1085, 1998.
- (Resnik 99) Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
- (Sparck-Jones 86) Karen Sparck-Jones. *Synonymy and semantic classification*. Edinburgh University Press, 1986.
- (vanRijsbergen 79) C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, U.K., 1979.
- (Weiss & Kulikowski 91) Sholom M. Weiss and Casimir A. Kulikowski. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.