Tracking Morphological and Semantic Co-occurrences in Spontaneous Dialogues

Mark Seligman
Université Joseph Fourier
GETA, CLIPS, IMAG-campus, BP 53
385, rue de la Bibliothèque
38041 Grenoble Cedex 9, France
seligman@cerfnet.com

Jan Alexandersson
German Research Institute of Computer Science,
DFKI GmbH
Stuhlsatzenausweg 3
66 123 Saarbrücken, Germany
jan.alexanderson@dfki.de

Kristiina Jokinen
Flanders Language Valley
Sint Krispijnstraat 7
8900 Ieper, Belgium
Kristiina.Jokinen@flv.be

Abstract

In the processing of spontaneous language, information concerning discourse-level co-occurrences of words or morphemes — relatively long-term predictions on the scale of several utterances — may help to reduce perplexity in speech recognition, facilitate lexical disambiguation, and contribute to topic tracking. This working paper describes a new set of facilities for tracking lexical co-occurrences. The major innovation is the use of semantic smoothing: we track co-occurrences of semantic tokens associated with words or morphs in addition to co-occurrences of the words or morphs themselves. Such smoothing offers an approach to the problem of data sparseness: it is possible to retrieve reasonable semantically-mediated associations for morphs not in the training corpus. We report on preliminary experiments with a corpus of morphologically-tagged transcripts of 16 spontaneous Japanese dialogues concerning direction-finding and hotel arrangements. We close with discussion of lexical disambiguation and topic tracking as they relate to co-occurrence networks.

1 Introduction

In the processing of spontaneous language, predictions at the morphological or lexical level can be useful in several ways. First, they can aid speech recognition: to weigh recognition candidates appropriately, it is crucial to know which morphological or lexical items are most likely, given the words recently seen in a discourse. Second, lexical predictions can facilitate lexical disambiguation: to distinguish the meanings of an ambiguous word like *bank*, for instance, it is helpful to know which meaning is most predictable, given the words and meanings recently seen in a discourse. Both uses of lexical prediction can be seen as aspects of topic tracking.

The classical mechanism for lexical prediction is the use of N-gram statistics for the surface forms of the relevant lexical items. For the purposes of speech recognition and disambiguation in spontaneous language, however, this technique is unsatisfactory in two respects.

First, the range of predictions is too short, as predictions are usually made over a distance of no more than five words [Church, 1990]. To support bottom-up recognition and analysis of noisy material containing gaps and fragments, longer-rang predictions are needed as well. Long-range predictions should have the advantage of being stronger than very short-range predictions, since predicting what will come "soon" is in general easier than predicting what will come next; and they should require less data, since examples of occurrence "soon" will be found more often in a corpus than examples of consecutive occurrence.

Second, data for spontaneous language is often too sparse to support accurate and consistent predictions based on the surface items alone. A biased corpus may, for instance, indicate a strong association between *bus* and *street*, yet fail to associate *car* and *street*, even though this association would be intuitively expected and potentially useful.

To obtain predictions at longer range than N-grams statistics can provide, we might consider stochastic grammars [Black *et al.*, 1993]. However, these predict only within utterances, while our interest extends to predictions on the scale of several utterances. We might also consider discourse-oriented mechanisms such as centering and global focusing models [Grosz and Sidner, 1986], [Walker *et al.*, 1992]; but in fact these are not designed to predict the lexical items that will be seen a bit later in the dialogue.

Instead, we propose to permit the flexible definition of windows in a transcribed corpus within which co-occurrences of morphological or lexical elements can be examined. In this respect, our approach is similar to that of e.g. [Ferret and Grau, 1998].

With respect to the problem of data sparseness, however, our approach is, as far as we know, unprecedented. In addition to standard statistical smoothing procedures, we propose new techniques for semantic smoothing: we track co-occurrences of semantic tokens associated with words or morphs in addition to co-occurrences of the words or morphs themselves. The benefits of this smoothing technique appear especially in the possibility of retrieving reasonable semantically-mediated associations for morphs not in a training corpus.

In Section 2, we describe our definitions of segments and windows and the statistical approach which follows from them. In Section 3, we describe the spontaneous corpus used in our preliminary experiments. Section 4 presents our semantic smoothing approach, the central innovation of the current work. Section 5 evaluates our early experimental results. Section 6 is reserved for discussion of lexical disambiguation and topic tracking as it relates to co-occurrence networks. We conclude in Section 7 by outlining our plans for further experimentation.

2 Segments and Windows

We first permit the investigator to define minimal segments within the corpus: these may be utterances, sections bounded by pauses or significant morphemes such as conjunctions, hesitations, postpositions, etc. Windows composed of several successive minimal segments can then be recognized: let Si be the current segment and N be the number of additional segments in the window as it extends to the right. N = 2 would, for instance, give a window three segments long with Si as its first segment. Then if a given word or morpheme M1 occurs (at least once) in the initial segment, Si, we attempt to predict the other words or morphemes which will co-occur (at least once) anywhere in the window.

Specifically, a conditional probability Q can be defined as follows: Q(M1, M2) = P(M2 element of Si U Si+1 U Si+2 ... Si+N | M1 element of Si), where M1, M2 ... are morphemes, S1, S2 ... are minimal segments, and N is the width of window in segments. Q is thus the conditional probability that M2 is an element of the union of segments Si, Si+1, Si+2, and so on up to Si+N, given that M1 is an element of Si.

If N = 0, the window is a single segment. In this case, Q indicates the probability that M2 co-occurs in segment Si, given that M1 occurs there. If n is greater than 1, Q is the probability that M2 will be found in any of the window's several segments. (Thus, while Q usually predicts M2 later in a window, M2 may sometimes precede M1 if it occurs in window-initial segment Si.) Both the segment definition and the number of segments in a window can be adjusted to vary the range over which co-occurrence predictions are attempted.

3 Corpus and Early Experiments

For initial experiments, we used a morphologically-tagged corpus of 16 spontaneous Japanese dialogues concerning direction-finding and hotel arrangements [Loken-Kim and Yato, 1993]. We collected common-noun/common-noun, common-noun/verb, verb/common-noun, and verb/verb conditional probabilities in a three-segment window (n = 2). Conditional probability Q is computed among all morph pairs for these classes and stored to a database; pairs scoring below a threshold (0.1 for the initial experiments) were discarded. We also compute and store the mutual information for each morph pair, using the standard definition as in [Fano, 1961].

Fast queries of the database are then enabled. A central function is GET-MORPH-WINDOW-MATES, which provides all the window mates for a specified morph which belong to a specified class and have scores above a specified threshold for the specified co-occurrence measure (conditional probability or mutual information).

The intent is to use such queries in real time to support bottom-up, island-driven speech recognition and analysis. To support the establishment of island centers for such parsing, we also collect information on each corpus morph in isolation: its hit count and the segments it appears in, its unigram probability and probability of appearance in a segment, its probability of appearance in any given segment, etc. Once island hypotheses have been established based on this foundation, co-occurrence predictions will come into play for island extension. Global information concerning morphs is also recorded, showing that our present 16-dialogue corpus, in which a minimal segment has been defined as a single utterance, contains 1743 segments; is 19250 morphs long; has 949 different morphs; and has a morph unigram entropy of 6.8982.

4 Semantic Smoothing

It was suggested that sparse data should be somewhat less problematic for long-range than for short-range predictions. Still, there is never quite enough data; so abstraction, or smoothing, of the data will remain desirable. As a statistical smoothing measure, we support the use of standard techniques [Nadas, 1985] for smoothing both conditional probability and mutual information. We make no commitment to any particular technique, however.

In addition, we enable semantic smoothing in an innovative way. Thesaurus categories — *cats* for short — are sought for each corpus morph (and stored in a corpus-specific customized thesaurus for fast access). The common-noun *eki* (station), for instance, has among others the cat label "725a" (representing a semantic class of posts-or-stations) in the standard Kadokawa Japanese thesaurus [Ohno and Hamanish, 1981].

Equipped with such information, we can study the co-occurrence within windows of cats as well as morphs. For example, using n = 2, GET-CAT-WINDOW-MATES finds 36 cats co-occurring with "725a", one of the cats associated with *eki* (station), with a conditional probability Q greater than 0.10, including "459a" (*sewa*, taking-care-of or looking-after), "216a" (*henkou*, transfer), and "315b" (*ori*, getting-off). Since we have prepared an indexed reverse thesaurus for our corpus, we can quickly find the corpus morphs which have these cat labels, respectively *miru*, "look", *mieru*, "can see, visible"; *magaru*, "turn"; and *oriru*, "get off". The resulting morphs are related to the input morph *eki* via semantic rather than morph-specific co-occurrence. They thus form a broader, smoothed group.

This semantic smoothing procedure — morph to related cats, cats to co-occurring category window-mates, cats to related morphs — has been encapsulated in the function GET-MORPH-WINDOW-MATES-VIA-CATS. It permits filtering, so that morphs are output only if they belong to a desired morphological class and are mediated by cats whose co-occurrence likelihood is above a specified threshold.

Thesaurus categories are normally arranged in a type hierarchy. In the Kadokawa thesaurus, there are four levels of specificity: "725a" (posts-or-stations), mentioned above, belongs to a more general category "725" (stations-and-harbors), which in turn belongs to "72" (institutions), which belongs to "7" (society). Accordingly, we need not restrict co-occurrence investigation to cats at the level given by the thesaurus. Instead, knowing that "725a" occurred in a segment Si, we can infer that all of its ancestor cats occurred there as well; and can seek and record semantic co-occurrences at every level of specificity. This has been done; and GET-MORPH-WINDOW-MATES-VIA-CATS has a parameter permitting specification of the desired level of semantic smoothing. The more abstract the level of smoothing, the broader the resulting group of semantically-mediated morpheme co-occurrences.

The most desirable level for semantic smoothing is a matter for future experimentation. However, we can anticipate a general preference for the most specific predictions available: we would resort to semantic smoothing only when no morph-specific co-occurrences could be predicted at a certain threshold, and would resort to more abstract semantic smoothing only when more specific smoothing failed. Thus we provide a function GET-MORPH-WINDOW-MATES-MOST-SPECIFIC with this behavior. Its value for robustness appears especially in cases when the input is a morph which did not occur in the training corpus. Without semantic-smoothing-in-case-of-need, the attempt to make co-occurrence predictions would certainly fail; but with this possibility, reasonable predictions can often be made. An entry for the new morph is sought dynamically in the relevant thesaurus; any cats thus found are checked for likely co-occurring cats; and morphs associated with these cats in the training corpus can be delivered as output. For instance, *kuruma*, "car, auto", does not appear in our corpus. However, the Kadokawa thesaurus does list this morph with codes "997" (vehicles) and "985" (wheels), yielding a wide range of associated verbs from our corpus, including *iku*, "go", *tuku*, "arrive", and 44 others; and of common-nouns, including *shibasu*, "city bus", *ikikata* "(street) directions", and 52 others. For each morph retrieved in this way, the conditional probability of the mediating cat co-occurrence can be recovered.

5 Evaluation

We are presently reporting the implementation of facilities intended to enable many experiments concerning morphological and morpho-semantic co-occurrence; the experiments themselves remain for the future. Nevertheless, some indication of the basic usability of the data is in order.

Tools have been provided for comparing two corpora with respect to any of the fields in the records relating to morphs, morph co-occurrences, cats, or cat co-occurrences. Using these, we treated 15 of our dialogues as a training corpus, and the one remaining dialogue as a test corpus. We compared the two corpora in terms of unigram probabilities for morphs, and in terms of conditional probabilities for morph co-occurrences. (In both cases, statistically unsmoothed scores were used for simplicity of interpretation.)

Considering all morphological classes, we found 898 different morphs in the training corpus and 365 in the test. 314 morphs were found in both corpora; so our training corpus of 15 dialogues covered 314 out of 365 morphs in the test dialogue, or about 86 percent. Table 1 compares the corpora for 25 shared common-nouns. Morphs are listed in order of least difference between corpora. While most common-nouns occurred too rarely over both corpora to allow reasonable comparison of probabilities, even in this subset of a small corpus there were several which did give combined counts in the thirties or forties, and for these the closeness of probability scores between corpora seems encouraging. Results for verbs were comparable.

As to morph co-occurrences, we found 5162 co-occurrence pairs above a conditional probability threshold of 0.10 in the training corpus and 1552 in the test. Since 509 pairs occurred in both corpora, the training corpus covered 509 out of 1552, or 33 percent, of the test corpus. That is, one third of the morph co-occurrences with conditional probabilities above 0.10 in the test corpus were anticipated by the training corpus.

This coverage seems respectable, considering that the training corpus was small and that neither statistical nor semantic smoothing was used. More important than coverage, however, is the presence of numerous pairs for which good co-occurrence predictions were obtained. Such predictions differ from those made using n-grams in that they need not be chained, and thus need not cover the input to be useful: if consistently good co-occurrence predictions can be recognized, they can be exploited selectively.

Table 2 shows pair comparisons for the 35 pairs which occurred most often, taking the sum of counts in both corpora. Pairs are ordered by least difference between corpora, so that the best predictions appear first. Again, agreement seems encouraging for pairs occurring often enough to allow meaningful comparison. The figures obtained for cats and cat co-occurrences are comparable.

Among the morph co-occurrences with the highest counts, some clearly reflect grammatical patterns. For example, *hou* and *goza* are associated (141 times) because both elements are politeness markers characteristic of agents' speech when addressing customers. Other co-occurrences apparently reflect a common topic (see further below), as for the pair *deguchi* ("exit") and *densha* ("train"), with 40 hits.

6 Discussion: Disambiguation, Topics

We have mentioned two possible uses of lexical prediction: to constrain speech recognition, perhaps in an island-driven style; and to facilitate lexical disambiguation. Having already sketched an approach to the first application area, we will do likewise for the second. We will then turn to more general discussion of topic tracking and its relation to co-occurrence studies. Finally, we will assess several problems related to co-occurrence networks, disambiguation, and topic tracking.

Co-occurrence and Lexical Disambiguation. A weighted co-occurrence between morphemes or lexemes can be viewed as an association between these items; so the set of co-occurrences which CO-OC discovers can be viewed as an associative or semantic network. Spreading activation within such networks is often proposed as a method of lexical disambiguation. (For example, if the concept MONEY has been observed, then the lexical item *bank* has the meaning closest to MONEY in the network: "savings institution" rather than "edge of river", etc.)

[Schütze, 1998] and [Kikui, 1999] have employed co-occurrence networks for lexical disambiguation in this way – the first in the context of information retrieval, and the second in the service of machine translation. Their approaches differ from ours in two principal respects: (1) in defining windows within which to seek co-occurrences, they do not segment the corpus into utterances, pause units, etc. as we do, but instead simply count running words; and (2) they do not attempt semantic smoothing as we do.

Co-occurrence and Topic. We now turn as promised to the study of topic, since this broad field embraces both speech recognition and lexical disambiguation applications of co-occurrence statistics.

In one sense, the notion of topic is implicit whenever the attempt is made to predict upcoming words based on words already seen. One could plausibly claim, after all, that knowledge about typical word or sense groupings *is* knowledge about topics. Thus topic tracking would be implicit in any use of co-occurrence networks to support speech recognition or lexical disambiguation.

But of course one can also attempt to recognize and track topics explicitly. For example, in a corpus of conversations ranging from street directions to hotel reservations, we can try to explicitly mark the shift between the first and second topic. If the boundary can be reliably recognized, computational resources specialized for specific topics – for example, specialized sub-grammars – can be brought to bear.

CO-OC's co-occurrence networks might indeed be utilized in this way: topic boundaries can be hypothesized at spans within a dialogue where relatively few co-occurrence predictions are fulfilled. (Compare e.g. [Morris and Hirst, 1991], [Hearst, 1994], [Nomoto and Nitta, 1994], or [Kozima and Furugori, 1994].)

There is also the possibility of explicitly recognizing topics as sub-networks or clusters within an associative network like CO-OC's. The aim would be to formalize the above-mentioned intuition that knowledge about typical word or sense groupings *is* knowledge about topics. Individual topics might then be imagined as nebulae, in which words or concepts are the stars. Topic transitions would be seen as movement of activation from one nebula, or group of nebulae, to another. Since there can in theory be groups within groups, such analyses might support efforts to explicitly track movement among topics and sub-topics, e.g. using tree diagrams as in [Jokinen and Tanaka, 1998].

Visualization of topics as groupings within semantic networks may also inform the design of human interfaces. [Veling and van der Weerd, 1999] present an interesting application of this approach to information retrieval. (The paper is also valuable for its straightforward yet effective techniques for topologically analyzing co-occurrence networks in order to seek meaningful groupings within them.) In online searches, most queries contain few words. Since these are often ambiguous, many unusable hits are often returned. The proposed remedy is to discover topical word groupings in which the query words participate, so that users can select among them. More constrained queries can then be automatically composed. For instance, the query word *satellite* is found to belong to the word clusters (*programming*, *entertainment*) and (*rocket*, *orbit*, *space*, *NASA*). No labels are presently assigned to the groups, but the first group might be hand-labeled as a TELEVISION topic and the second as a SPACEFLIGHT topic, each with its own meaning of *satellite.*

Topical groupings based upon co-occurrence statistics can helpfully constrain information retrieval even when no lexical ambiguity is at issue. For instance, the query word *bomb* yields the groupings (*injured*, *explosion*, *injuries*), (*soldiers*, *wounded*, *officers*), and (*hospital*). Veling and van der Weerd interpret the first two clusters as "bombing by terrorists", "bombing in a war". The interpretation of the third cluster is left unclear, but we might venture to characterize it as "the aftermath of bombing".

When using co-occurrence networks to attempt automatic discovery of explicit topics, researchers need not always seek clusters or nebulae within the networks themselves. Another approach is proposed by [Ferret and Grau, 1998], who use such networks only to guide the segmentation of corpora, as suggested above. They then identify the most significant words within segments – those which proved most relevant to segmentation – and establish these as tentative topic word groups. An iterative merging of tentative groups yields final topic word groupings.

Problems and Issues. In any attempt to interpret groupings within co-occurrence networks as topics, some difficulties are to be expected. A particular problem is raised by ambiguous words which participate in more than one otherwise unrelated topic. Do they make their several topics overlap, thus inappropriately bringing into proximity words which should remain distant? For instance, assume that English *bank* is closely associated with *money* on one hand, and with *river* on the other. Now that *money* and *river* have been dragged close together by their mutual association with *bank*, are they henceforward to be considered topic-mates? Clearly not; rather, since they themselves rarely co-occur, we have *prima facia* evidence that two separate topics are involved, and that *bank* can belongs to both of them precisely because it is in fact polysemous. We might then consider the creation of two separate tokens for the word, say $bank_1$ and $bank_2$. Of course, the evidence for the topic distinction will be strengthened if many other words also independently support it (e.g. *interest*, *rate*, etc. on one hand and *boat*, *shore*, etc. on the other hand).

In the effort to avoid spurious interpretation of co-occurrence networks, it may also prove helpful to work with associations among concepts or semantic tokens, rather than exclusively among lexical elements. For instance, assume that instead of the lexical pairs just discussed, (*money*, *bank*) and (*bank*, *river*), the following pairs of semantic tokens were under examination: (CURRENCY, FINANCIAL-INSTITUTION) and (GEOGRAPHICAL-LOCATION, BODY-OF-WATER). In this case, the ambiguity of the surface item *bank* would be invisible, and there would be no problematic memberships in multiple clusters. As explained, CO-OC does seek just such semantic associations, on multiple levels of abstraction. (In fact, the program creates not one but several co-occurrence networks: one network for literal surface elements and additional networks at each level of semantic abstraction.)

On the other hand, it must be granted that the tracking of co-occurrences among semantic tokens brings its own problems with ambiguity. During a CO-OC run, a thesaurus will suggest multiple semantic tokens for polysemous lexical items. And so, for e.g. *bank*, which semantic token should we track: FINANCIAL-INSTITUTION (for the money sense) or GEOGRAPHICAL-LOCATION (for the river sense)? If we must solve this sort of ambiguity before we can use the resulting network to disambiguate (or define topics), do we not face a vicious circularity?

Three solutions to this last difficulty are under consideration:

〈 We can ignore the problem, hoping that when the co-occurrence network is used, good associations will outvote the bad ones: that is, that the associations of non-ambiguous topic-mate words and tokens will overwhelm the ambiguous associations.

〈 We can select topic labels by hand for the corpus of interest. An efficient interface, similar to that for a spell-checker, could make this selection tolerably efficient. As in spell-checking, a selection may remain valid for all occurrences throughout a sub-corpus; so the interface could include a SELECT-THROUGHOUT-DOCUMENT command which could save considerable time.

〈 We might bootstrap from clues in available resources. For instance, WordNet [Fellbaum, 1998], one potential source of semantic tokens, contains its own topic groupings, and these might be exploited to help select semantic tokens for tracking.

Attempts to exploit co-occurrence networks for lexical disambiguation and topic tracking certainly raise numerous issues. Nevertheless, the area remains a promising one for continued research.

7 Conclusions and Prospects

We have described a new set of facilities for tracking lexical co-occurrences within flexibly-definable windows and have given a preliminary demonstration of the basic usability of the information obtained. To address the problem of scarce data, we have proposed innovative techniques for semantic smoothing of co-occurrence information: we track co-occurrences of semantic tokens associated with morphs in addition to co-occurrences of the morphs themselves. Such smoothing enables the retrieval of reasonable semantically-mediated co-occurrence predictions for morphs which are rare or absent in the original corpus.

We have briefly discussed several possible applications of co-occurrence networks, with special interest in speech recognition, lexical disambiguation, and topic tracking.

To date, we have implemented all of the necessary programs for experimentation with the Japanese corpus described above, and have undertaken the proof-of-concept experiments reported here. We can now complete and extend our work in two directions.

First, we must apply our techniques to larger corpora in order to more fully evaluate their reliability. We are especially interested to determine the corpus size at which our predictions become stable. The approach we have in mind is similar to that of [Reithinger, 1995].

Secondly, we plan to extend our experiments to languages other than Japanese. English and German are of particular interest. We anticipate that only a moderate effort will be required to adapt our programs to accommodate new thesauri and new corpora.

References

[Black *et al.*, 1993] Black, E., R. Garside, and G. Leech. 1993. *Statistically-driven Computer Grammars of English: The IBM/Lancaster Approach. Language and Computers: Studies in Practical Linguistics* No. 8. Rodopi. Amsterdam, Atlanta GA. 1993.

[Church, 1990] Church, K. 1990. “Word Association Norms, Mutual Information, and Lexicography.” *Computational Linguistics*, 16, Number 1, March 1990.

[Fano, 1961] Fano, R. 1961. *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

[Fellbaum, 1998] Fellbaum, Christiane, ed. 1998. *Wordnet: An Electronic Lexical Database (Language, Speech and Communication).* MIT Press, Cambridge, March, 1998.

[Ferret and Grau, 1998] Ferret, O. and B. Grau. 1998. “Structuration d’un réseau de cooccurrences lexicales en domaines sémantiques par analyse de textes.” In Natural Language Processing and Industrial Applications (NLP+IA-98), Aug. 18-21, 1998. Moncton, New Brunswick, Canada.

[Grosz and Sidner, 1986] Grosz, B. and C. Sidner. 1986. “Attention, Intentions, and the Structure of Discourse.” *Computational Linguistics*, 12, pages 175-204.

[Hearst, 1994] Hearst, M. 1994. "Multi-Paragraph Segmentation of Expository Text." In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, NM, June 27-30, 1994.

[Jokinen and Tanaka, 1998] Jokinen, K. and H. Tanaka. 1998. "Context Management with Topics for Spoken Dialogue Systems." In *Proceedings of COLING-ACL 1998*, pages 631-637, Montreal, Quebec, Canada, August 10-14, 1998.

[Kikui, 1999] Kikui, Gen-ichiro. 1999. "Resolving Translation Ambiguity using Non-parallel Bilingual Corpora". In *ACL'99 Workshop on Unsupervised Learning in Natural Language Processing*. University of Maryland, June 21, 1999.

[Kozima and Furugori, 1994] Kozima, H. and T. Furugori. 1994. "Segmenting Narrative Text into Coherent Scenes." *Literary and Linguistic Computing*, Volume 9, Number 1.

[Loken-Kim and Yato, 1993] Loken-Kim, K. and F. Yato. 1993. *EMMI-ATR Environment for Multi-modal Interaction.* ATR Interpreting Telephony Research Laboratories, ATR Technical Report TR-I-0118. Also in EACL-89.

[Morris and Hirst, 1991] Morris, J. and G. Hirst. 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text." *Computational Linguistics*, 17, pages 21-48.

[Nadas, 1985] Nadas, A. 1985. "On Turings's Formula for Word Probabilities." *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-33, paragraph. 1414-1416, December, 1985.

[Nomoto and Nitta, 1994] Nomoto, T. and Y. Nitta. 1994. "A Grammatico-statistical Approach to Discourse Partitioning." In *Proceedings of COLING-94*, Aug. 5-9, 1994, Kyoto, Japan.

[Ohno and Hamanish, 1981] Ohno, S. and M. Hamanish. 1981. *Kadokawa Ruigo Shin-jiten* (Kadokawa New Word Category Dictionary). Kadokawa Shoten. January 30, 1981.

[Reithinger, 1995] Reithinger, Norbert. 1995. "Some Experiments in Speech Act Prediction." In *Empirical Methods in Discourse: Interpretation &Generation: Papers from the 1995 AAAI Symposium*. Johanna Moore and Marilyn Walker, Co-chairs. Stanford Technical Report SS-95-06, page 126.

[Schütze, 1998] Schütze, Hinrich. 1998. "Automatic Word Sense Discrimination". *Computational Linguistics*, volume 24, number 1, pages 97-124.

[Veling and van der Weerd, 1999] Veling, Anne and Peter van der Weerd. 1999. "Conceptual Grouping in Word Co-occurrence Networks." In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) '99*, August, 1999, Stockholm, Sweden.

[Walker *et al.*, 1992] Walker, M., M. Iida, and S. Cote. 1992. *Japanese Discourse and the Process of Centering.* University of Pennsylvania Technical Report, IRCS Report 92-14.

Table 1: Training and test corpora compared: 25 shared common-nouns

MORPH	Q1	Q2	DIFFNC	COUNT1	COUNT2	SUMMED COUNT
ato [hira.]	4.0E-4	4.0E-4	0.0	7	1	8
kinpen	4.0E-4	4.0E-4	0.0	7	1	8
toko	4.0E-4	4.0E-4	0.0	6	1	7
jitaku	4.0E-4	4.0E-4	0.0	6	1	7
ato [kanji]	4.0E-4	4.0E-4	0.0	6	1	7
nanika	4.0E-4	4.0E-4	0.0	6	1	7
deguchi	0.0024	0.0025	1.0E-4	41	6	47
shibasu	7.0E-4	8.0E-4	1.0E-4	11	2	13
houmen	5.0E-4	4.0E-4	1.0E-4	8	1	9
gamen	3.0E-4	4.0E-4	1.0E-4	5	1	6
atari	3.0E-4	4.0E-4	1.0E-4	5	1	6
atari	3.0E-4	4.0E-4	1.0E-4	5	1	6
konkai	3.0E-4	4.0E-4	1.0E-4	5	1	6
i	0.0023	0.0025	2.0E-4	39	6	45
basu	0.0018	0.0016	2.0E-4	31	4	35
hidarite	0.001	0.0012	2.0E-4	16	3	19
tatemono	0.001	8.0E-4	2.0E-4	16	2	18
ue	2.0E-4	4.0E-4	2.0E-4	4	1	5
yuugata	2.0E-4	4.0E-4	2.0E-4	4	1	5
furonto	2.0E-4	4.0E-4	2.0E-4	4	1	5
naname	2.0E-4	4.0E-4	2.0E-4	4	1	5
ichi	2.0E-4	4.0E-4	2.0E-4	3	1	4
shita	2.0E-4	4.0E-4	2.0E-4	3	1	4
katachi	2.0E-4	4.0E-4	2.0E-4	3	1	4
T	2.0E-4	4.0E-4	2.0E-4	3	1	4

Table 2: The 35 most frequent pairs over both corpora

MORPH1	POS	MORPH2	POS	Q1	Q2	DIFFNC	COUNT1	COUNT2	SUM
michi	C-NOUN	ji	C-NOUN	1.0	1.0	0.0	21	1	22
noritsu	VERB	densha	C-NOUN	1.0	1.0	0.0	7	1	8
me	C-NOUN	mae	C-NOUN	1.0	1.0	0.0	5	2	7
jitaku	C-NOUN	bango	C-NOUN	1.0	1.0	0.0	6	1	7
wakari	C-NOUN	na	VERB	1.0	1.0	0.0	4	2	6
gamen	C-NOUN	chizu	C-NOUN	1.0	1.0	0.0	5	1	6
furonto	C-NOUN	you	C-NOUN	1.0	1.0	0.0	4	1	5
furonto	C-NOUN	namae	C-NOUN	1.0	1.0	0.0	4	1	5
meda	VERB	waka	VERB	1.0	1.0	0.0	3	1	4
meda	VERB	tatemono	C-NOUN	1.0	1.0	0.0	3	1	4
T	C-NOUN	mie	VERB	1.0	1.0	0.0	3	1	4
T	C-NOUN	maga	VERB	1.0	1.0	0.0	3	1	4
T	C-NOUN	aru	VERB	1.0	1.0	0.0	3	1	4
T	C-NOUN	michi	C-NOUN	1.0	1.0	0.0	3	1	4
T	C-NOUN	hidari	C-NOUN	1.0	1.0	0.0	3	1	4
T	C-NOUN	ji	C-NOUN	1.0	1.0	0.0	3	1	4
T	C-NOUN	hou	C-NOUN	1.0	1.0	0.0	3	1	4
kae	VERB	chizu	C-NOUN	1.0	1.0	0.0	1	2	3
ike	VERB	i	VERB	1.0	1.0	0.0	2	1	3
konkousu	C-NOUN	deguchi	C-NOUN	1.0	1.0	0.0	1	1	2
ikutsu	C-NOUN	basutei	C-NOUN	1.0	1.0	0.0	1	1	2
gakkai	C-NOUN	oshie	VERB	1.0	1.0	0.0	1	1	2
gakkai	C-NOUN	basho	C-NOUN	1.0	1.0	0.0	1	1	2
yoru	C-NOUN	shingaru	C-NOUN	1.0	1.0	0.0	1	1	2
ma	VERB	ka	VERB	0.2024	0.202	4.0E-4	25	5	30
i	VERB	de	VERB	0.1683	0.1688	5.0E-4	77	6	83
namae	C-NOUN	you	C-NOUN	0.2537	0.2532	5.0E-4	20	4	24
oshie	VERB	hou	C-NOUN	0.3354	0.3361	7.0E-4	24	3	27
hou	C-NOUN	goza	VERB	0.198	0.2	0.002	131	10	141
deguchi	C-NOUN	na	VERB	0.2	0.202	0.002	35	5	40
deguchi	C-NOUN	densha	C-NOUN	0.2	0.202	0.002	35	5	40
you	C-NOUN	furonto	C-NOUN	0.3333	0.3361	0.0028	12	3	15
de	VERB	shoumen	C-NOUN	0.1749	0.1688	0.0061	69	6	75
de	VERB	i	C-NOUN	0.1749	0.1688	0.0061	69	6	75
hidari	C-NOUN	aru	VERB	0.5	0.5063	0.0063	26	2	28