Ordering Among Premodifiers

James Shaw and **Vasileios Hatzivassiloglou**

Department of Computer Science
Columbia University
New York, N.Y. 10027, USA
{shaw, vh}@cs.columbia.edu

Abstract

We present a corpus-based study of the sequential ordering among premodifiers in noun phrases. This information is important for the fluency of generated text in practical applications. We propose and evaluate three approaches to identify sequential order among premodifiers: direct evidence, transitive closure, and clustering. Our implemented system can make over 94% of such ordering decisions correctly, as evaluated on a large, previously unseen test corpus.

1 Introduction

Sequential ordering among premodifiers affects the fluency of text, e.g., "large foreign financial firms" or "zero-coupon global bonds" are desirable, while "foreign large financial firms" or "global zero-coupon bonds" sound odd. The difficulties in specifying a consistent ordering of adjectives have already been noted by linguists [Whorf 1956; Vendler 1968]. During the process of generating complex sentences by combining multiple clauses, there are situations where multiple adjectives or nouns modify the same head noun. The text generation system must order these modifiers in a similar way as domain experts use them to ensure fluency of the text. For example, the description of the age of a patient precedes his ethnicity and gender in medical domain as in "a 50 year-old white female patient". Yet, general lexicons such as WordNet [Miller *et al.* 1990] and COMLEX [Grishman *et al.* 1994], do not store such information.

In this paper, we present automated techniques for addressing this problem of determining, given two premodifiers A and B, the preferred ordering between them. Our methods rely on and generalize empirical evidence obtained from large corpora, and are evaluated objectively on such corpora. They are informed and motivated by our practical need for ordering multiple premodifiers in the MAGIC system [Dalal *et al.* 1996]. MAGIC utilizes co-ordinated text, speech, and graphics to convey information about a patient's status after coronary bypass surgery; it generates concise but complex descriptions that frequently involve four or more premodifiers in the same noun phrase.

To demonstrate that a significant portion of noun phrases have multiple premodifiers, we extracted all the noun phrases (NPs, excluding pronouns) in a two million word corpus of medical discharge summaries and a 1.5 million word Wall Street Journal (WSJ) corpus (see Section 4 for a more detailed description of the corpora). In the medical corpus, out of 612,718 NPs, 12% have multiple premodifiers and 6% contain solely multiple adjectival premodifiers. In the WSJ corpus, the percentages are a little lower, 8% and 2%, respectively. These percentages imply that one in ten NPs contains multiple premodifiers while one in 25 contains just multiple adjectives.

Traditionally, linguists study the premodifier ordering problem using a *class-based* approach. Based on a corpus, they propose various semantic classes, such as color, size, or nationality, and specify a sequential order among the classes. However, it is not always clear how to map premodifiers to these classes, especially in domain-specific applications. This justifies the exploration of empirical, corpus-based alternatives, where the ordering between A and B is determined either from direct prior evidence in the corpus or indirectly through other words whose relative order to A and B has already been established. The corpus-based approach lacks the ontological knowledge used by linguists, but uses a much larger amount of direct evidence, provides answers for many more premodifier orderings, and is portable to different domains.

In the next section, we briefly describe prior linguistic research on this topic. Sections 3 and 4 describe the methodology and corpus used in our analysis, while the results of our experiments are presented in Section 5. In Section 6, we demonstrate how we incorporated our ordering results in a general text generation system. Finally, Section 7 discusses possible improvements to our current approach.

2 Related Work

The order of adjectives (and, by analogy, nominal premodifiers) seems to be outside of the grammar; it is influenced by factors such as polarity [Malkiel 1959], scope, and collocational restrictions [Bache 1978]. Linguists [Goyvaerts 1968; Vendler 1968; Quirk and Greenbaum 1973; Bache 1978; Dixon 1982] have performed manual analyses of (small) corpora and pointed out various tendencies, such as the facts that underived adjectives often precede derived adjectives, and shorter modifiers precede longer ones. Given the difficulty of adequately describing all factors that influence the order of premodifiers, most earlier work is based on placing the premodifiers into broad semantic classes, and specifying an order among these classes. More than ten classes have been proposed, with some of them further broken down into subclasses. Though not all these studies agree on the details, they demonstrate that there is fairly rigid regularity in the ordering of adjectives. For example, Goyvaerts [1968, p. 27] proposed the order quality $\prec$ size/length/shape $\prec$ old/new/young $\prec$ color $\prec$ nationality $\prec$ style $\prec$ gerund $\prec$ denominal[1]; Quirk and Greenbaum [1973, p. 404] the order general $\prec$ age $\prec$ color $\prec$ participle $\prec$ provenance $\prec$ noun $\prec$ denominal; and Dixon [1982, p. 24] the order value $\prec$ dimension $\prec$ physical property $\prec$ speed $\prec$ human propensity $\prec$ age $\prec$ color.

Researchers have also looked at adjective ordering across languages [Dixon 1982; Frawley 1992]. Frawley [1992], for example, observed that English, German, Hungarian, Polish, Turkish, Hindi, Persian, Indonesian, and Basque, all order value before size and both of those before color.

As with most manual analyses, the corpora used in these analyses are relatively small compared with modern corpora-based studies. Furthermore, different criteria were used to arrive at the classes. To illustrate, the adjective "beautiful" can be classified into at least two different classes because the phrase "beautiful dancer" can be transformed from either the phrase "dancer who is beautiful", or "dancer who dances beautifully".

Several deep semantic features have been proposed to explain the regularity among the positional behavior of adjectives. Teyssier [1968] first proposed that adjectival functions, i.e. identification, characterization, and classification, affect adjective order. Martin [1970] carried out psycholinguistic studies of adjective ordering. Frawley [1992] extended the work by Kamp [1975] and proposed that intensional modifiers precede extensional ones. However, while these studies offer insights at the complex phenomenon of adjective ordering, they cannot be directly mapped to a computational procedure.

On the other hand, recent computational work on sentence planning [Bateman *et al.* 1998; Shaw 1998b] indicates that generation research has progressed to a point where hard problems such as ellipsis, conjunctions, and ordering of paradigmatically related constituents are addressed. Computational corpus studies related to adjectives were performed by [Justeson and Katz 1991; Hatzivassiloglou and McKeown 1993; Hatzivassiloglou and McKeown 1997], but none was directly on the ordering problem. [Knight and Hatzivassiloglou 1995] and [Langkilde and Knight 1998] have proposed models for incorporating statistical information into a text generation system, an approach that is similar to our way of using the evidence obtained from corpus in our actual generator.

3 Methodology

In this section, we discuss how we obtain the premodifier sequences from the corpus for analysis and the three approaches we use for establishing ordering relationships: direct corpus evidence, transitive closure, and clustering analysis. The result of our analysis is embodied in a

[1] Where $A \prec B$ stands for "A precedes B".

function, *compute_order*(A, B), which returns the sequential ordering between two premodifiers, word A and word B.

To identify orderings among premodifiers, premodifier sequences are extracted from simplex NPs. A simplex NP is a maximal noun phrase that includes premodifiers such as determiners and possessives but not post-nominal constituents such as prepositional phrases or relative clauses. We use a part-of-speech tagger [Brill 1992] and a finite-state grammar to extract simplex NPs. The noun phrases we extract start with an optional determiner (DT) or possessive pronoun (PRP$), followed by a sequence of cardinal numbers (CDs), adjectives (JJs), nouns (NNs), and end with a noun. We include cardinal numbers in NPs to capture the ordering of numerical information such as age and amounts. Gerunds (tagged as VBG) or past participles (tagged as VBN), such as "heated" in "heated debate", are considered as adjectives if the word in front of them is a determiner, possessive pronoun, or adjective, thus separating adjectival and verbal forms that are conflated by the tagger. A morphology module transforms plural nouns and comparative and superlative adjectives into their base forms to ensure maximization of our frequency counts. There is a regular expression filter which removes obvious concatenations of simplex NPs such as "takeover bid last week" and "Tylenol 40 milligrams".

After simplex NPs are extracted, sequences of premodifiers are obtained by dropping determiners, genitives, cardinal numbers and head nouns. Our subsequent analysis operates on the resulting premodifier sequences, and involves three stages: direct evidence, transitive closure, and clustering. We describe each stage in more detail in the following subsections.

3.1 Direct Evidence

Our analysis proceeds on the hypothesis that the relative order of two premodifiers is fixed and independent of context. Given two premodifiers A and B, there are three possible underlying orderings, and our system should strive to find which is true in this particular case: either A comes before B, B comes before A, or the order between A and B is truly unimportant. Our first stage relies on frequency data collected from a training corpus to predict the order of adjective and noun premodifiers in an unseen test corpus.

To collect direct evidence on the order of premodifiers, we extract all the premodifiers from the corpus as described in the previous subsection. We first transform the premodifier sequences into *ordered pairs*. For example, the phrase "well-known traditional brand-name drug" has three ordered pairs, "well-known $\prec$ traditional", "well-known $\prec$ brand-name", and "traditional $\prec$ brand-name". A phrase with n premodifiers will have $\binom{n}{2}$ ordered pairs. From these ordered pairs, we construct a $w \times w$ matrix *Count*, where w the number of distinct modifiers. The cell $[A, B]$ in this matrix represents the number of occurrences of the pair "A $\prec$ B", in that order, in the corpus.

Assuming that there is a preferred ordering between premodifiers A and B, one of the cells $Count[A, B]$ and $Count[B, A]$ should be much larger than the other, at least if the corpus becomes arbitrarily large. However, given a corpus of a fixed size there will be many cases where the frequency counts will both be small. This data sparseness problem is exacerbated by the inevitable occurrence of errors during the data extraction process, which will introduce some spurious pairs (and orderings) of premodifiers. We therefore apply probabilistic reasoning to determine when the data is strong enough to decide that $A \prec B$ or $B \prec A$. Under the null hypothesis that the two premodifiers order is arbitrary, the number of times we have seen one of them follows the binomial distribution with parameter $p = 0.5$. The probability that we would see the actually observed number of cases with $A \prec B$, say m, among n pairs involving A and B is

$$\sum_{k=m}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{(n-k)} \qquad (1)$$

which for the special case $p = 0.5$ becomes

$$\sum_{k=m}^{n} \binom{n}{k} \cdot 0.5^k \cdot 0.5^{(n-k)} = \sum_{k=m}^{n} \binom{n}{k} \cdot 0.5^n \quad (2)$$

If this probability is low, we reject the null hypothesis and conclude that A indeed precedes (or follows, as indicated by the relative frequencies) B.

3.2 Transitivity

As we mentioned before, sparse data is a serious problem in our analysis. For example, the matrix of frequencies for adjectives in our training corpus from the medical domain is 99.8% empty—only 9,106 entries in the 2,232 × 2,232 matrix contain non-zero values. To compensate for this problem, we explore the transitive properties between ordered pairs by computing the transitive closure of the ordering relation. Utilizing transitivity information corresponds to making the inference that $A \prec C$ follows from $A \prec B$ and $B \prec C$, even if we have no direct evidence for the pair (A, C) but provided that there is no contradictory evidence to this inference either. This approach allows us to fill from 15% (WSJ) to 30% (medical corpus) of the entries in the matrix.

To compute the transitive closure of the order relation, we map our underlying data to special cases of *commutative semirings* [Pereira and Riley 1997]. Each word is represented as a node of a graph, while arcs between nodes correspond to ordering relationships and are labeled with elements from the chosen semiring. This formalism can be used for a variety of problems, using appropriate definitions of the two binary operators (*collection* and *extension*) that operate on the semiring's elements. For example, the all-pairs shortest-paths problem in graph theory can be formulated in a *min-plus* semiring over the real numbers with the operators *min* for collection and + for extension. Similarly, finding the transitive closure of a binary relation can be formulated in a *max-min* semi-ring or a *or-and* semiring over the set $\{0, 1\}$. Once the proper operators have been chosen, the generic Floyd-Warshall algorithm [Aho *et al.* 1974] can solve the corresponding problem without modifications.

We explored three semirings appropriate to our problem. First, we apply the statistical decision procedure of the previous subsection and assign to each pair of premodifiers either 0 (if we don't have enough information about their preferred ordering) or 1 (if we do). Then we use the *or-and* semiring over the {0,1} set; in the transitive closure, the ordering $A \prec B$ will be present if at least one path connecting A and B via ordered pairs exists. Note that it is possible for both $A \prec B$ and $B \prec A$ to be present in the transitive closure.

This model involves conversions of the corpus evidence for each pair into hard decisions on whether one of the words in the pair precedes the other. To avoid such early commitments, we use a second, refined model for transitive closure where the arc from A to B is labeled with the probability that A precedes indeed B. The natural extension of the ({0, 1}, *or*, *and*) semiring when the set of labels is replaced with the interval [0, 1] is then ([0, 1], *max*, *min*). We estimate the probability that A precedes B as one minus the probability of reaching that conclusion in error, according to the statistical test of the previous subsection (i.e., one minus the sum specified in equation (2). We obtained similar results with this estimator and with the maximal likelihood estimator (the ratio of the number of times A appeared before B to the total number of pairs involving A and B).

Finally, we consider a third model in which we explore an alternative to transitive closure. Rather than treating the number attached to each arc as a probability, we treat it as a *cost*, the cost of erroneously assuming that the corresponding ordering exists. We assign to an edge (A, B) the negative logarithm of the probability that A precedes B; probabilities are estimated as in the previous paragraph. Then our problem becomes identical to the all-pairs shortest-path problem in graph theory; the corresponding semiring is ((0, $+\infty$), *min*, +). We use logarithms to address computational precision issues stemming from the multiplication of small probabilities, and negate the logarithms so that we cast the problem as a minimization task (i.e., we find the path in the graph the minimizes the total sum of negative log probabilities, and therefore maximizes the product of the original probabilities).

3.3 Clustering

As noted earlier, earlier linguistic work on the ordering problem puts words into semantic classes and generalizes the task from ordering between specific words to ordering the corresponding classes. We follow a similar, but evidence-based, approach for the pairs of words that neither direct evidence nor transitivity can resolve. We compute an *order similarity* measure between any two premodifiers, reflecting whether the two words share the same pattern of relative order with other premodifiers for which we have sufficient evidence. For each pair of premodifiers A and B, we examine every other premodifier in the corpus, X; if both $A \prec X$ and $B \prec X$, or both $A \succ X$ and $B \succ X$, one point is added to the similarity score between A and B. If on the other hand $A \prec X$ and $B \succ X$, or $A \succ X$ and $B \prec X$, one point is subtracted. X does not contribute to the similarity score if there is not sufficient prior evidence for the relative order of X and A, or of X and B. This procedure closely parallels non-parametric distributional tests such as Kendall's τ [Kendall 1938].

The similarity scores are then converted into dissimilarities and fed into a non-hierarchical clustering algorithm [Späth 1985], which separates the premodifiers in groups. This is achieved by minimizing an *objective function*, defined as the sum of within-group dissimilarities over all groups. In this manner, premodifiers that are closely similar in terms of sharing the same relative order with other premodifiers are placed in the same group.

Once classes of premodifiers have been induced, we examine every pair of classes and decide which precedes the other. For two classes C_1 and C_2, we extract all pairs of premodifiers (x, y) with $x \in C_1$ and $y \in C_2$. If we have evidence (either direct or through transitivity) that $x \prec y$, one point is added in favor of $C_1 \prec C_2$; similarly, one point is subtracted if $x \succ y$. After all such pairs have been considered, we can then predict the relative order between words in the two clusters which we haven't seen together earlier. This method makes (weak) predictions for any pair (A, B) of words, except if (a) both A and B are placed in the same cluster; (b) no ordered pairs (x, y) with one element in the class of A and one in the class of B have been identified; or (c) the evidence for one class preceding the other is in the aggregate equally strong in both directions.

4 The Corpus

We used two corpora for our analysis: hospital discharge summaries from 1991 to 1997 from the Columbia-Presbyterian Medical Center, and the January 1996 part of the Wall Street Journal corpus from the Penn TreeBank [Marcus *et al.* 1993]. To facilitate comparisons across the two corpora, we intentionally limited ourselves to only one month of the WSJ corpus, so that approximately the same amount of data would be examined in each case. The text in each corpus is divided into a training part (2.3 million words for the medical corpus and 1.5 million words for the WSJ) and a test part (1.2 million words for the medical corpus and 1.6 million words for the WSJ).

All domain-specific markup was removed, and the text was processed by the MXTERMINATOR sentence boundary detector [Reynar and Ratnaparkhi 1997] and Brill's part-of-speech tagger [Brill 1992]. Noun phrases and pairs of premodifiers were extracted from the tagged corpus according to the methods of Section 3. From the medical corpus, we retrieved 934,823 simplex NPs, of which 115,411 have multiple premodifiers and 53,235 multiple adjectives only. The corresponding numbers for the WSJ corpus were 839,921 NPs, 68,153 NPs with multiple premodifiers, and 16,325 NPs with just multiple adjectives.

We separately analyze two groups of premodifiers: adjectives, and adjectives plus nouns modifying the head noun. Although our techniques are identical in both cases, the division is motivated by our expectation that the task will be easier when modifiers are limited to adjectives, because nouns tend to be harder to match correctly with our finite-state grammar and the input data is sparser for nouns.

5 Results

We applied the three ordering algorithms proposed in this paper to the two corpora separately for adjectives and adjectives plus nouns. For our first technique of directly using evidence from a separate training corpus, we filled the *Count* matrix (see Section 3.1) with the frequencies of each ordering for each pair of premodifiers using the training corpora. Then, we calculated which of those pairs correspond to a true underlying order relation, i.e., pass the statistical test of Section 3.1 with the probability given by equation (2) less than or equal to 50%. We then examined each *instance* of ordered premodifiers in the corresponding test corpus, and counted how many of those the direct evidence method could predict correctly. Note that if A and B occur sometimes as $A \prec B$ and some-

Corpus	Test pairs	Direct evidence	Transitivity (max-min)	Transitivity (min-plus)
Medical/ adjectives	27,670	**92.67%** (88.20%–98.47%)	**89.60%** (94.94%–91.79%)	**94.93%** (97.20%–96.16%)
Financial/ adjectives	9,925	**75.41%** (53.85%–98.37%)	**79.92%** (72.76%–90.79%)	**80.77%** (76.36%–90.18%)
Medical/ adjectives and nouns	74,664	**88.79%** (80.38%–98.35%)	**87.69%** (90.86%–91.50%)	**90.67%** (91.90%–94.27%)
Financial/ adjectives and nouns	62,383	**65.93%** (35.76%–95.27%)	**69.61%** (56.63%–84.51%)	**71.04%** (62.48%–83.55%)

Table 1: Accuracy of direct-evidence and transitivity methods on different data strata of our test corpora. In each case, overall accuracy is listed first in bold, and then, in parentheses, the percentage of the test pairs that the method has an opinion for (rather than randomly assign a decision because of lack of evidence) and the accuracy of the method within that subset of test cases.

times as $B \prec A$, no prediction method can get all those instances correct. We elected to follow this evaluation approach, which lowers the apparent scores of our method, rather than forcing each pair in the test corpus to one unambiguous category ($A \prec B$, $B \prec A$, or arbitrary).

Under this evaluation method, stage one of our system achieves on adjectives in the medical domain 98.47% correct decisions on pairs for which a determination of order could be made. Since 11.80% of the total pairs in the test corpus involve previously unseen combinations of adjectives and/or new adjectives, the overall accuracy is 92.67%. The corresponding accuracy on data for which we can make a prediction and the overall accuracy is 98.35% and 88.79% for adjectives plus nouns in the medical domain, 98.37% and 75.41% for adjectives in the WSJ data, and 95.27% and 65.93% for adjectives plus nouns in the WSJ data. Note that the WSJ corpus is considerably more sparse, with 64.24% unseen combinations of adjective and noun premodifiers in the test part. Using lower thresholds in equation (2) results in a lower percentage of cases for which the system has an opinion but a higher accuracy for those decisions. For example, a threshold of 25% results in the ability to predict 83.72% of the test adjective pairs in the medical corpus with 99.01% accuracy for these cases.

We subsequently applied the transitivity stage, testing the three semiring models discussed in Section 3.2. Early experimentation indicated that the *or-and* model performed poorly, which we attribute to the extensive propagation of decisions (once a decision in favor of the existence of an ordering relationship is made, it cannot be revised even in the presence of conflicting evidence). Therefore we report results below for the other two semiring models. Of those, the *min-plus* semiring achieved higher performance. That model offers additional predictions for 9.00% of adjective pairs and 11.52% of adjective-plus-noun pairs in the medical corpus, raising overall accuracy of our predictions to 94.93% and 90.67% respectively. Overall accuracy in the WSJ test data was 80.77% for adjectives and 71.04% for adjectives plus nouns. Table 1 summarizes the results of these two stages.

Finally, we applied our third, clustering approach on each data stratum. Due to data sparseness and computational complexity issues, we clustered the most frequent words in each set of premodifiers (adjectives or adjectives plus nouns), selecting those that occurred at least 50 times in the training part of the corpus being analyzed. We report results for the adjectives selected in this manner (472 frequent adjectives from the medical corpus and 307 adjectives from the WSJ corpus). For these words, the information collected by the first two stages of the system covers most pairs. Out of the 111,176 (=472·471/2) possible pairs in the medical data, the direct evidence and transitivity stages make predictions for 105,335 (94.76%); the corresponding number for the WSJ data is 40,476 out of 46,971 possible pairs (86.17%).

The clustering technique makes ordering predictions for a part of the remaining pairs—on average, depending on how many clusters are created, this method produces answers for 80% of the ordering cases that remained unanswered after the first two stages in the medical corpus, and for 54% of the unanswered cases in the WSJ corpus. Its accuracy on these predictions is 56% on the medical corpus, and slightly worse than the baseline 50% on the WSJ corpus; this latter, aberrant result is due to a single, very frequent pair, *chief executive*, in which *executive* is consistently mistagged as an adjective by the part-of-speech tagger.

Qualitative analysis of the third stage's output indicates that it identifies many interesting relationships between premodifiers; for example, the pair of most similar premodifiers on the basis of positional information is *left* and *right*, which clearly fall in a class similar to the semantic classes manually constructed by linguists. Other sets of adjectives with strongly similar members include {*mild*, *severe*, *significant*} and {*cardiac*, *pulmonary*, *respiratory*}.

We conclude our empirical analysis by testing whether a separate model is needed for predicting adjective order in each different domain. We trained the first two stages of our system on the medical corpus and tested them on the WSJ corpus, obtaining an overall prediction accuracy of 54% for adjectives and 52% for adjectives plus nouns. Similar results were obtained when we trained on the financial domain and tested on medical data (58% and 56%). These results are not much better than what would have been obtained by chance, and are clearly inferior to those reported in Table 1. Although the two corpora share a large number of adjectives (1,438 out of 5,703 total adjectives in the medical corpus and 8,240 in the WSJ corpus), they share only 2 to 5% of the adjective *pairs*. This empirical evidence indicates that adjectives are used differently in the two domains, and hence domain-specific probabilities must be estimated, which increases the value of an automated procedure for the prediction task.

6 Using Ordered Premodifiers in Text Generation

Extracting sequential ordering information of premodifiers is an off-line process, the results of which can be easily incorporated into the overall generation architecture. We have integrated the function *compute_order*(A, B) into our multimedia presentation system MAGIC [Dalal *et al.* 1996] in the medical domain and resolved numerous premodifier ordering tasks correctly. Example cases where the statistical prediction module was helpful in producing a more fluent description in MAGIC include placing age information before ethnicity information and the latter before gender information, as well as specific ordering preferences, such as "thick" before "yellow" and "acute" before "severe". MAGIC's output is being evaluated by medical doctors, who provide us with feedback on different components of the system, including the fluency of the generated text and its similarity to human-produced reports.

(a) "`John is a diabetic male white 74-year-old hypertensive patient with a red swollen mass in the left groin.`"

(b) "`John is a 74-year-old hypertensive diabetic white male patient with a swollen red mass in the left groin.`"

Figure 1: (a) Output of the generator without our ordering module, containing several errors. (b) Output of the generator with our ordering module.

Lexicalization is inherently domain dependent, so traditional lexica cannot be ported across domains without major modifications. Our approach, in contrast, is based on words extracted from a domain corpus and not on concepts, therefore it can be easily applied to new domains. In our MAGIC system, aggregation operators, such as conjunction, ellipsis, and transformations of clauses to adjectival phrases and relative clauses, are performed to combine related clauses together and increase conciseness [Shaw 1998a; Shaw 1998b]. We wrote a function, *reorder_premod(...)*, which is called after the aggregation operators, takes the whole lexicalized semantic representation, and reorders the premodifiers right before the linguistic realizer is invoked. Figure 1 shows the difference in the output produced by our generator with and without the ordering component.

7 Conclusions and Future Work

We have presented three techniques for exploring prior corpus evidence in predicting the order of premodifiers within noun phrases. Our methods expand on observable data, by inferring new relationships between premodifiers even for combinations of premodifiers that do not occur in the training corpus. We have empirically validated our approach, showing that we can predict order with more than 94% accuracy when enough corpus data is available. We have also implemented our procedure in a text generator, producing more fluent output sentences.

We are currently exploring alternative ways to integrate the classes constructed by the third stage of our system into our generator. In the future, we will experiment with semantic (rather than positional) clustering of premodifiers, using techniques such as those proposed in [Hatzivassiloglou and McKeown 1993; Pereira *et al.* 1993]. The qualitative analysis of the output of our clustering module shows that frequently positional and semantic classes overlap, and we are interested in measuring the extent of this phenomenon quantitatively. Conditioning the premodifier ordering on the head noun is another promising approach, at least for very frequent nouns.

8 Acknowledgments

We are grateful to Kathy McKeown for numerous discussions during the development of this work. The research is supported in part by the National Library of Medicine under grant R01-LM06593-01 and the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York State Science and Technology Foundation). Any opinions, findings, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the above agencies.

References

Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Massachusetts, 1974.

Carl Bache. *The Order of Premodifying Adjectives in Present-Day English.* Odense University Press, 1978.

John A. Bateman, Thomas Kamps, Jorg Kleinz, and Klaus Reichenberger. Communicative Goal-Driven NL Generation and Data-Driven Graphics Generation: An Architectural Synthesis for Multimedia Page Generation. In *Proceedings of the 9th International Workshop on Natural Language Generation.*, pages 8–17, 1998.

Eric Brill. A Simple Rule-Based Part of Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processsing*, Trento, Italy, 1992. Association for Computational Linguistics.

Mukesh Dalal, Steven K. Feiner, Kathleen R. McKeown, Desmond A. Jordan, Barry Allen, and Yasser al Safadi. MAGIC: An Experimental System for Generating Multimedia Briefings about Post-Bypass Patient Status. In *Proceedings of the 1996 Annual Fall Symposium of the American Medical Informatics Association (AMIA-96)*, pages 684–688, Washington, D.C., October 26–30 1996.

R. M. W. Dixon. *Where Have All the Adjectives Gone?* Mouton, New York, 1982.

William Frawley. *Linguistic Semantics.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

D. L. Goyvaerts. An Introductory Study on the Ordering of a String of Adjectives in Present-Day English. *Philologica Pragensia*, **11**:12–28, 1968.

Ralph Grishman, Catherine Macleod, and Adam Meyers. COMLEX Syntax: Building a Computational Lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, Kyoto, Japan, 1994.

Vasileios Hatzivassiloglou and Kathleen McKeown. Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 172–

182, Columbus, Ohio, June 1993. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 174–181, Madrid, Spain, July 1997. Association for Computational Linguistics.

John S. Justeson and Slava M. Katz. Co-occurrences of Antonymous Adjectives and Their Contexts. *Computational Linguistics*, **17**(1):1–19, 1991.

J. A. W. Kamp. Two Theories of Adjectives. In E. L. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press, Cambridge, England, 1975.

Maurice G. Kendall. A New Measure of Rank Correlation. *Biometrika*, **30**(1–2):81–93, June 1938.

Kevin Knight and Vasileios Hatzivassiloglou. Two-Level, Many-Paths Generation. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 252–260, Boston, Massachusetts, June 1995. Association for Computational Linguistics.

Irene Langkilde and Kevin Knight. Generation that Exploits Corpus-Based Statistical Knowledge. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (ACL/COLING-98)*, pages 704–710, Montreal, Canada, 1998.

Yakov Malkiel. Studies in Irreversible Binomials. *Lingua*, **8**(2):113–160, May 1959. Reprinted in [Malkiel 1968].

Yakov Malkiel. *Essays on Linguistic Themes.* Blackwell, Oxford, 1968.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**:313–330, 1993.

J. E. Martin. Adjective Order and Juncture. *Journal of Verbal Learning and Verbal Behavior*, **9**:379–384, 1970.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-Line Lexical Database. *International Journal of Lexicography (special issue)*, **3**(4):235–312, 1990.

Fernando C. N. Pereira and Michael D. Riley. Speech Recognition by Composition of Weighted Finite Automata. In Emmanuel Roche and Yves Schabes, editors, *Finite-State Language Processing*, pages 431–453. MIT Press, Cambridge, Massachusetts, 1997.

Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional Clustering of English Words. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 183–190, Columbus, Ohio, June 1993. Association for Computational Linguistics.

Randolph Quirk and Sidney Greenbaum. *A Concise Grammar of Contemporary English.* Harcourt Brace Jovanovich, Inc., London, 1973.

Jeffrey C. Reynar and Adwait Ratnaparkhi. A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Proc. of the 5th Applied Natural Language Conference (ANLP-97)*, Washington, D.C., April 1997.

James Shaw. Clause Aggregation Using Linguistic Knowledge. In *Proceedings of the 9th International Workshop on Natural Language Generation.*, pages 138–147, 1998.

James Shaw. Segregatory Coordination and Ellipsis in Text Generation. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (ACL/COLING-98)*, pages 1220–1226, Montreal, Canada, 1998.

Helmuth Späth. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples.* Ellis Horwood, Chichester, England, 1985.

J. Teyssier. Notes on the Syntax of the Adjective in Modern English. *Behavioral Science*, **20**:225–249, 1968.

Zeno Vendler. *Adjectives and Nominalizations.* Mouton and Co., The Netherlands, 1968.

Benjamin Lee Whorf. *Language, Thought, and Reality; Selected Writings.* MIT Press, Cambridge, Massachusetts, 1956.