Exciting and interesting: issues in the generation of binomials

Ann Copestake
Computer Laboratory,
University of Cambridge,
15 JJ Thomson Avenue,
Cambridge, CB3 0FD, UK
ann.copestake@cl.cam.ac.uk

Aurélie Herbelot
Institut für Linguistik,
Universität Potsdam,
Karl-Liebknecht-Straße 24-25
D-14476 Golm, Germany
herbelot@uni-potsdam.de

Abstract

We discuss the preferred ordering of elements of binomials (e.g., conjunctions such as *fish and chips*, *lager and lime*, *exciting and interesting*) and provide a detailed critique of Benor and Levy's probabilistic account of English binomials. In particular, we discuss the extent to which their approach is suitable as a model of language generation. We describe resources we have developed for the investigation of binomials using a combination of parsed corpora and very large unparsed corpora. We discuss the use of these resources in developing models of binomial ordering, concentrating in particular on the evaluation issues which arise.

1 Introduction

Phrases such as *exciting and interesting* and *gin and tonic* (referred to in the linguistics literature as **binomials**) are generally described as having a semantics which makes the ordering of the conjuncts irrelevant. For instance, *exciting and interesting* might correspond to $\text{exciting}'(x) \wedge \text{interesting}'(x)$ which is identical in meaning to $\text{interesting}'(x) \wedge \text{exciting}'(x)$. However, in many cases, the binomial is realized with a preferred ordering, and in some cases this preference is so strong that the reverse is perceived as highly marked and may even be difficult to understand. For example, *tonic and gin* has a corpus frequency which is a very small fraction of that of *gin and tonic*. Such cases are referred to as **irreversible binomials**, although the term is sometimes used only for the fully lexicalised, non-compositional examples, such as *odds and ends*.

Of course, realization techniques that utilize very large corpora to decide on word ordering will tend to get the correct ordering for such phrases if they have been seen sufficiently frequently in the training data. But the phenomenon is nevertheless of some practical interest because rare and newly-coined phrases can still demonstrate a strong ordering preference. For instance, the ordering found in the names of mixed drinks, where the alcoholic component comes first, applies not just to the conventional examples such as *gin and tonic*, but also to *brandy and coke*, *lager and lime*, *sake and grapefruit* and (hopefully) unseen combinations such as *armagnac and blackcurrant*.[1] A second issue is that data from an unparsed corpus can be misleading in deciding on binomial order. Furthermore, our own interest is predominantly in developing plausible computational models of human language generation, and from this perspective, using data from extremely large corpora to train a model is unrealistic. Binomials are a particularly interesting construction to look at because they raise two important questions: (1) to what extent does lexicalisation/establishment of phrases play a role in determining order? and (2) is a detailed lexical semantic classification required to accurately predict order?

As far as we are aware, the problem of developing a model of binomial ordering for language generation has not previously been addressed. However, Benor and Levy (2006) have published an important and detailed paper on binomial ordering which we draw on extensively in this work. Their research has the objective of determining how the various constraints which have been proposed in the linguistic literature might interact to determine bino-

[1] One of our reviewers very helpfully consulted a bartender about this generalization, and reports the hypothesis that the alcohol always comes first because it is poured first. However, there is the counter-example *gin and bitters* (another name for pink gin), where the bitters are added first (unless the drink is made in a cocktail shaker, in which case ordering is irrelevant).

Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop, pages 45–53,
Edinburgh, Scotland, UK, July 31, 2011. ©2011 Association for Computational Linguistics

mial ordering as observed in a corpus. We present a critical evaluation of that work here, in terms of the somewhat different requirements for a model for language generation.

The issues that we concentrate on in this paper are necessary preliminaries to constructing corpus-based models of binomial reversibility and ordering. These are:

1. Building a suitable corpus of binomials.
2. Developing a corpus-based technique for evaluation.
3. Constructing an initial model to test the evaluation methodology.

In §2, we provide a brief overview of some of the factors affecting binomial ordering and discuss Benor and Levy's work in particular. §3 discusses evaluation issues and motivates some of the decisions we made in deciding on the resources we have developed, described in §4. §5 illustrates the evaluation of a simple model of binomial ordering.

2 Benor and Levy's account

We do not have space here for a proper discussion of the extensive literature on binomials, or indeed for a full discussion of Benor and Levy's paper (henceforth B+L) but instead summarise the aspects which are most important for the current work.

For convenience, we follow B+L in referring to the elements of an ordered binomial as A and B. They only consider binomials of the form 'A and B' where A and B are of the same syntactic category. Personal proper names were excluded from their analysis. Because they required tagged data, they used a combination of Switchboard, Brown and the Wall Street Journal portion of the Penn Treebank to extract binomials, selecting 411 binomial types and all of the corresponding tokens (692 instances).

B+L investigate a considerable number of constraints on binomial ordering which have been discussed in the linguistics literature. They group the features they use into 4 classes: semantic, word frequency, metrical and non-metrical phonological. We will not discuss the last class here, since they found little evidence that it was relevant once the other features were taken into account. The metrical constraints were **lapse** (2 consecutive weak syllables are generally avoided), **length** (A should not have more syllables than B) and **stress** (B should not have ultimate (primary) stress: this feature was actually found to overlap almost entirely with lapse and length). The frequency constraint is that B should not be more frequent than A, based on corpus specific counts of frequency (unsurprisingly, frequency correlates with the length feature).

The semantic constraints are less straightforward since the linguistics literature has discussed many constraints and a variety of possible generalisations. B+L use:

Markedness Divided into **Relative formal**, which includes cases like *flowers and roses* (more general term first) among others and **Perception-based**, which is determined by extra-linguistic knowledge, including cases like *see and hear* (seeing is more salient). B should not be less marked than A. Unfortunately markedness is too complex to summarise adequately here. It is clear that it overlaps with other constraints in some cases, including frequency, since unmarked terms tend to be more frequent.

Iconicity Sequence ordering of events, numbered entities and so on (e.g., *shot and killed*, *eighth and ninth*). If there is such a sequence, the binomial ordering should mirror it.

Power Power includes gender relationships (discussed below), hierarchical relationships (e.g., *clergymen and parishioners*), the 'condiment rule' (e.g., *fish and chips*) and so on. B should not be more powerful than A.

Set Open Construction This is used for certain conventional cases where a given A may occur with multiple Bs: e.g., *nice and*.

Pragmatic A miscellaneous context-dependent constraint, used, for instance, where the binomial ordering mirrors the ordering of other words in the sentence.

B+L looked at the binomials in sentential context to assign the semantic constraints. The iconicity constraint, in particular, is context-dependent. For example, although the sequence *ninth and eighth* looks as though it violates iconicity, we found that a Google search reveals a substantial number of instances, many of which refer to the ninth and eighth centuries BC. In this case, iconicity is actually observed, if we assume that temporal ordering determines the constraint, rather than the ordering of the ordinals.

The aspect of binomials which has received most attention in the literature is the effect of gender: words which refer to (human) males tend to precede those referring to females. For instance (with Google 3-gram percentages for binomials with the masculine term first): *men and women* (85%), *boys and girls* (80%), *male and female* (91%) (exceptions are *father and mother* (51%) and *mothers and fathers* (33%)). There is also an observed bias towards predominantly male names preceding female names. B+L, following previous authors, take gender as an example of the Power feature. For reasons of space we can only touch on this issue very superficially, but it illustrates a distinction between semantic features which we think important. Iconicity generally refers to a sequence of real world events or entities occuring in a particular order, hence its context-dependence. For verbs, at least, there is a truth conditional effect of the ordering of the binomial: *shot and killed* does not mean the same thing as *killed and shot*. Power, on the other hand, is supposed to be about a conventional relationship between the entities. Even if we are currently more interested in chips rather than fish or biscuits rather than tea, we will still tend to refer to *fish and chips* and *tea and biscuits*. The actual ordering may depend on culture,[2] but the assumption is that, within a particular community, the power relationship which the binomial ordering depends on is fixed.

B+L analyse the effects of all the features in detail, and look at a range of models for combining features, with logistic regression being the most successful. This predicts the ordering of 79.2% of the binomial tokens and 76.7% of the types. When semantic constraints apply, they tend to outrank the metrical constraints. B+L found that iconicity, in particular, is a very strong predictor of binomial order.

B+L's stated assumption is that a speaker/writer knows they want to generate a binomial with the words A and B and decides on the order based on the words and the context. It is this order that they are trying to predict. Of course, it is clear that some binomials are non-compositional multiword expressions (e.g., *odds and ends*) which are listed in conventional dictionaries. These can be thought of as 'words with spaces' and, we would argue that the speaker does not have a choice of ordering in such cases. B+L argue that using a model which listed the fixed phrases would be valid in the prediction of binomial tokens, but not binomial types. We do not think this holds in general and return to the issue in §3.

B+L's work is important in being the first account which examines the effect of the postulated constraints in combination. However, from our perspective (which is of course quite different from theirs), there are a number of potential problems. The first is data sparsity: the vast majority of binomial types in their data occur only once. It is impossible to know whether both orderings are frequent for most types. Furthermore, the number of binomial types is rather small for full investigation of semantic features: e.g., Power is marked on only 26 types. The second issue is that the combined models which B+L examine are, in effect, partially trained on the test data, in that the relative contribution of the various factors is optimized on the test data itself. Thirdly, the semantic factors which B+L consider have no independent verification: they were assigned by the authors for the binomials under consideration, a methodology which makes it impossible to avoid the possibility of bias. There was some control over this, in that it was done independently by the two authors with subsequent discussion to resolve disagreements. However, we think that it would be hard to avoid the possibility of bias in the 'Set open' and 'Pragmatic' constraints in particular. Some of the choices seem unintuitive: e.g., we are unsure why there is a Power annotation on *broccoli and cauliflower*, and why *go and vote* would be marked for Iconicity while *went and voted* is not. It seems to us that the definition of some of these semantic factors in the literature (markedness and power in particular) is suf-

[2] Our favourite example is an English-French parallel text where the order of *Queen Elizabeth and President Mitterand* is reversed in the French.

ficiently unclear for reproducible annotation of the type now expected in computational linguistics to be extremely difficult.

Both for practical and theoretical reasons, we are interested in investigating alternative models which rely on a corpus instead of explicit semantic features. Native speakers are aware of some lexicalised and established binomials (see (Sag et al, 2002) for a discussion of lexicalisation vs establishment in multiword expressions), and will tend to generate them in the familiar order. Instead of explicit features being learned for the unseen cases, we want to investigate the possible role of analogy to the known binomials. For instance, if *tea and biscuits* is known, *coffee and cake* might be generated in that ordering by semantic analogy. The work presented in this paper is essentially preparatory to such experiments, although we will discuss an extremely simple corpus-based model in §5.

3 Evaluating models of binomial ordering

In this section, we discuss what models of binomial ordering should predict and how we might evaluate those predictions.

The first question is to decide precisely what we are attempting to model. B+L take the position that the speaker/writer has in mind the two words of the binomial and chooses to generate them in one order or other in a particular context, but this seems problematic for the irreversible binomials and, in any case, is not directly testable. Alternatively we can ask: Given a corpus of sentences where the binomials have been replaced with unordered pairs of AB, can we generate the ordering actually found? Both of these are essentially token-based evaluations, although we could additionally count binomial types, as B+L do.

One problem with these formulations is that, to do them justice, our models would really have to incorporate features from the surrounding context. Factors such as postmodification of the binomial affect the ordering. This type of evaluation would clearly be the right one if we had a model of binomials incorporated into a general realisation model, but it is not clear it is suitable for looking at binomials in isolation.

Perhaps more importantly, to model the irreversible or semi-irreversible binomials, we should take into account the order and degree of reversibility of particular binomial types. It seems problematic to formulate the generation of a lexicalised binomial, such as *odds and ends*, as a process of deciding on the order of the components, since the speaker must have the term in mind as a unit. In terms of the corpus formulation, given the pair AB, the first question in deciding how to realise the phrase is whether the order is actually fixed. The case of established but compositional binomials, such as *fish and chips*, is slightly less clear, but there still seem good grounds for regarding it as a unit (Cruse, 1986). Furthermore, in evaluating a token-based realisation model, we should not penalise the wrong ordering of a reversible binomial as severely as if the binomial were irreversible. From these perspectives, developing a model of ordering of binomial types should be a preliminary to developing a model of binomial tokens. Context would be important in properly modelling the iconicity effect, but is less of an issue for the other ordering constraints. And even though iconicity is context-dependent, there is a very strongly preferred ordering for many of the binomial types where iconicity is relevant.

Thus we argue that it is appropriate to look at the question: Given two words A, B which can be conjoined, what order do we find most frequently in a corpus? Or, in order to look at degree of reversibility: What proportion of the two orderings do we find in a corpus? This means that we require relatively large corpora to obtain good estimates in order to evaluate a model.

Of course, if we are interested in analogical models of binomial ordering, as mentioned at the end of §2, we need a reasonably large corpus of binomials to develop the model. Ideally this should be a different corpus from the one used for evaluation. We note that some experiments on premodifier ordering have found a considerable drop in performance when testing on a different domain (Shaw and Hatzivassiloglou, 1999). Using a single corpus split into training and test data would, of course, be problematic when working with binomial types. We have thus developed a relatively novel methodology of using an automatically parsed corpus in combination with frequencies from Web data. This is discussed in the next section.

4 Binomial corpora and corpus investigation

In this section, we describe the resources we have developed for investigating binomials and addressing some of the evaluation questions introduced in the previous section. We then present an initial analysis of some of the corpus data.

4.1 Benor and Levy data

The appendix of B+L's paper[3] contains a list of the binomials they looked at, plus some of their markup. Although the size of the B+L dataset is too small for many purposes, we found it useful to consider it as a clean source of binomial types for our initial corpus investigation and evaluation. We produced a version of this list excluding the 10 capitalised examples: some of these seem to arise from sentence initial capitals while others are proper names which we decided to exclude from this study. We produced a manually lemmatised version of the list, which results in a slightly reduced number of binomial types: e.g., *bought and sold* and *buy and sell* correspond to a single type. The issue of lemmatisation is slightly problematic in that a few examples are lexicalised with particular inflections, such as *been and gone*. However, our use of parsed data meant that we had to use lemmatization decisions which were compatible with the parser.

4.2 Wikipedia and the Google n-gram corpus

In line with B+L, we assume that binomials are made of two conjuncts with the same part of speech. It is not possible to use an unparsed corpus to extract such constructions automatically: first, the raw text surrounding a conjunction may not correspond to the actual elements of the coordination (e.g., the trigram *dictionary and phrase* in *She bought a dictionary and phrase book*); second, the part of speech information is not available. Using a parsed corpus, however, has disadvantages: in particular, it limits the amount of data available and, consequently, the number of times that a given type can be observed. In this section, we discuss the use of Wikipedia, which is small enough for parsing to be tractable but which turns out to have a fairly representative distribution of binomials. The latter point is demonstrated by comparison with a large dataset: the Google n-gram corpus (Brants and Franz, 2006). Although the Google data is not suitable for the actual task of extracting binomials, because it is not parsed, we hypothesize it is usable to predict the preferred order of a given binomial and to estimate the extent to which it is reversible.

In order to build a corpus of binomials, we process the parsed Wikipedia dump produced by Kummerfeld et al (2010). The parse consists of grammatical relations of the following form:

$$(gr\ word_1\_x\ word_2\_y\ ...\ word_n\_z)$$

where gr is the name of the grammatical relation, $word_{1...n}$ are the arguments of the relation, and $x, y...z$ are the positions of the arguments in the sentence. The lemmatised forms of the arguments, as well as their part of speech, are available separately.

We used the first one million *and* coordinations in the corpus in these experiments. The conjuncts are required to have the same part of speech and to directly precede and follow the coordination. The latter requirement ensures that we retrieve true binomials (phrases, as opposed to distant coordinates). For each binomial in this data, we record a frequency and whether it is found in the reverse order in the same dataset. The frequency of the reverse ordering is similarly collected. Since we intend to compare the Wikipedia data to a larger, unparsed corpus, we merge the counts of all possible parts of speech for a given type in a given ordering, so the counts for *European and American* as nouns and as adjectives, for instance, are added together. We also record the preferred ordering (the one with the highest frequency) of the binomial and the ratio of the frequencies as an indication of (ir)reversibility. In line with our treatment of the B+L data, we disregarded the binomials that coordinate proper names, but noted that a large proportion of proper names found in the Wikipedia data cannot be found in the Google data.[4] The Google corpus also splits (most) hyphen-

[3] http://idiom.ucsd.edu/~rlevy/papers/binomials-sem-alpha-formatted

[4] This suggests that the Google n-gram corpus does not contain much (if any) of the Wikipedia data: the particular dump of Wikipedia from which the parsed data is extracted being in any case several years later than the date that the Google n-gram corpus was produced.

ated words. Since hyphenation is notoriously irregular in English, we disregarded all binomials containing hyphenated words. The resulting data contains 279136 unique binomial types. Around 7600 of those types have a frequency of 10 or more in our Wikipedia subset. As expected, this leaves a large amount of data with low frequency.

We then attempt to verify how close the sparse Wikipedia data is to the Google 3-gram corpus. For each binomial obtained from Wikipedia, we retrieve the frequency of both its orderings in the Google data and, as before, calculate the ratio of the frequencies in the larger corpus. The procedure involves converting the lemmatised forms in the Wikipedia parse back into surface forms. Rather than using a morphological generator, which would introduce noise in our data, we search for the surface forms as they appeared in the original Wikipedia data, as well as for the coordinated base forms (this ensures high recall in cases where the original frequency is low). So for example, given the one instance of the binomial 'sadden and anger' in Wikipedia, appearing as *Saddened and angered* in the corpus, we search for *Saddened and angered*, *sadden and anger* and *anger and sadden*.

Around 30% of the Wikipedia binomials are not in the Google data. We manually spot checked a number of those and confirmed that they were unavailable from the Google data, regardless of inflection. Examples of binomials not found in the n-gram corpus include *dagger and saber*, *sagacious and firm* and (rather surprisingly) *gay and flamboyant*. 19% of the Wikipedia binomials have a different preferred order in the Google corpus. As expected, most of those have a low frequency in Wikipedia. For the binomials with an occurrence count over 40, the agreement on ordering is high (around 96%). Furthermore, many of those disagreements are not 'real' in that they concern binomials found with a high dispreferred to preferred order ratio. Disregarding cases where this ratio is over 0.3 lowers the initial disagreement figure to 7%. We will argue in §4.4 that true irreversibility can be shown to roughly correspond to a ratio of 0.1. At this cutoff, the percentage of disagreements between the two corpora is only 2%. Thus we found no evidence that the encyclopaedic nature of Wikipedia has a significant skewing effect on the frequencies. We thus believe that Wikipedia is a suitable dataset for training an automatic binomial ordering system.

4.3 Lexicalisation

Our basic methodology for investigation of lexicalisation was to check online dictionaries for the phrases. However, deciding whether a binomial should be regarded as a fixed phrase is not entirely straightforward. For instance, consider *warm and fuzzy*. At first sight, it might appear compositional, but the particular use of *fuzzy*, referring to feelings, is not the usual one. While *warm and fuzzy* is not listed in most dictionaries we have examined, it has an entry in the *Urban Dictionary*[5] and is used in examples illustrating that particular usage of *fuzzy* in the online Merriam-Webster.[6] Another case from the B+L data is *nice and toasty*, which again is used in a Merriam-Webster example.[7]

We therefore used a manual search procedure to check for lexicalisation of the B+L binomials. We used a broad notion of lexicalisation, treating a phrase as lexicalised if it occurred as an entry in one or more online English dictionaries using Google search. We included a few phrases as semi-lexicalised when they were given in examples in dictionaries produced by professional lexicographers, but this was, to some extent, a subjective decision. Since such a search is time-consuming, we only checked examples which one of us (a native British English speaker) intuitively considered might be lexicalised. We first validated that this would not cause too great a loss of recall by checking a small subset of the B+L data exhaustively: this did not reveal any additional examples.

Using these criteria, we found 39 lexicalised binomial types in the B+L data, of which 7 were semi-lexicalised.[8] The phrases *backwards and forwards*, *backward and forward*, *day and night*, *salt and pepper* and *in and out* are lexicalised (or semi-lexicalised) in both orders.

[5] `http://www.urbandictionary.com/`

[6] `http://www.merriam-webster.com/`

[7] The convention of indicating semi-fixed phrases in examples is quite common in lexicography, especially in dictionaries intended for language learners.

[8] There are 40 tokens, because *cut and dry* and *cut and dried* are both lexicalised. An additional example, *foot-loose and fancy-free*, might be included, but we did not find it in any dictionary with that hyphenation.

4.4 Reversibility and corpus evidence

There are a number of possible reasons why a particular binomial type AB might (almost) always appear in one ordering (*A and B* or *B and A*):

1. The phrase *A and B* (*B and A*) might be fully lexicalised (word with spaces).

2. The binomial might have a compositional meaning, but have a conventional ordering. A particular binomial AB might be established with that ordering (e.g., *gin and tonic* is established for most British and American speakers) or might belong to a conventional pattern (e.g., *armagnac and blackcurrant*, *sole and artichokes*).

3. The binomial could refer to a sequence of real world events or entities which almost invariably occur in a particular order. For example, *shot and killed* has a frequency of 241675 in the Google 3-gram corpus, as opposed to 158 for *killed and shot*. This ratio is larger that that of many of the lexicalised binomials.

Relatively few of the binomials from the B+L data are completely irreversible according to the Google 3-gram data. There are instances of the reverse of even obviously fixed phrases, such as *odds and ends*. Of course, there is no available context in the 3-gram data, but we investigated some of these cases by online search for the reversed phrases. This indicates a variety of sources of noise, including wordplay (e.g., Beckett's play *Ends and Odds*), different word senses (e.g., *toasty and nice* occurs when *toasty* is used to describe wine) and false positives from hyphenated words etc.

We can obtain a crude estimate of extent to which binomials which should be irreversible actually turn up in the 'wrong' order by looking at the clearly lexicalised phrases discussed in §4.3. Excluding the cases where both orders are lexicalised, the mean proportion of inverted cases is about 3%. There are a few outliers, such as *there and back* and *now and then* which have more than 10% inverted: however, these all involve very frequent closed class words which are more likely to show up in spurious contexts. We therefore tentatively conclude that up to 10% of the tokens of a open-class irreversible binomial could be inverted in the 3-gram corpus, but that we can take higher ratios as evidence for a degree of genuine reversibility.

5 An initial model

We developed an initial n-gram-based model for ordering using the Wikipedia-derived counts. The approach is very similar to that presented in (Malouf, 2000) for adjective ordering. We use the observed order of binomials where possible and back off to counts of a lexeme's position as first or second conjunct over all binomials (i.e., we use what Malouf refers to as **positional probabilities**).

To be more precise, assume that the task is to predict the order $a \prec b$ or $b \prec a$ for a given lexeme pair a,b. We use the notation $C(\text{a and b})$ and $C(\text{b and a})$ to refer to the counts in a given corpus of the two orderings of the binomial (i.e., we count all inflections of a and b). $C(\text{a and})$ refers to the count of all binomials with the lexeme a as the first conjunct, $C(\text{and a})$ all binomials with a as the second conjunct, and so on. We predict $a \prec b$

$$\begin{array}{ll} \text{if} & C(\text{a and b}) > C(\text{b and a}) \\ \text{or} & C(\text{a and b}) = C(\text{a and b}) \\ & \ \text{and} \\ & C(\text{a and})C(\text{and b}) > C(\text{b and})C(\text{and a}) \end{array}$$

and conversely for $b \prec a$. Most of the cases where the condition $C(\text{a and b}) = C(\text{a and b})$ is true occur when $C(\text{a and b}) = C(\text{a and b}) = 0$ but we also use the positional probabilities to break ties in the counts. We could, of course, define this in terms of probability estimates and investigate various forms of smoothing and interpolation, but for our initial purposes it is adequate to see how this very simple model behaves.

We obtained counts for the model from the Wikipedia-derived data and evaluated it on the binomial types derived from B+L (as described in §4.1). There were only 9 cases where there was no prediction, so for the sake of simplicity, we default to alphabetic ordering in those cases. In Table 1, we show the results evaluating against the B+L majority decision and against the Google 3-gram majority. Because not all the B+L binomials are found in the Google data, the numbers of binomial types evaluated against the Google data is slightly lower. In addition to the overall figures, we also show the relative accuracy of the bigram prediction vs the backoff and the different accuracies on the lexicalised and non-lexicalised data. In Table 2, we group the results according to the ratio of the less frequent order in the Google data and by frequency.

Unsurprisingly, performance on more frequent binomials and lexicalised binomials is better and the bigram performance, where available, is better than the backoff to positional probabilities. The scores when evaluated on the Google corpus are generally higher than those on the B+L counts, as expected given the noise created by the data sparsity in B+L combined with the effect of frequency.

One outcome from our experiments is that it does not seem essential to treat the lexicalised examples separately from the high frequency, low reversibility cases. Since determining lexicalisation is time-consuming and error-prone, this is a useful result.

The model described does not predict whether or not a given binomial is irreversible, but our analysis of the data strongly suggests that this would be important in developing more realistic models. An obvious extension would be to generate probability estimates of orderings and to compare these with the observed Google 3-gram data.

Although n-gram models are completely standard in computational linguistics, their applicability to modelling human performance on a task is not straightforward. Minimally, if we were to propose that humans were using such a model as part of their decision on binomial ordering, it would be necessary to demonstrate that the counts we are relying on correspond to data which it is plausible to assume that a human could have been exposed to. This is not a trivial consideration. We would, of course, expect to obtain higher scores on this task by using counts derived from the Google n-gram corpus rather than from Wikipedia, but this would be completely unrealistic from a psycholinguistic perspective. We should emphasize, therefore, that the model presented here is simply intended as an initial exercise in developing distributional models of binomial ordering, which allows us to check whether the resources we have developed might be an adequate basis for more serious modelling and whether the evaluation schemes are reasonable.

6 Conclusion

We have demonstrated that we can make use of a combination of corpora to build resources for development and evaluation of models of binomial ordering.[9] One novel aspect is our use of an automatically parsed corpus, another is the use of combined corpora. If binomial ordering is primarily determined by universal linguistic factors, we would not expect the relative frequency to differ very substantially between large corpora. The cases where we did observe differences in preferred ordering between the Wikipedia and Google data are predominantly ones where the Wikipedia frequency is low or the binomial is highly reversible. We have investigated several properties of binomials using this data and produced a simple initial model. We tested this on the relatively small number of binomials used by Benor and Levy (2006), but in future work we will evaluate on a much larger subset of our corpus. Our intention is to develop further models which use analogy (morphological and distributional semantic similarity) to known binomials to predict degree of reversibility and ordering. This will allow us to investigate whether human performance can be modelled without the use of explicit semantic features.

We briefly touched on Malouf's (2000) work on prenominal adjective ordering in our discussion of the initial model. There are some similarities between these tasks, and in fact adjectives in binomials tend to occur in the same order when they appear as prenominal adjectives (e.g., *cold and wet* and *cold wet* are preferred over the inverse orders). However, the binomial problem is considerably more complex. Binomials are much more variable because they involve all the main syntactic categories. Furthermore, adjective ordering is considerably easier to investigate because an unparsed corpus can be used, the semantic features which have been postulated are more straightforward than for binomials and lexicalisation of adjective sequences is not an issue. We hypothesize that it should be possible to develop similar analogical models for adjective ordering and binomials which could be relevant for other constructions where ordering is only partially determined by syntax. In the long term, we would like to in-

[9] Available from `http://www.cl.cam.ac.uk/research/nl/nl-download/binomials/`

	n B+L	n Google	accuracy B+L (%)	accuracy Google (%)
Overall	380	305	69	79
Bigram	187	185	79	89
Pos Prob	184	117	61	65
Unknown	9	3	33	0
Lexicalised	34	34	87	94
Non-lexicalised	346	271	67	77

Table 1: Evaluation of initial model, showing effects of lexicalisation. (n B+L and n Google indicates the number of binomial types evaluated)

		n	accuracy B+L (%)	accuracy Google (%)
Google count	0	75	59	-
	1–1000	71	56	68
	1001–10000	81	70	67
	> 10000	153	80	91
Google ratio	0	11	64	64
	0–0.1	41	94	93
	0.1–0.25	33	75	85
	> 0.25	220	68	76

Table 2: Evaluation of initial model, showing effects of frequency and reversibility.

vestigate using such models in conjunction with a grammar-based realizer (cf (Velldal, 2007), (Cahill and Riester, 2009)). However, for an initial investigation of the role of semantics and lexicalisation, looking at the binomial construction in isolation is more tractable.

Acknowledgments

This work was partially supported by a fellowship to Aurélie Herbelot from the Alexander von Humboldt Foundation. We are grateful to the reviewers for their comments.

References

Sarah Benor and Roger Levy. 2006. *The Chicken or the Egg? A Probabilistic Analysis of English Binomials.* Language, **82** 233–78.

Thorsten Brants and Alex Franz. 2006. *The Google Web 1T 5-gram Corpus Version 1.1.* LDC2006T13.

Aoife Cahill and Arndt Riester. 2009. *Incorporating Information Status into Generation Ranking.* In Proceedings of the 47th Annual Meeting of the ACL, pp. 817-825, Suntec, Singapore. Association for Computational Linguistics.

D. Alan Cruse. 1986. Lexical Semantics. Cambridge University Press.

Jonathan K. Kummerfeld, Jessika Rosener, Tim Dawborn, James Haggerty, James R. Curran, Stephen Clark. 2010. *Faster parsing by supertagger adaptation* Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pages 345–355

Rob Malouf. 2000. The order of prenominal adjectives in natural language generation. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong.

Ivan Sag, Tim Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. *Multiword expressions: A pain in the neck for NLP.* In Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), pages 1–15, Mexico City, Mexico.

James Shaw and Vasileios Hatzivassiloglou. 1999. *Ordering among premodifiers.* In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 135–143, College Park, Maryland.

Eric Velldal. 2007. *Empirical Realization Ranking.* Ph.D. thesis, University of Oslo, Department of Informatics.