Automatically Identifying the Source Words of Lexical Blends in English

Paul Cook*
University of Toronto

Suzanne Stevenson**
University of Toronto

Newly coined words pose problems for natural language processing systems because they are not in a system's lexicon, and therefore no lexical information is available for such words. A common way to form new words is lexical blending, as in cosmeceutical*, a blend of* cosmetic *and* pharmaceutical*. We propose a statistical model for inferring a blend's source words drawing on observed linguistic properties of blends; these properties are largely based on the recognizability of the source words in a blend. We annotate a set of 1,186 recently coined expressions which includes 515 blends, and evaluate our methods on a 324-item subset. In this first study of novel blends we achieve an accuracy of 40% on the task of inferring a blend's source words, which corresponds to a reduction in error rate of 39% over an informed baseline. We also give preliminary results showing that our features for source word identification can be used to distinguish blends from other kinds of novel words.*

1. Lexical Blends

Neologisms—newly coined words or new senses of an existing word—are constantly being introduced into a language (Algeo 1980; Lehrer 2003), often for the purpose of naming a new concept. Domains that are culturally prominent or that are rapidly advancing, such as electronic communication and the Internet, often contain many neologisms, although novel words arise throughout a language (Ayto 1990, 2006; Knowles and Elliott 1997). Consequently, any natural language processing (NLP) system operating on recently produced text will encounter new words. Because lexical resources are often a key component of an NLP system, performance of the entire system will likely suffer due to missing lexical information for neologisms. Ideally, an NLP system could identify neologisms as such, and then infer various aspects of their syntactic or semantic properties necessary for the computational task at hand. Recent approaches to this kind of lexical acquisition task typically infer the target lexical information from statistical distributional properties of the terms. However, this technique is generally not

* Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4, Canada, E-mail: pcook@cs.toronto.edu.
** Department of Computer Science, University of Toronto, 6 King's College Rd., Toronto, ON M5S 3G4, Canada, E-mail: suzanne@cs.toronto.edu.

Submission received: 4 November 2008; revised submission received: 23 May 2009; accepted for publication: 24 June 2009.

applicable to neologisms, which are relatively infrequent due to their recent introduction into the language. Fortunately, linguistic observations regarding neologisms—namely, that they are formed through specific word formation processes—can give insights for automatically learning their lexical properties.

New words come about through a variety of means, including derivational morphology, compounding, and borrowing from another language (Algeo 1980; Bauer 1983; Plag 2003). Computational work on neologisms has largely focused on particular word formation processes, and has exploited information about the formation process to learn aspects of the semantic properties of words (Means 1988; Nadeau and Turney 2005; Baker and Brew 2008, for example). Subtractive word formations—words formed from partial orthographic or phonological content from existing words—have received a fair amount of attention recently in computational linguistics, particularly under the heading of inferring the long form of acronyms, especially in the bio-medical domain (e.g., Schwartz and Hearst 2003; Nadeau and Turney 2005; Okazaki and Ananiadou 2006, for example).

Lexical blends—the focus of this study—also known as blends, are another common type of subtractive word formation. Most blends are formed by combining a prefix of one source word with a suffix of another source word, as in *brunch* (*breakfast* and *lunch*). There may be overlap in the contribution of the source words, as in *fantabulous* (*fantastic* and *fabulous*). It is also possible that one or both source words are entirely present, for example, *gaydar* (*gay radar*) and *jetiquette* (*jet etiquette*). We refer to blends such as these as simple two-word sequential blends, and focus on this common type of blend in this article. Blends in which (part of) a word is inserted within another (e.g., *entertoyment*, a blend of *entertainment* and *toy*) and blends formed from more than two source words (e.g., *nofriendo* from *no*, *friend*, and *Nintendo*) are rare. In Algeo's (1991) study of new words, approximately 5% were blends. However, in our analysis of 1,186 words taken from a popular neologisms Web site, approximately 43% were blends. Clearly, computational techniques are needed that can augment lexicons with knowledge of novel blends.

The precise nature and intended use of a computational lexicon will determine the degree of processing required of a novel blend. In some cases it may suffice for the lexical entry for a blend to simply consist of its source words. For example, a system that employs a measure of distributional similarity may benefit from replacing occurrences of a blend—likely a recently coined and hence low frequency item—by its source words, for which distributional information is likely available. In other cases, further semantic reasoning about the blend and its source words may be required (e.g., determining the semantic relationship between the source words as an approximation of the meaning of the blend). However, any approach to handling blends will need to recognize that a novel word is a blend and identify its source words. These two tasks are the focus of this article. Specifically, we draw on linguistic knowledge of how blends are formed as the basis for automatically determining the source words of a blend. Language users create blends that tend to be interpretable by others. Tapping into properties of blends believed to contribute to the recognizability of their source words—and hence the interpretability of the resulting blend—we develop statistical measures that indicate whether a word pair is likely the source words for a given blend. Moreover, the fact that a novel word is determined to have a "good" source word pair may be evidence that it is in fact a blend, because we are unlikely to find two words that are a "good" source word pair for a non-blend. Thus, the statistical measures we develop for source word identification may also be useful in recognizing a novel word as a blend.

To our knowledge, the only computational treatment of blends is our earlier work that presents preliminary statistical methods and results for the two tasks of recognizing an unknown word as a blend and identifying its source words (Cook and Stevenson 2007). Here we extend that work in a number of important directions. We expand the statistical features to better capture co-occurrence patterns of the source words that can indicate the likelihood of their combination into a blend. We present experimental results confirming that the extended features provide a substantial improvement over the earlier work on source word identification. We further propose a means, based on linguistic factors of the source words, for pruning the number of word pairs that are considered for a blend. This filtering heuristic greatly reduces the number of candidate source words processed, while giving modest gains in performance, even though this method excludes the correct word pair from consideration for a number of blends. We then consider the use of a much larger lexicon of candidate source words, which could potentially improve performance greatly as it contains the correct source word pair for many more blends than a smaller lexicon.

We also make improvements to the earlier experimental data set and methods. In our earlier study, we use a data set consisting of a list of blends extracted from a dictionary. In the current work, we annotate a set of 324 blends (with their source words) from a recent database of neologisms, to enable a more legitimate testing of our method, on truly novel blends. Experiments on this new data set show that the recent blends differ from established blends in terms of their statistical properties, and emphasize the need for further resources of neologisms. We also experiment with a machine learning approach to combine the information from the statistical features in a more sophisticated manner than in our previous work. Finally, we perform more extensive experiments on distinguishing blends from other kinds of novel words.

2. A Statistical Model of Lexical Blends

We present statistical features that are used to automatically infer the source words of a word known to be a lexical blend, and show that the same features can be used to distinguish blends from other types of neologisms. First, given a blend, we generate all word pairs that could have formed the blend. This set is termed the candidate set, and the word pairs it contains are referred to as candidate pairs (Section 2.1). Next, we extract a number of linguistically motivated statistical features for each candidate pair, as well as filter from the candidate sets those pairs that are unlikely to be source words due to their linguistic properties (Section 2.2). Later, we explain how we use the features to rank the candidate pairs according to how likely they are the source words for that blend. Interestingly, the “goodness” of a candidate pair is also related to how likely the word is actually a blend.

2.1 Candidate Sets

To create the candidate set for a blend, we first generate each prefix–suffix pair such that the blend is composed of the prefix followed by the suffix. (In this work, *prefix* and *suffix* refer to the beginning or ending of a string, regardless of whether those portions are affixes.) We restrict the prefixes and suffixes to be of length two or more. This heuristic reduces the size of the candidate sets, yet generally does not exclude a blend’s source words from its candidate set since it is uncommon for a source word to contribute less than two letters. For example, for *brunch* (*breakfast+lunch*) we consider the following prefix–suffix pairs: *br*, *unch*; *bru*, *nch*; *brun*, *ch*. For each prefix–suffix pair, we then find in a lexicon all words beginning with the prefix and all words ending in the suffix, ignoring hyphens and whitespace, and take the Cartesian product of the prefix words and suffix words to form a list of candidate word pairs. The candidate set for the blend is the union of the candidate word pairs for all its prefix–suffix pairs. Note that in this example, the candidate pair *brute crunch* would be included twice: once for the prefix–suffix pair *br*, *unch*; and once again for *bru*, *nch*. Unlike in our previous study, we remove all such duplicate pairs from the final candidate set. A candidate set for *architourist*, a blend of *architecture* and *tourist*, is given in Table 1.

2.2 Statistical Features

Our statistical features are motivated by properties of blends observed in corpus-based studies, and by cognitive factors in human interpretation of blends, particularly relating to how easily humans can recognize a blend's source words. All the features are formulated to give higher values for more likely candidate pairs. We organize the features into four groups—frequency; length, contribution, and phonology; semantics; and syllable structure—and describe each feature group in the following subsections.

2.2.1 Frequency. Various frequency properties of the source words influence how easily a language user recognizes the words that form a blend. Because blends are most usefully coined when the source words can be readily deduced, we hypothesize that frequency-based features will be useful in identifying blends and their source words. We propose ten features that draw on the frequency of candidate source words.

Lehrer (2003) presents a study in which humans are asked to give the source words for blends. Among her findings are that frequent source words are more easily recognizable. Our first two features—the frequency of each candidate word, $freq(w_1)$ and $freq(w_2)$—reflect this finding. Lehrer also finds that the recognizability of a source word is further affected by both the number of words in its neighborhood—the set of words which begin/end with the prefix/suffix which that source word

Table 1
A candidate set for *architourist*, a blend of *architecture* and *tourist*.

archimandrite	tourist
archipelago	tourist
architect	behaviourist
architect	tourist
architectural	behaviourist
architectural	tourist
architecturally	behaviourist
architecturally	tourist
architecture	behaviourist
architecture	tourist
archives	tourist
archivist	tourist

contributes—and the frequencies of those words. (Gries [2006] reports a similar finding.) Our next two features capture this insight:

$$\frac{freq(w_1)}{freq(prefix)} \qquad \frac{freq(w_2)}{freq(suffix)} \tag{1}$$

where *freq*(*prefix*) is the sum of the frequency of all words beginning with *prefix*, and similarly for *freq*(*suffix*).

These four features were used in our previous study; the following six features are new in this study.

Because we observe that blends are often formed from two words that co-occur in language use, our previous study (Cook and Stevenson 2007) included a feature, the pointwise mutual information of w_1 and w_2, to reflect this. However, this feature provides only a weak indication that there is a semantic relation between two words sufficient to lead to them being blended. Here we propose six new features that capture various co-occurrence frequencies as follows.

A blend's source words often correspond to a common sequence of words, for example, *camouflanguage* is *camouflaged language*. We therefore include two features based on Dice's co-efficient to capture the frequency with which the source words occur consecutively:

$$\frac{2 \times freq(w_1\ w_2)}{freq(w_1) + freq(w_2)} \qquad \frac{2 \times freq(w_2\ w_1)}{freq(w_1) + freq(w_2)} \tag{2}$$

Because many blends can be paraphrased by a conjunctive phrase—for example, *broccoflower* is *broccoli and cauliflower*—we also use a feature that reflects how often the candidate words are used in this way:

$$\frac{2 \times (freq(w_1\ and\ w_2) + freq(w_2\ and\ w_1))}{freq(w_1\ and) + freq(and\ w_1) + freq(w_2\ and) + freq(and\ w_2)} \tag{3}$$

Furthermore, some blends can be paraphrased by a noun modified by a prepositional phrase, for example, a *nicotini* is a *martini with nicotine*. Lauer (1995) suggests eight prepositional paraphrases for identifying the semantic relationship between the modifier and head in a noun compound. Using the same paraphrases, the following feature measures how often two candidate source words occur with any of the following prepositions *P* between them: *about*, *at*, *for*, *from*, *in*, *of*, *on*, *with*:

$$\frac{2 \times (freq(w_1\ P\ w_2) + freq(w_2\ P\ w_1))}{freq(w_1\ P) + freq(P\ w_1) + freq(w_2\ P) + freq(P\ w_2)} \tag{4}$$

where *freq*(w P v) is the sum of the frequency of w and v occurring with each of the eight prepositions between w and v, and *freq*(w P) is the sum of the frequency of w occurring with each of the eight prepositions immediately following w.

Because the previous three features target the source words occurring in very specific patterns, we also count the candidate source words occurring in any of the patterns in an effort to avoid data sparseness problems.

$$\frac{\begin{array}{c}2 \times (\mathit{freq}(w_1\ w_2) + \mathit{freq}(w_2\ w_1) + \mathit{freq}(w_1\ \mathit{and}\ w_2) \\ + \mathit{freq}(w_2\ \mathit{and}\ w_1) + \mathit{freq}(w_1\ P\ w_2) + \mathit{freq}(w_2\ P\ w_1))\end{array}}{\mathit{freq}(w_1) + \mathit{freq}(w_2)} \tag{5}$$

Finally, because the above patterns are very specific, and do not capture general co-occurrence information which may also be useful in identifying a blend's source words, we include the following feature which counts the candidate source words co-occurring within a five-word window.

$$\frac{2 \times \mathit{freq}(w_1, w_2 \ \text{ in a 5 word window})}{\mathit{freq}(w_1) + \mathit{freq}(w_2)} \tag{6}$$

2.2.2 Length, Contribution, and Phonology. Ten features tap into properties of the orthographic or phonetic composition of the source words and blend. In our previous work on blends, we found such features unhelpful in source word identification. Here, we propose revised versions of our old features, and a few new ones. Note that although we use information about the phonological and/or syllabic structure of the source words, we do not assume such knowledge for the blend itself, since it is a neologism for which such lexical information is typically unavailable.

The first word in a conjunct tends to be shorter than the second, and this also seems to be the case for the source words in blends (Kelly 1998; Gries 2004). The first three features therefore capture this tendency based on the graphemic, phonemic, and syllabic length of w_2 relative to w_1, respectively:

$$\frac{\mathit{len}_{\mathit{graphemes}}(w_2)}{\mathit{len}_{\mathit{graphemes}}(w_1) + \mathit{len}_{\mathit{graphemes}}(w_2)} \tag{7}$$

$$\frac{\mathit{len}_{\mathit{phonemes}}(w_2)}{\mathit{len}_{\mathit{phonemes}}(w_1) + \mathit{len}_{\mathit{phonemes}}(w_2)} \tag{8}$$

$$\frac{\mathit{len}_{\mathit{syllables}}(w_2)}{\mathit{len}_{\mathit{syllables}}(w_1) + \mathit{len}_{\mathit{syllables}}(w_2)} \tag{9}$$

A blend and its second source word also tend to be similar in length, possibly because, similar to compounds, the second source word of a blend is often the head; therefore it is this word that determines the overall phonological structure of the resulting blend (Kubozono 1990). The following feature captures this property using graphemic length as an approximation to phonemic length, because as stated previously, we assume no phonological information about the blend.

$$1 - \frac{|len_{graphemes}(blend) - len_{graphemes}(w_2)|}{max(len_{graphemes}(blend), len_{graphemes}(w_2))} \quad (10)$$

We hypothesize that a candidate source word is more likely if it contributes more graphemes to a blend. We use two ways to measure contribution in terms of graphemes: $cont_{seq}(w, b)$ is the length of the longest prefix/suffix of word w which blend b begins/ends with, and $cont_{lcs}(w, b)$ is the longest common subsequence (LCS) of w and b. This yields four features:

$$\frac{cont_{seq}(w_1, b)}{len_{graphemes}(w_1)} \qquad \frac{cont_{seq}(w_2, b)}{len_{graphemes}(w_2)} \quad (11)$$

$$\frac{cont_{lcs}(w_1, b)}{len_{graphemes}(w_1)} \qquad \frac{cont_{lcs}(w_2, b)}{len_{graphemes}(w_2)} \quad (12)$$

Note that for some blends, such as *spamdex* (*spam index*), $cont_{seq}$ and $cont_{lcs}$ will be equal; however, this is not the case in general, as in the blend *tomacco* (*tomato* and *tobacco*) in which *tomato* overlaps with the blend not only in its prefix *toma*, but also in the final *o*.

In order to be recognizable in a blend, the shorter source word will tend to contribute more material, relative to its length, than the longer source word (Gries 2004). We formulate the following feature which is positive only when this is the case:

$$\left(\frac{cont_{seq}(w_1, b)}{len_{graphemes}(w_1)} - \frac{cont_{seq}(w_2, b)}{len_{graphemes}(w_2)}\right) \times \left(\frac{len_{graphemes}(w_2) - len_{graphemes}(w_1)}{len_{graphemes}(w_1) + len_{graphemes}(w_2)}\right) \quad (13)$$

For this feature we don't have strong motivation for choosing one measure of contribution over the other, and therefore use $cont_{seq}$, the simpler version of contribution.

Finally, the source words in a blend are often phonologically similar, as in *sheeple* (*sheep people*); the following feature captures this (Gries 2006):

$$LCS_{phonemes}(w_1, w_2) \quad (14)$$

2.2.3 Semantics. We include two semantic features from our previous study that are based on Lehrer's (2003) observation that people can more easily identify the source words of a blend when there is a semantic relation between them.

As noted, blends are often composed of two semantically similar words, reflecting a conjunction of their concepts. For example, a *pug* and a *beagle* are both a kind of dog, and can be combined to form the blend *puggle*. Similarly an *exergame* is a blend of *exercise* and *game*, both of which are types of activity. Our first semantic feature captures similarity using an ontological similarity measure, which is calculated over an ontology populated with word frequencies from a corpus.

The source words of some blends are not semantically similar (in the sense of their relative positions within an ontology), but are semantically related. For example, the source words of *slanguist*—*slang* and *linguist*—are related in that *slang* is a type of language and a *linguist* studies language. Our second semantic feature is a measure of semantic relatedness using distributional similarity between word co-occurrence vectors.

2.2.4 Syllable Structure. Kubozono (1990) notes that the split of a source word—into the prefix/suffix it contributes to the blend and the remainder of the word—occurs at a syllable boundary or immediately after the onset of the syllable. Because this syllable structure property holds sufficiently often, we use it as a filter over candidate pairs (rather than as an additional statistical feature) in an effort to reduce the size of the candidate sets. Candidate sets can be very large, and we expect that our features will be more successful at selecting the correct source word pair from a smaller candidate set. In our subsequent results, we analyze the reduction in candidate set size using this syllable structure heuristic, and its impact on performance.

3. Creating a Data Set of Recent Blends

The data set used in our previous work on blends contains dictionary words whose etymological entry indicates they were formed from a blend of two words. Using a dictionary in this way provides an objective method for selecting experimental expressions and indicating their gold standard source words. However, it results in a data set of blends that are sufficiently established in the language to appear in a dictionary. Truly novel blends—neologisms which have been recently added to the language—may have differing properties from fully established forms in a dictionary. In particular, many of our features are based on properties of the source words, both individually and in relation to each other, that may not hold for expressions that entered the language some time ago. For example, although *meld* is a blend of *melt* and *weld*, the current frequency of the phrase *melt and weld* may not be as common as the source word co-occurrences for newly coined expressions. Thus, an important step to support further research on blends is to develop a data set of recent neologisms that are judged to be lexical blends.

To develop a data set of recently coined blends we drew on `www.wordspy.com`, a popular Web site documenting English neologisms (and a small number of rare or specialized terms) that have been recently used in a recordable medium such as a newspaper or book, and that (typically) are not found in currently available dictionaries. A (partial) sample entry from Wordspy is given in Table 2. The words on this Web site satisfy our goal of being new; however, they include many kinds of neologisms, not just blends. We thus annotated the data set to identify the blends and their source

Table 2
The Wordspy definition, and first citation given, for the blend *staycation*.

staycation n. A stay-at-home vacation. Also: stay-cation.
—staycationer n.

Example Citation:
Amy and Adam Geurden of Hollandtown, Wis., had planned a long summer of short, fun getaways with their kids, Eric, 6, Holly, 3, and Jake, 2. In the works were water-park visits, roller-coaster rides, hiking adventures and a whirlwind weekend in Chicago. Then Amy did the math: their Chevy Suburban gets 17 miles to the gallon and, with gas prices topping $4, the family would have spent about $320 on fill-ups alone. They've since scrapped their plans in favor of a "staycation" around the backyard swimming pool.
—Linda Stern, "Try Freeloading Off Friends!," Newsweek, May 26, 2008

words. (In cases where multiple source words were found to be equally acceptable, all source words judged to be valid were included in the annotation.) Most expressions in Wordspy include both a definition and an example usage, making the task fairly straightforward.

As of 17 July 2008 Wordspy contained 1,186 single word entries. One author annotated each of these words as a blend or not a blend, and indicated the source words for each blend. To ensure validity of the annotation task, the other author similarly annotated 100 words randomly sampled from the 1,186. On this subset of 100 words, observed agreement on both the blend/non-blend annotation and the component source word identification was 92%, with an unweighted Kappa score of .84. On four blends, the judges gave different variants of the same source word; for example, *fuzzy buzzword* and *fuzz buzzword* for the blend *fuzzword*. These items were counted as agreements, and all variants were considered correct source words.

Given the high level of agreement between the annotators, only one person annotated all 1,186 items. A total of 515 words were judged to be blends, with 351 being simple two-word sequential blends whose source words are not proper nouns (this latter type of blend being the focus of this study). Table 3 shows the variety of blends encountered in the Wordspy data, organized according to a categorization scheme we devised. Of the simple two-word sequential blends, we restrict our experimental data set to the 324 items whose entries included a citation of their usage, as we have evidence that they have in fact been used; moreover, such blends may be less likely to be nonce formations—expressions which are used once but do not become part of the language. The usage data in the citations can also be used in the future for semantic features based on contextual information. We refer to this new data set of 324 items as WORDSPLEND (a blend of *Wordspy* and *blend*).

4. Materials and Methods

4.1 Experimental Expressions

The data set used in our previous study of blends consisted of expressions from the Macquarie Dictionary (Delbridge 1981) with an etymology entry indicating that they are blends. All of our statistical features were devised using the development portion of this data set, enabling us to use the full WORDSPLEND data set for testing. To compare our results to those in our earlier study, we also perform experiments on a subset of the previous data set. We are uncertain as to whether a number of the blends from the Macquarie Dictionary are in fact blends. For example, it does not match our intuition that *clash* is a blend of *clap* and *dash*. We created a second data set of confirmed blends, MAC-CONF, consisting of only those blends from Macquarie that are found in at least one of two additional dictionaries with an etymology entry indicating that they are blends. We report results on the 30 expressions in the unseen test portion of MAC-CONF.

4.2 Experimental Resources

We generate candidate sets using two different lexicons: the CELEX lexicon (Baayen, Piepenbrock, and Gulikers 1995),[1] and a wordlist created from the Web 1T 5-gram

[1] From CELEX, we use lemmas as potential source words, as it is uncommon for a source word to be an inflected form—there are no such examples in our development data.

Table 3
Types of blends and their frequency in Wordspy data.

Blend type	Freq.	Example
Simple two-word sequential blends	351	*digifeiter* (*digital counterfeiter*)
Proper nouns	50	*Japanimation* (*Japanese animation*)
Affixes	61	*prevenge* (*pre-revenge*)
Common one-letter prefix	10	*e-business* (*electronic business*)
Non-source word material	7	*aireoke* (*air guitar karaoke*)
w_2 contributes a prefix	10	*theocon* (*theological conservative*)
Foreign word	4	*sousveillance* (French *sous*, meaning under, and English *surveillance*)
Non-sequential blends	6	*entertoyment* (*entertainment* blended with *toy*)
w_1 contributes a suffix	5	*caponomics* (*salary cap economics*)
Multiple source words	6	MoSoSo (*mobile social software*)
Other	5	*CUV* (*car* blended with initialism *SUV*)

Corpus (Brants and Franz 2006). These are discussed further herein. The frequency information needed to calculate the frequency features is extracted from the Web 1T 5-gram Corpus. The length, contribution, and phonology features, as well as the syllable structure filter, are calculated on the basis of the source words themselves, or are derived from information in CELEX (when CELEX is the lexicon in use).[2] We compute semantic similarity between the source words using Jiang and Conrath's (1997) measure in the WordNet::Similarity package (Pedersen, Patwardhan, and Michelizzi 2004), and we compute semantic relatedness of the pair using the cosine between word co-occurrence vectors using software provided by Mohammad and Hirst (2006).

We conduct separate experiments with the two different lexicons for candidate set creation. We began by using CELEX, because it contains rich phonological information that some of our features draw on. However, in our analysis of the results, we noted that for many expressions the correct candidate pair is not in the candidate set. Many of the blends in WORDSPLEND are formed from words which are themselves new words, often coined for concepts related to the Internet, such as *download*, for example; such words are not listed in CELEX. This motivated us to create a lexicon from a

[2] Note that it would be possible to automatically infer the phonological and syllabic information required for our features using automatic approaches for text-to-phoneme conversion and syllabification (Bartlett, Kondrak, and Cherry 2008, for example). Although such techniques currently provide noisy information, phonological and syllabic information for the blend itself could also be inferred, allowing the development of features that exploit this information. We leave exploring such possibilities for future work.

recent data set (the Web 1T 5-gram Corpus) that would be expected to contain many of these new coinages. To form a lexicon from this corpus, we extract the 100K most frequent words, restricted to lowercase and all-alphabetic forms. Using this lexicon we expect the correct source word pair to be in the candidate set for more expressions. However, this comes at the expense of potentially larger candidate sets, due to the larger lexicon size. Furthermore, since this lexicon does not contain phonological or syllabic representations of each word, we cannot extract three features: the feature for the syllable heuristic, and the two features that capture the tendency for the second source word to be longer than the first in terms of phonemes and syllables. (We do calculate the phonological similarity between the two candidate source words, in terms of graphemes.)

4.3 Experimental Methods

Because each of our features is designed to have a high value for a correct source word pair and a low value otherwise, we can simply sum the features for each candidate pair to get a score for each pair indicating its degree of goodness as a source word pair for the blend under consideration. However, because our various features have values falling on differing ranges, we first normalize the feature values by subtracting the mean of that feature within that candidate set and dividing by the corresponding standard deviation. We also take the arctan of each resulting feature value to reduce the influence of outliers. We then sum the feature values for each candidate pair, and order the pairs within each candidate set according to this sum. This ranks the pairs in terms of decreasing degree of goodness as a source word pair. We refer to this method as the **feature ranking approach**.

We also use a machine learning approach applied to the features in a training regimen. Our task can be viewed as a classification problem in which each candidate pair is either a positive instance (the correct source word pair) or a negative instance (an incorrect source word pair). However, a standard machine learning algorithm does not directly apply because of the structure of the problem space. In classification, we typically look for a hyperplane that separates the positive and negative training examples. In the context of our problem, this corresponds to separating all the correct candidate pairs (for all blends in our data set) from all the incorrect candidate pairs. However, such an approach is undesirable as it ignores the structure of the candidate sets; it is only necessary to separate the correct source word pair for a given blend from the corresponding incorrect candidate pairs (i.e., for the same blend). This is also in line with the formulation of our features, which are designed to give relatively higher values to correct candidate pairs than incorrect candidate pairs within the candidate set for a given blend; it is not necessarily the case that the feature values for the correct candidate pair for a given blend will be higher than those for an incorrect candidate pair for another blend. In other words, the features are designed to give values that are relative to the candidates for a particular blend.

To address this issue, we use a version of the perceptron algorithm similar to that proposed by Shen and Joshi (2005). In this approach, the classifier is trained by only adjusting the perceptron weight vector when the correct candidate pair is not scored higher than the incorrect pairs *for the target blend* (not across all the candidate pairs for all blends). Furthermore, to accommodate for the large variation in candidate set size we use an uneven margin—in this case the distance between the weighted sum of the feature vector for a correct and incorrect candidate pair—of

$\frac{1}{\text{\#correct cand. pairs} \times \text{\#incorrect cand. pairs}}$. We therefore learn a single weight vector such that, within each candidate set, the correct candidate pairs are scored higher than the incorrect candidate pairs by a factor of this margin. When updating the weight vector, we multiply the update that we add to the weight vector by a factor of this margin to prevent the classifier from being overly influenced by large candidate sets. During testing, each candidate pair is ranked according to the weighted sum of its feature vector. To evaluate this approach, on each of WORDSPLEND and MAC-CONF we perform 10-fold cross-validation with 10 random restarts. In these experiments, we use our syllable heuristic as a feature, rather than as a filter, to allow the learner to weight it appropriately.

4.4 Evaluation Metrics

We evaluate our methods according to two measures: accuracy and mean reciprocal rank (MRR). Under the accuracy measure, the system is scored as correct if it ranks one of the correct source word pairs for a given blend first, and as incorrect otherwise. The MRR gives the mean of the rank of the highest ranked correct source word pair for each blend. Although accuracy is more stringent than MRR, we are interested in MRR to see where the system ranks the correct source word pair in the case that it is not ranked first. We compare the accuracy of our system against a chance (random) baseline, and an informed baseline in which the feature ranking approach is applied using just two of our features, the frequency of each candidate source word.

5. Experimental Results

5.1 Candidate Sets

We begin by examining some properties of the candidate sets created using CELEX as the lexicon, also referred to as the CELEX candidate set, in rows 2–4 of Table 4. First, in the second row of this table, we observe that only 78–83% of expressions have both source words in CELEX. For the other 17–22% of expressions, our system is always incorrect, because the CELEX candidate set cannot contain the correct source words.

Table 4
Percent of expressions (% exps) with their source words in each lexical resource and candidate set (CS), and after applying the syllable heuristic filter on the CELEX CS, as well as median CS size, for both the WORDSPLEND and MAC-CONF data sets.

Lexical resource or CS	WORDSPLEND		MAC-CONF	
	% exps	Med. CS size	% exps	Med. CS size
CELEX	78	-	83	-
CELEX CS	76	117	83	121
CELEX CS after syllable filter	71	71	77	92
Web 1T lexicon	92	-	-	-
Web 1T CS	89	442	-	-

The percentages reported in this row thus serve as an upper bound on the task for each data set.

The third row of Table 4 shows the percentage of expressions for which the CELEX candidate set contains the correct source words. Note that in most cases, if the source words are in CELEX, they are also in the CELEX candidate set. The only expressions in WORDSPLEND for which that is not the case are those in which a source word contributes a single letter to the blend. We could remove our restriction that each source word contribute at least two letters; however, this would cause the candidate sets to be much larger and likely reduce accuracy.

We now look at the effect of filtering the CELEX candidate sets to include only those candidate pairs that are valid according to our syllable heuristic. This process results in a 24–39% reduction in median candidate set size, but only excludes the source words from the candidate set for a relatively small number of expressions (5–6%), as shown in the fourth row of Table 4. We will further examine the effectiveness of this heuristic in the following subsection.

Now we examine the candidate sets created using the lexicon derived from the Web 1T 5-gram Corpus.[3] In the final two rows of Table 4 we see that, as expected, many more expressions have their source words in the Web 1T lexicon than in CELEX, and furthermore, more expressions have their source words in the candidate sets created using the Web 1T lexicon than in the candidate sets formed from CELEX. This means that the upper bound for our task is much higher when using the Web 1T lexicon than when using CELEX. However, this comes at the cost of creating much larger candidate sets; we examine this trade-off more thoroughly herein.

5.2 Source Word Identification

In the following subsections we present results using the feature ranking approach (Section 5.2.1), and analyze some of the errors the system makes in these experiments (Section 5.2.2). We then consider results using the modified perceptron algorithm (Section 5.2.3), and finally we compare our results to our previous study and human performance (Section 5.2.4).

5.2.1 Feature Ranking. Table 5 gives the accuracy using the feature ranking approach for both the random and informed baselines, each feature group, and the combination of all features, on each data set, using both the CELEX and Web 1T lexicons in the case of WORDSPLEND. Feature groups and combinations marked with an asterisk are significantly better than the informed baseline at the 0.05 confidence level using McNemar's Test.[4]

We first note that the informed baseline is an improvement over the random baseline in all cases, which points to the importance of word frequency in blend formation. We also see that the informed baseline is quite a bit higher on WORDSPLEND than MAC-CONF. Inspection of candidate sets—created from the CELEX lexicon—that include the correct source words reveals that the average source word frequency for WORDSPLEND

[3] Syllable structure information is not available for all words in the Web 1T lexicon; therefore, we do not apply the syllable heuristic filter to the pairs in these candidate sets (see Section 4.2). We do not create candidate sets for MAC-CONF using the Web 1T lexicon since this lexicon was constructed specifically in response to the kinds of new words found in WORDSPLEND.

[4] McNemar's Test is a non-parametric test that can be applied to correlated, nominal data.

Table 5
Percent accuracy on blends in WORDSPLEND and MAC-CONF using the feature ranking approach. The size of each data set is given in parentheses. The lexicon employed (CELEX or WEB 1T) is indicated. The best accuracy obtained using this approach for each data set and lexicon is shown in **boldface**. * = results that are significantly better than the informed baseline.

Features	WORDSPLEND (324) CELEX	 WEB 1T	MAC-CONF (30) CELEX
Random Baseline	6	3	1
Informed Baseline	27	27	7
Frequency	32*	32*	30*
Len./Cont./Phono.	20	20	7
Semantic	15	13	20
All	38*	**42***	**37***
All+Syllable	**40***	-	**37***

is much higher than for MAC-CONF (118M vs. 34M). On the other hand, the average for *non*-source words in the candidate sets is similar across these data sets (11M vs. 9M). Thus, although source words are more frequent than non-source words for both data sets, frequency is a much more reliable indicator of being a source word for truly novel blends than for established blends. This finding emphasizes the need for a data set such as WORDSPLEND to evaluate methods for processing neologisms.

All of the individual feature groups outperform the random baseline. We also see that our frequency features are better than the informed baseline. Although source word frequency (the informed baseline) clearly plays an important role in forming interpretable blends, this finding confirms that additional aspects of source word frequency beyond their unigram counts also play an important role in blend formation. Also note that the semantic features are substantially better than the informed baseline—although not significantly so—on MAC-CONF, but not on WORDSPLEND. This result demonstrates the importance of testing on true neologisms to have an accurate assessment of a method. It also supports our future plan to explore alternative semantic features, such as those that draw on the context of usage of a blend (as provided in our new data set).

We expect using all the features to provide an improvement in performance over any individual feature group, because they tap into very different types of information about blends. Indeed, the combination of all features (All) does perform better than the frequency features, supporting our hypothesis that the information provided by the different feature groups is complementary.[5]

Looking at the results on WORDSPLEND using the Web 1T lexicon, we see that as expected, due to the larger candidate sets, the random baseline is lower than when using the CELEX lexicon. However, the informed baseline, and each feature group used on its own, give very similar results, with only a small difference observed for the semantic features. The combination of all features gives slightly higher performance

[5] This difference is significant ($p < 0.01$) according to McNemar's test for the WORDSPLEND data set using both the CELEX and Web 1T lexicons. The difference was not significant for MAC-CONF.

using the Web 1T lexicon than the CELEX lexicon, although again this difference is rather small.

Recall that we wanted to see if the use of our syllable heuristic filter to reduce candidate set size would have a negative impact on performance. Table 5 shows that the accuracy on all features when we apply our syllable heuristic filter (All+Syllable) is at least as good as when we do not apply the filter (All). This is the case even though the syllable heuristic filter removes the correct source word pairs for 5–6% of the blends (see Table 4). It seems that the words this heuristic excludes from consideration are not those that the features rank highly, indicating that it is a reasonable method for pruning candidate sets. Moreover, reducing candidate set size will enable future work to explore features that are more expensive to extract than those currently used. Given the promising results using the Web1T lexicon, we also intend to examine ways to automatically estimate the syllable filtering heuristic for words for which we do not have syllable structure information.

5.2.2 Error Analysis. We now examine some cases where the system ranks an incorrect candidate pair first, to try to determine why the system makes the errors it does. We focus on the expressions in WORDSPLEND using the CELEX lexicon, as we are able to extract all of our features for this experimental setup. First, we observe that when considering feature groups individually, the frequency features perform best; however, in many cases, they also contribute to errors. This seems to be primarily due to (incorrect) candidate pairs that occur very frequently together. For example, in the case of *mathlete* (*math athlete*), the candidate pair *male* and *athlete* co-occurs much more frequently than the correct source word pair, causing the system to rank the incorrect source word pair first. We observe a similar situation for *cutensil* (*cute utensil*), where the candidate pair *cup* and *utensil* often co-occur. In both these cases, phonological information for the blend itself could help as, for example, *cute* ([kjut]) contributes more phonemes to *cutensil* ([kjutɛnsl̩]) than *cup* ([kʌp]).

Turning to the length, contribution, and phonology features, we see that although many blends exhibit the properties on which these features are based, there are also many blends which do not. For example, our first feature in this group captures the property that the second source word tends to be longer than the first; however, this is not the case for some blends, such as *testilie* (*testify* and *lie*). Furthermore, even for blends for which the second source word is longer than the first, there may exist a candidate pair that has a higher value for this feature than the correct source word pair. In the case of *banalysis*—*banal analysis*—*banal electrolysis* is a better source word pair according to this feature. These observations, and similar issues with other length, contribution, and phonology features, likely contribute to the poor performance of this feature group. Moreover, such findings motivate approaches such as our modified perceptron algorithm—discussed in the following subsection—that learn a weighting for the features.

Finally, for the semantic features, we find cases where a blend's source words are similar and related, but there is another (incorrect) candidate pair which is more similar and related according to these features. For example, *puggle*, a blend of *pug* and *beagle*, has the candidate source words *push* and *struggle* which are more semantically similar and related than the correct source word pair. In this case, the part-of-speech of the candidate source words, along with contextual knowledge indicating the part-of-speech of the blend, may be useful; blending *pug* and *beagle* would result in a noun, while a blend of *push* and *struggle* would likely be a verb. Another example is *camikini*, a blend of *camisole* and *bikini*. Both of these source words are women's garments, so we would expect them to have a moderately high similarity. However, the semantic similarity feature assigns this candidate pair the lowest possible score, since these words do not occur in the corpus from which this feature is estimated.

5.2.3 Modified Perceptron. Table 6 shows the average accuracy of the modified perceptron algorithm for the informed baseline and the combination of all features plus the feature corresponding to the syllable heuristic, on each data set, using both the CELEX and Web 1T lexicons in the case of WORDSPLEND. We don't compare this method directly against the results using the feature ranking approach because our perceptron experiments are conducted using cross-validation, rather than a held-out test set methodology. Examining the results using the combination of All+Syllable, we see that for each data set and lexicon the mean accuracy over the 10-fold cross-validation is significantly higher than that obtained using the informed baseline, according to an unpaired t-test ($p < 0.0001$ in each case).

Interestingly, on WORDSPLEND using the combination of all features, we see higher performance using the CELEX lexicon than the Web 1T lexicon. We hypothesize that this is due to the training data in the latter case containing many more negative examples (incorrect candidate pairs—due to the larger candidate sets). It is worth noting that, despite the differing experimental methodologies, the results are in fact not very different from those obtained in the feature ranking approach. One limitation of this perceptron algorithm is that it assumes that the training data is linearly separable. In future work, we will try other machine learning techniques that do not make this assumption.

5.2.4 Discussion. We now compare the feature ranking results on MAC-CONF here of 37% accuracy, to our previous best results on this data set of 27% accuracy, also using feature ranking (Cook and Stevenson 2007). To make this comparison, we should consider the differing baselines and upper bounds across the experiments. The informed baseline in our previous study on MAC-CONF is 13%, substantially higher than the 7% in the current study. Recall that the first row of Table 4 shows the upper bound using the CELEX lexicon on this data set to be 83%. By contrast, in our previous work we only use blends whose source words appear in the lexicon we used there (Macquarie), so the upper bound for that study is 100%. Taking these factors into account, the best results in our previous study correspond to a reduction in error rate (RER) over the informed baseline of 16%, while the feature ranking method here using

Table 6
Percent accuracy on blends in WORDSPLEND and MAC-CONF using the modified perceptron algorithm. The size of each data set is given in parentheses. The lexicon employed (CELEX or WEB 1T) is indicated. * = results that are significantly better than the informed baseline.

Features	WORDSPLEND		MAC-CONF
	(324)		(30)
	CELEX	WEB 1T	CELEX
Informed Baseline	23	24	7
All+Syllable	40*	37*	35*

the combination of all features and the syllable heuristic filter achieves a much higher RER of 39%.[6]

Lehrer (2003) finds human performance for determining the source words of blends to be 34% to 79%—depending on the blends considered—which indicates the difficulty of this task.[7] Our best accuracy on each data set of 37%–42% is quite respectable in comparison. These accuracies correspond to mean reciprocal ranks of 0.47–0.51, and the random baseline on WORDSPLEND and MAC-CONF in terms of this measure is 0.03–0.07. This indicates that even when our system is incorrect, the correct source word pair is still ranked fairly high. Such information about the best interpretations of a blend could be useful in semi-automated methods, such as computer-aided translation, where a human may not be familiar with a novel blend in the source text.

6. Blend Identification

The statistical features we have developed may also be informative about whether or not a word is in fact a blend—that is, we expect that if a novel word has "good" candidate source words, then the word is more likely to be a blend than the result of another word formation process. Because our features are designed to be high for a blend's source words and low for other word pairs, we hypothesize that the highest scoring candidate pairs for blends will be higher than those of non-blends.

To test this hypothesis, we first create a data set of non-blends from our earlier annotation, which found 671 non-blends out of the 1,186 Wordspy expressions (see Section 3). From these words, we eliminate all those beginning with a capital letter (to exclude words formed from proper nouns) or containing a non-letter character (to exclude acronyms and initialisms). This results in 663 non-blends.

We create candidate sets for the non-blends using the CELEX lexicon. Using the CELEX lexicon allows us to extract—and consider the contribution of—all of our length, contribution, and phonology features, some of which are not available when using the Web 1T lexicon. The candidate sets resulting from using the CELEX lexicon were also much smaller than when using the Web 1T lexicon. We calculate the features for the non-blends as we did for the blends, and then order all expressions (both blends and non-blends) according to the sum of the features for their highest-scoring candidate source word pair. We use the same feature groups and combinations presented in Table 5. Rather than set an arbitrary cut-off to distinguish blends from non-blends, we instead give receiver operating characteristic (ROC) curves for some of these experiments. ROC curves plot true positive rate versus false positive rate as the cut-off is varied (see Figure 1). The top-left corner represents perfect classification, with points further towards the top-left from the diagonal (a random classifier) being "better." We see that the informed baseline is a substantial improvement over a random classifier, and the combination All+Syllable is a further improvement over the informed baseline. The individual feature groups (not shown in Figure 1) do not perform as well as All+Syllable.

[6] Reduction in error rate = $\frac{\text{accuracy}-\text{baseline}}{\text{upper bound}-\text{baseline}}$.

[7] Note that the high level of interannotator agreement achieved in our annotation task (Section 3) may seem surprising in the context of Lehrer's results. However, our task is much easier, because our annotators were given a definition of the blend, whereas Lehrer's subjects were not.

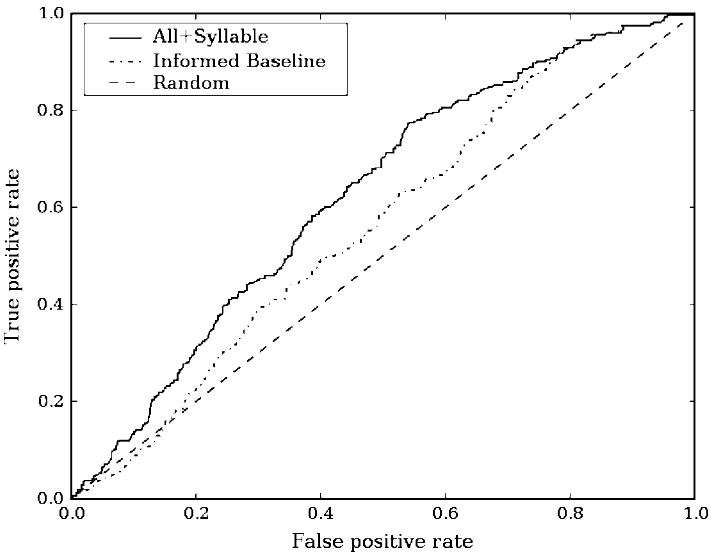

Figure 1
ROC curves for blend identification.

In future work, we plan to re-examine this task and develop methods specifically for identifying blends and other types of neologism.

7. Related Work

As discussed in Section 1, techniques generally used in the automatic acquisition of syntactic and semantic properties of words are not applicable here, because they use corpus statistics that cannot be accurately estimated for low frequency items, such as the novel lexical blends considered in this study (Hindle 1990; Lapata and Brew 2004; Joanis, Stevenson, and James 2008, for example). Other work has used the context in which an unknown word occurs, along with domain-specific knowledge, to infer aspects of its meaning and syntax (Granger 1977; Cardie 1993; Hastings and Lytinen 1994, for example). These studies have been able to learn properties of an unknown word from just one usage, or a small number of usages; however, the domain-specific resources that these studies rely on limit their applicability to general text.

Techniques for inferring lexical properties of neologisms can make use of information that is typically not available in other lexical acquisition tasks—specifically, knowledge of the processes through which neologisms are formed. Computational work on neologisms has tended to focus on tasks pertaining to a specific type of neologism, such as identifying and inferring the long form of acronyms (Schwartz and Hearst 2003; Nadeau and Turney 2005; Okazaki and Ananiadou 2006, for example), recognizing loanwords (Baker and Brew 2008; Alex 2008, for example), and identifying and expanding clippings (Means 1988, for example). This study focuses on the tasks of identifying, and inferring the source words of, lexical blends, a common type of neologism, which have been previously unaddressed except for our preliminary work in Cook and Stevenson (2007).

In addition to knowledge about a word's formation process, for many types of neologism, information about its phonological and orthographic content can be used to infer aspects of its syntactic and semantic properties. This is the case for neologisms that are composed of existing words or affixes (e.g., compounds and derivations) or partial orthographic or phonological material from existing words or affixes (e.g., acronyms, clippings, and blends). For example, in the case of part-of-speech tagging, information about the suffix of an unknown word can be used to determine its part-of-speech (Brill 1994; Ratnaparkhi 1996; Mikheev 1997, for example). For the task of inferring the long form of an acronym, the letters which compose a given acronym can be used to determine the most likely long form (Schwartz and Hearst 2003; Nadeau and Turney 2005; Okazaki and Ananiadou 2006, for example).

The latter approach to acronyms is somewhat similar to the way in which we use knowledge of the letters that make up a blend to form candidate sets and determine the most likely source words. However, in the case of acronyms, each word in a long form typically contributes only one letter to the acronym, while for blends, a source word usually contributes more than one letter. At first glance, it may appear that this makes the task of source word identification easier for blends, since there is more source word material available to work with. However, acronyms have two properties that help in their identification. First, there is less uncertainty in the "split" of an acronym, because each letter is usually contributed by a separate word. By contrast, due to the large variation in the amount of material contributed by the source words in blends, one of the challenges in blend identification is to determine which material in the blend belongs to each source word. Second, and more importantly, acronyms are typically introduced in regular patterns (e.g., the long form followed by the acronym capitalized and in parentheses) which can be exploited in acronym identification and long form inference; in the case of blends there is no counterpart for this information.

8. Conclusions

We propose a statistical model for inferring the source words of lexical blends—a very frequent class of new words—based largely on properties related to the recognizability of their source words. We also introduce a method based on syllable structure for reducing the number of words that are considered as possible source words. We evaluate our methods on two data sets, one consisting of novel blends, the other containing established blends; in both cases our features significantly outperform an informed baseline. Moreover, the results in this study are substantially better than those reported previously (Cook and Stevenson 2007). We further show that our methods for source word identification can also be used to distinguish blends from other word types. In addition, we annotate a data set of newly coined expressions which will support future research not only on lexical blends, but on neologisms in general.

Our future plans include expanding our techniques for identifying blends to address the more general problem of determining the formation process of a novel word. We further intend to apply our source word identification methods to other types of neologisms formed from material from existing words, such as clippings (e.g., *lab* for *laboratory*).

Acknowledgments
This article is an extended and updated version of a paper that appeared in the *Proceedings of the Tenth Conference of the Pacific Association for Computational Linguistics* (PACLING-2007). We thank the anonymous reviewers of this article for their comments, which have helped us to improve the quality of this work. We also thank the members of the computational linguistics group at the University of Toronto for their comments and feedback on our research. This work was financially supported by the National Sciences and Engineering Research

Council of Canada, the Ontario Graduate Scholarship program, and the University of Toronto.

References

Alex, Beatrice. 2008. Comparing corpus-based to Web-based lookup techniques for automatic English inclusion detection. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, pages 2693–2697, Marrakech.

Algeo, John. 1980. Where do all the new words come from? *American Speech*, 55(4):264–277.

Algeo, John, editor. 1991. *Fifty Years Among the New Words*. Cambridge University Press, Cambridge.

Ayto, John, editor. 1990. *The Longman Register of New Words*, volume 2. Longman, London.

Ayto, John. 2006. *Movers and Shakers: A Chronology of Words that Shaped our Age*. Oxford University Press, Oxford.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX Lexical Database (release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [distributor].

Baker, Kirk and Chris Brew. 2008. Statistical identification of English loanwords in Korean using automatically generated training data. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, pages 1159–1163, Marrakech.

Bartlett, Susan, Grzegorz Kondrak, and Colin Cherry. 2008. Automatic syllabification with structured SVMs for letter-to-phoneme conversion. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08): Human Language Technologies*, pages 568–576, Columbus, OH.

Bauer, Laurie. 1983. *English Word-formation*. Cambridge University Press, Cambridge.

Brants, Thorsten and Alex Franz. 2006. Web 1T 5-gram Corpus version 1.1. Linguistic Data Consortium, Philadelphia, PA.

Brill, Eric. 1994. Some advances in transformation-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 722–727, Seattle, WA.

Cardie, Claire. 1993. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 798–803, Washington, DC.

Cook, Paul and Suzanne Stevenson. 2007. Automagically inferring the source words of lexical blends. In *Proceedings of the Tenth Conference of the Pacific Association for Computational Linguistics (PACLING-2007)*, pages 289–297, Melbourne.

Delbridge, Arthur, editor. 1981. *The Macquarie Dictionary*. Macquarie Library, Sydney.

Granger, Richard H. 1977. FOUL-UP: A program that figures out the meanings of words from context. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pages 172–178, Cambridge, MA.

Gries, Stefan Th. 2004. Shouldn't it be breakfunch? A quantitative analysis of the structure of blends. *Linguistics*, 42(3):639–667.

Gries, Stefan Th. 2006. Cognitive determinants of subtractive word-formation processes: A corpus-based perspective. *Cognitive Linguistics*, 17(4):535–558.

Hastings, Peter M. and Steven L. Lytinen. 1994. The ups and downs of lexical acquisition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 754–759, Seattle, WA.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA.

Jiang, Jay J. and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING X)*, pages 19–33, Taiwan.

Joanis, Eric, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

Kelly, Michael H. 1998. To "brunch" or to "brench": Some aspects of blend structure. *Linguistics*, 36(3):579–590.

Knowles, Elizabeth and Julia Elliott, editors. 1997. *The Oxford Dictionary of New Words*. Oxford University Press, New York.

Kubozono, Haruo. 1990. Phonological constraints on blending in English as a case for phonology-morphology interface. *Yearbook of Morphology*, 3:1–20.

Lapata, Mirella and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Lauer, Mark. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University, Sydney.

Lehrer, Adrienne. 2003. Understanding trendy neologisms. *Italian Journal of Linguistics*, 15(2):369–382.

Means, Linda G. 1988. Cn yur cmputr raed ths? In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 93–100, Austin, TX.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Mohammad, Saif and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 35–43, Sydney.

Nadeau, David and Peter D. Turney. 2005. A supervised learning approach to acronym identification. In *Proceedings of the Eighteenth Canadian Conference on Artificial Intelligence (AI'2005)*, pages 319–329, Victoria.

Okazaki, Naoaki and Sophia Ananiadou. 2006. A term recognition approach to acronym recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL 2006)*, pages 643–650, Sydney.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::Similarity—Measuring the relatedness of concepts. In *Demonstration Papers at the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pages 38–41, Boston, MA.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge University Press, Cambridge.

Ratnaparkhi, Adwait. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, Philadelphia, PA.

Schwartz, Ariel S. and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical texts. In *Proceedings of the Pacific Symposium on Biocomputing (PSB 2003)*, pages 451–462, Lihue, HI.

Shen, Libin and Aravind K. Joshi. 2005. Ranking and reranking with perceptron. *Machine Learning*, 60(1):73–96.