Measuring Amok

Term Paper for CS 224U: Natural Language Understanding

Richard Futrell
Department of Linguistics
Stanford University
futrell@stanford.edu

Samuel Bowman
Department of Linguistics
Stanford University
sbowman@stanford.edu

Abstract

We propose and compare a number of metrics to capture the degree to which words are restricted in the contexts in which they can occur. We re-frame the problem of contextual restrictedness, and introduce the use of vector space models based on syntactic dependencies. We show that our most successful metric, residualized entropy, is quite successful in selecting highly collocationally restricted words, and is predictive of animacy.

1 Introduction: Collocational restrictedness

Words can freely co-occur with one another to express novel meanings, resulting in a combinatorial explosion for strings of more than one word, and sparse attestation for many strings even in large corpora. This productivity is a defining property of human language.

Productivity in natural language, however, is not absolute. Unlike formal languages, natural languages impose complex gradient constraints on the combination of terms. For instance, the adverb *amok* appears to have a highly restricted distribution: it can only modify the verb *run*. The phrase *run amok* is common and easily interpreted by humans, while the phrase *blossom amok* is unattested and unlikely to be understood by humans without some effort. Its distribution is nevertheless not categorical: in the NYT Gigaword corpus, *amok* is attested very rarely modifying the verb *go*, as in *goes amok*. In contrast, an adverb of comparable meaning, *insanely*, can and does appear modifying a much broader range of verbs. Productivity is limited in that some words have more restricted distributions than others.

We explore productivity in language by developing a measure of the distributional freedom or restrictedness of words. Previous work, under the headings of collocation detection and selectional preferences, has focused on characterizing the *relationship* between words. Building on this work, we develop and evaluate several summary measures to describe contextual restrictedness as a quantitative lexical property of individual words. Our metrics are based on vector space models with labeled grammatical dependencies as features.

Equipped with a general, reliable measure of contextual restrictedness, it should be possible to explore scientific questions about productivity and its correlates in languages. For instance, one can determine whether the grammars of different languages include more gradient or more categorical contextual restrictions, perhaps necessitating different processing strategies, or one can investigate if collocationally restricted words are more or less likely to change in meaning over time. In this project, we investigate one potential correlate of contextual restrictedness: animacy. We hypothesize that animate nouns, since they represent entities capable of a variety of actions, may have freer distributions, while inanimate nouns may have more restricted distributions.

1.1 Background: Previous attempts to capture the distributional properties of individual words

1.1.1 Contextual Distinctiveness

McDonald and Shillcock (2001) propose a contextual distinctiveness (CD) metric—a measure of how much a word's contextual distribution differs from that of a typical word. It appears to constitute the first attempt to extract a meaningful property of individual words from their co-occurrence distributions, so it is the most immediate precedent for our work.

Their metric defines the context of a word as a vector of counts of lemmas appearing in a k-word window around the target word, as in Lund and Burgess (1996). They then define the CD of a lemma w as the Kullback-Leibler divergence from the overall distribution of lemmas to the distribution of w, both calculated as maximum-likelihood estimators over their corpus.

$$\begin{aligned} CD(w) &= D(P(c)||P(c|w)) \\ &= \sum_{i=1}^{n} P(c_i|w) \log_2 \frac{P(c_i|w)}{P(c_i)} \end{aligned} \quad (1)$$

Strictly speaking, this measures contextual *distinctiveness*, or the unusualness of a word's contexts as compared to the average word, not contextual restrictedness, though the two may end up correlated empirically. It is possible, for example, that a word can co-occur freely with a wide range of infrequent and otherwise restricted words, but few frequent ones, giving it a high contextual distinctiveness and but relatively little contextual restrictedness.

McDonald and Shillcock (2001) motivate their metric with two studies. The first successfully shows that CD is much more closely correlated with subject response times in a lexical decision task (rapidly differentiating words from non-words) than pure word frequency. This result is replicated in Baayen (2010). The authors also compare CD with six other lexical properties defined without reference to a corpus—Concreteness, Context Availability, Number of Contexts, Ambiguity, Age of Acquisition and Familiarity—and find only an inverse relationship with ambiguity.

1.1.2 Selectional Strength

A similar measure has been applied to describe the selectional preferences of verbs. For instance, Resnik (1997) calculates a "selectional strength" $SelStr$ for each verb, a measure of how restricted its objects are. The equation is:

$$\begin{aligned} SelStr(v,r) &= D(P(c|v,r)||P(c|r)) \\ &= \sum_{c \in C} P(c|v,r) log \frac{P(c|v,r)}{P(c|r)} \end{aligned} \quad (2)$$

where c is a noun's WordNet class, v is the verb or predicate, and r is the relation between verb and noun (in this case the direct object relation). This measures the extent to which a knowing a verb and its relation to a noun changes the probability distribution of semantic classes for that noun. The verb's selectional preference for a particular object is just that object's contribution to the selectional strength.

Erk et al. (2010) show that Resnik (1997)'s $SelStr$ generalizes well to describe restictedness in other grammatical relations. They calculate "inverse selectional preferences", the extent to which knowing a noun and a relation change the probability distribution of verbs, measuring the distribution over lemmas W rather than WordNet classes C. The primary advantage of these approaches over those of McDonald and Shillcock (2001) is the use of grammatical relations in the vectors describing the contexts of a word, an idea originating from in Grefenstette (1993), who shows that vector space models incorporating relations are more able to select words that have the same syntactic and semantic categories.

1.1.3 The Frequency Problem

An ever-present confound for these information-theoretic measures of restrictedness is frequency. Low-frequency words are likely to appear spuriously distinct because of measurement error. McDonald and Shillcock (2001) deal with this issue by throwing out much of the data; they only consider the 500 most frequent words as context words, and do not consider target words with frequency less than 25. We aim for a measure that is more able to describe the contexts of low-frequency words.

2 Methods

2.1 Vector Space Model

We model the context of a target word as a vector of counts of context words with labeled grammatical relations. We only include context words that appear in some grammatical relation with the target word. We believe the use of labeled dependencies is crucial to the quality of our results. The word *beholder* may appear adjacent to any number of words, but in nearly all its appearances, it stands in a specific grammatical relation to the word *eye* as in *eye of the beholder*. Collapsed labeled dependencies capture that relation as *prep_of(eye, X)* and allow a more precise account of the restrictedness of *beholder*. Two sample vectors in this space are shown here:

	prep_of(eye, X)	nsubj(smell, X)	dobj(wear, X)
beholder	100	0	0
undershirt	0	50	25

Table 1: Two words represented with invented counts in a simplified version of our vector space.

In addition to using raw counts, we also apply two weighting schemes to the counts in our vectors: positive pointwise mutual information (Turney et al., 2010, PPMI) and the t statistic of Curran and Moens (2002).

2.2 Data

We test our model on the New York Times section of the Gigaword English Text Corpus (Graff et al., 2007), a collection of 914 million words of news text from 1994–2006. The corpus is not a perfectly balanced sample—it contains a substantial number of duplicate texts which we were not able to filter out.

We began with a version of the corpus that had been parsed by Nate Chambers at Stanford using the Stanford Parser (Klein and Manning, 2003), and extracted lemmas using the accompanying Stanford lemmatizer. We then converted the parses to collapsed typed dependency form (de Marneffe et al., 2006), annotated with part of speech tags and lemmas, yielding representations of the following form:

nsubj('re|be∧VBP-14, you|you∧PRP-13)

Lemmatizing fits our intuition that contextual restrictedness should hold equally of every inflectional variant of a word (but not every derivational variant: *totem=totems≠totemic*), and also helps to reduce the considerable problem of data sparsity for the relatively infrequent words that we are interested in. POS tagging enables us to at least partially alleviate the serious problem of homonyms, which show different interactions with context. Full-fledged word sense disambiguation may have better suited this task, but it was too unreliable and too computationally intensive to be practical for this project.

The lemmatization and dependency building operations were sufficiently time-intensive that we plan to make our version of the corpus available within the NLP group.

2.2.1 Filtering the Data

When building our final vector space model, we excluded all collapsed prepositional relations. We had experimented with including them, but found that this introduced considerable noise due to parse errors, and that freedom in prepositional relations tended to drown out restrictedness in other relations. For example, the word *wreak* should receive a high restrictedness rating because its direct object is almost always *havoc*, but one can also talk about *wreaking havoc in* any place, or *wreaking havoc with* any instrument. Each of these prepositional phrases would be coded as an independent context dimension, and we found that this resulted in surprisingly low restrictedness scores.

Furthermore, several nouns with truly restricted distributions with respect to prepositions do not have restricted distributions with respect to collapsed prepositional relations. The word *lot* appears in the binomial quantifier phrase *a lot of* followed by any noun. Using relations including *prep_of*, we find that *lot* is one of the freest words in the language. While this is an interesting observation about the construction *a lot of X*, it does not represent the word-based restrictedness which we are attempting to measure here.

2.3 Proposed Metrics

We apply the KL distinctiveness measures of (McDonald and Shillcock, 2001; Resnik, 1997; Erk et al., 2010) to our data as well as a simple measure of entropy over the context counts. Entropy is a more direct measure of freedom and restrictedness, rather than distinctiveness, as it simply quantifies the un-

certainty about a word's context. In order to get a value which increases with increased restrictedness, we add the (negative) raw entropy to 20. Because these information-theoretic measures are all highly correlated with frequency, we also calculate the residuals of these measures after controlling for log frequency in a linear regression.

In addition to the information-theoretic measures, we also test cosine distinctiveness, which is the cosine similarity of a lemma from the centroid in distributional space (the sum of all contextual vectors), subtracted from one to normalize directionality. This is an analogue of KL distinctiveness which we suspected to be less influenced by frequency.

We finally adapt a measure of morphological productivity, Baayen's P (Baayen, 2001). Applied to measure the productivity of the English prefix *pre*, Baayen's P is the number of hapax legomena (words occuring exactly once) with the prefix *pre*, divided by the token frequency of all words with the prefix. This is also known as vocabulary growth rate, and will be low for restricted prefixes and high for unrestricted ones. To measure the growth rate of the contexts of a word, we count the number of contexts with frequency 1 and divided by the sum count of all contexts.

2.4 Evaluation

In order to evaluate our metrics, we compiled up a list of 18 clearly restricted words (shown in blue), such as *beholder*, that appear almost exclusively in fixed phrases. We manually checked that the distribution of each word in the NYT Gigaword is categorical or overwhelmingly restricted. We also examine words of similar semantics to the restricted words (shown in black); in this case the semantic match for *beholder* is *observer*. Also we collected words of similar frequency and roughly similar semantics to the restricted words (shown in red), in this case *overseer*. The ability of the metrics to discriminate between restricted words and their low-frequency pairings is crucial.

For this task we use data that is not lemmatized, so that we can determine if inflectionally related words (i.e. *wreak* vs. *wreaks*) receive similar scores. The *prep_of* relation was included in order to capture certain idioms. A list of test words is provided below. In order to numerically evaluate the results of this task, we calculate the number of restricted words found in the top 15 results from each metric, and the number of restricted words found in the top 5.

Since the number of test items is so small, and since we are not tuning any statistical parameters, we did not explicitly divide the data into development and test sets for this task. We do, however, evaluate each metric independently on nouns and adjectives, and on verbs and adverbs, providing some sense of how well the performance of each metric generalizes to different cases. Furthermore, after selecting the metrics that perform reasonably on this toy task, we then sanity-check those metrics informally by inspecting the words given the highest restrictedness scores in a large lexicon.

Finally, we use the most successful metrics to predict the animacy class of nouns, and determine if distributional restrictedness is informative for this classification task.

3 Results

3.1 Distinguishing Restricted Words from Infrequent Words

Figure 1 shows our set of nouns and adjectives sorted by the scores assigned to each word by some of our metrics. Figure 2 shows scores assigned to verbs and adverbs.

Overall, our new measures, cosine distinctiveness and growth rate, do not stand out as better than other measures. We do not consider them in further analyses. The raw KL measure, which McDonald and Shillcock (2001) proposed for contextual distinctiveness, is also not among the best. The best results for a KL divergence-based metric come from weighting the counts according to the t-statistic, while the best overall scores come from the entropy of raw counts, achieving 5/5 accuracy in the top 5 words and 11/12 recall in the top 15.

The number of restricted words in the top 15 and top 5 ranked words for for metrics are given in table 2 and table 3.

The overall best measure appears to be the entropy of contexts. Raw KL distinctiveness does not perform especially well at distinguishing restricted words from infrequent ones, but its performance is competitive when counts are reweighted by PPMI or by the t statistic.

Top 15	Raw counts	PPMI	t-Test
KL Distinctiveness	7	10	10
Entropy	11	10	10
Cosine Dist.	8	10	6
Growth Rate (P)	7		
Top 5	**Raw counts**	**PPMI**	**t-Test**
KL Distinctiveness	3	3	4
Entropy	5	3	4
Cosine Dist.	3	3	4
Growth Rate (P)	3		

Table 2: The number of idiomatically restricted nouns and adjectives in the top 15 and top 5 most restricted words according to four metrics.

	Raw counts	PPMI	t-Test
KL Distinctiveness	3	4	4
Entropy	4	4	4
Cosine Dist.	3	4	4
Growth Rate (P)	0		

Table 3: The number of idiomatically restricted verbs and adverbs in the top 5 most restricted words according to four metrics. The scores for the top 15 are not shown, because all metrics place the 6 restricted words in this set into the top 15.

Baayen's P succeeds in not confusing infrequent nouns with restricted ones, but it confuses restricted words with the control words such as *damage* and *close*. Its performance is good but not better than entropy or t-test weighted KL distinctiveness. It also performs very poorly for ranking verbs and adverbs.

3.2 Finding Restricted Words in the Wild

Figure 3 shows the ten most restricted words in a large lexicon according to our two most successful metrics.

These metrics both find certain obviously restricted words, such as *cardiac* arrest, alma *mater*, and *vice* president, as well as rediscovering some of our original test words, such as *foregone*, *amok*, and *wreak*.

Upon examination, some of the more suspect words in the entropy list do turn out to be highly restricted in the corpus, for instance *loved* as an adjective appears highly frequently in the context *loved one*, and *wide* as an adverb appears mostly in the context *wide open*. The word *unearned* appears categorically in the phrase *unearned run*, a baseball term. *Olive*, misparsed as an adjective, appears overwhelmingly in the phrase *olive oil*.

PPMI KL:	**Pos. t-Test KL:**	**Raw Entropy:**
haywire	***roughshod***	***amok***
amok	***haywire***	***haywire***
roughshod	***wreaks***	***wreaking***
wreaks	***masquerade***	***wreaks***
crazily	***amok***	***harshly***
masquerade	***wreaking***	***wreak***
legalizing	***legalizing***	***legalizing***
wreaking	***pollute***	***roughshod***
harshly	***wreak***	***masquerade***
pollute	***gasp***	***crazily***
wreak	***crazily***	***pollute***
gasp	do	do
do	***harshly***	***gasp***
tosses	***tosses***	***tosses***
scream	***scream***	does

Figure 2: The top 15 restricted verbs and adverbs according to selected metrics. Blue words are highly restricted; red words are unrestricted but low-frequency words.

Pos. t-Test KL:	**Raw Entropy:**
oath∧NN	arthroscopic∧JJ
wide∧RB	starring∧JJ
arthroscopic∧JJ	unearned∧JJ
importantly∧RB	loved∧JJ
foregone∧JJ	integral∧JJ
hiding∧NN	foregone∧JJ
insatiable∧JJ	wide∧RB
saturated∧JJ	mater∧NN
mater∧NN	saturated∧JJ
pairing∧NN	unanswered∧JJ
cardiac∧JJ	vice∧NN
downfall∧NN	olive∧JJ
unearned∧JJ	cardiac∧JJ
stunned∧JJ	rectangular∧JJ
knock∧NN	amok∧RB

Figure 3: The top 15 most restricted words in our lexicon according to two of our best metrics.

The difference in function between the entropy and KL measure is apparent in these results. KL is a measure of distinctiveness; entropy is a measure of restrictedness. Thus the word *oath* receives a high KL score because it appears with an unusual set of words, such as *swear* and *take*, although it is relatively free to appear with any of these unusual words. Some of the highly-ranking KL results remain mysterious, such as *importantly*, which seems to have a relatively unremarkable distribution.

Raw KL:	PPMI KL:	Pos. t-Test KL:	Raw Entropy:	Raw Cos D.:	Growth Rate (*P*):
untrimmed	***bated***	***foggiest***	***bated***	***untrimmed***	close
bated	***foggiest***	***beholder***	***foregone***	***foggiest***	***foregone***
foggiest	***caboodle***	***foregone***	***foggiest***	***bated***	***foggiest***
dockyard	***dockyard***	***untrimmed***	***beholder***	***dockyard***	damage
beholder	***untrimmed***	***beeline***	***caboodle***	***beholder***	***beholder***
idealization	***beholder***	***lucre***	***beeline***	***undershirt***	range
foregone	***idealization***	***idealization***	***dockyard***	***foregone***	***bated***
undershirt	***foregone***	***umbrage***	***untrimmed***	***totem***	***fruition***
caboodle	***beeline***	***fruition***	***umbrage***	***fruition***	observer
totem	***lucre***	***undershirt***	***fruition***	***caboodle***	***gamut***
predetermined	***umbrage***	***totem***	***totem***	***idealization***	obscure
ineffectual	***undershirt***	***sharpshooter***	***idealization***	***gamut***	***havoc***
sharpshooter	***fruition***	***gamut***	***gamut***	***bulging***	***beeline***
lucre	***totem***	***overseer***	***lucre***	***sharpshooter***	***umbrage***
bulging	***sharpshooter***	***predetermined***	***undershirt***	***predetermined***	***totem***

Figure 1: The top 15 restricted nouns and adjectives according to selected metrics. Blue words are highly restricted; red words are unrestricted but low-frequency words.

3.3 Correlations with Frequency

Despite our efforts to select a metric robust to the effects of frequency, we still find a very strong correlation between the information-theoretic metrics and frequency. The raw entropy score is correlated with log frequency at r=-0.88, and the t-test weighted KL distinctiveness score is correlated with log frequency at r=-0.91.

In light of these strong correlations, we calculated another metric of restrictedness by simply taking the residual entropy score after controlling for log frequency in a linear regression. The results of this metric as applied to our test words are displayed in figure 4, in which the residual entropy score makes a clear distinction between infrequent and restricted words. The top ten most restricted words in the whole lexicon, by this metric, are displayed in figure 5.

foregone 5.3298745	***totem 0.8455303***
beholder 4.2269781	close 0.4840371
bated 3.5058180	damage 0.4196358
foggiest 3.0930278	***havoc -0.2222552***
fruition 1.9325606	***untrimmed -0.3404372***
beeline 1.8184322	range -0.6529436
gamut 1.3946669	displeasure -0.7873178
umbrage 1.3131754	

Figure 4: Top 15 restricted nouns and adjectives from the test list, sorted by residual entropy.

vice∧NN	since∧RB
last∧JJ	executive∧JJ
universal∧JJ	square∧JJ
already∧RB	longer∧RB
olive∧JJ	here∧RB
end∧VB	preliminary∧JJ
north∧JJ	no∧RB
prime∧JJ	

Figure 5: The top 15 most restricted words in our lexicon according the residualized entropy score.

The words selected by the residualized measure are markedly different from those selected by the other measures, in that they include several surprising high frequency adverbs such as *already*, *since*, and *no*. These seem at first to be in error, since they can occur in all sorts of semantic contexts. But upon examination, in the NYT corpus, the adverb *already* appears almost exclusively modifying the verb *to be* rather than other verbs, and *since*, when parsed as an adverb, appears almost exclusively modifying auxiliary *have* rather than verbs in the simple past. As far as we know, these contextual restrictedness for these adverbs has not been remarked upon previously. The adjective *last* appears primarily before time words, such as *week* or *year*, justifying its high rank in this listing. *No*, when parsed as an adverb, is nearly always in the phrase *no longer*.

Time words receive generally higher scores in the

residualized measure than otherwise; for instance, *year*, which appears almost always in time adverbials or after numbers, receives a residualized entropy score of 2.6, which means its restrictedness score is 2.6 bits higher than what one would expect from its frequency alone. It is ranked as the 6524th most restricted word by entropy, but as the 44th most restricted word by residual entropy. Similar patterns arise for *month*, *week*, and the season *spring*.

4 Contextual Restrictedness and Animacy

Here, we test the hypothesis that animate nouns are likely to be less restricted than inanimate nouns in the range of syntactic contexts in which they occur.

4.1 Data

We use data from the animacy hierarchy annotated section (Zaenen et al., 2004) of the NXT Switchboard Corpus (Calhoun et al., 2010). This corpus annotates noun phrases (NPs) for their position on an animacy hierachy containing the tags HUMAN, ORG (organizations), ANIMAL, PLACE, TIME, CONCRETE (physical objects), NONCONC (abstract entities), MAC (automata), VEH (vehicles), and MIX. In reducing these annotations to word–animacy pairs, we consider the animacy tag of an NP to hold of its lexical head (an assumption which seems to be fairly robust), and (for lack of any principled binary division) we consider the tags HUMAN, ANIMAL and MAC to denote animate entities.

4.2 Results

In order to examine possible correlations between animacy and our metrics of contextual restrictedness, we fit a logistic regression model predicting animacy (animate=1, inanimate=0) given various metrics. A model with log frequency and residual entropy score as features gives a significant negative coefficient to the entropy score frequency, indicating that highly restricted words are less likely to be animate (p ¡ 0.001). The entropy score feature does not, however, make the model a better fit than a model incorporating frequency alone. A model with frequency alone has precision (P) = 0.794 and recall (R) = 0.605 in predicting the data it was trained on; whereas a model incorporating entropy score has P = 0.773 and R = 0.603, a degradation in performance.

Models incorporating KL distinctiveness performed better. A model incorporating log frequency and KL distinctiveness, residualized on log frequency, achieves P = 0.853 and R = 0.603. Furthermore, a model incorporating frequency, KL distinctiveness, *and* entropy score (and all interactions among those two and frequency) achieves P = 0.881 and R = 0.605. In this model, the KL divergence was further residualized on entropy in order to avoid multicolinearity, since the two predictors were correlated at r = 0.47.

In order to ascertain that these positive results were not the result of overfitting, we split the data into a training set (90%) and a test set (10%), and trained our logistic regression model on the training set alone. On the training set, we find P = 0.887 and R = 0.603. On the test set, we find P = 0.844 and R = 0.653 (as opposed to P = 0.800 and R = 0.645 using only frequency as a feature). Though the values do fluctuate, the KL distinctiveness and entropy score together have good predictive value for animacy in unseen data.

The curious aspect of these models is the direction of their effects. The coefficients of the fitted logistic regression with frequency, KL, and entropy score as predictors are displayed below. The feature *freq.l* is log frequency; *h.rs* is entropy score residualized on frequency; and *kl.rs.rs* is KL distinctiveness residualized on frequency and on entropy. Interactions are indicated with colons.

Coefficients:	Est. Std.	Error	z value	Pr($>\|z\|$)
(Intercept)	-0.04987	0.196	-0.254	0.7995
freq.l	-0.26615	0.032	-8.234	>2e-16 ***
h.rs	-2.39103	0.260	-9.182	>2e-16 ***
kl.rs.rs	15.61553	0.508	30.754	>2e-16 ***
freq.l:h.rs	0.32498	0.035	9.205	>2e-16 ***
freq.l:kl.rs.rs	-2.35476	0.085	-27.680	>2e-16 ***
h.res:kl.rs.rs	-1.79567	0.871	-2.062	0.0392 *
freq.l:h.rs:kl.rs.rs	0.18067	0.126	1.432	0.1522

Table 4: Regression results.

The largest effect size is for KL distinctiveness; the *positive* effect size indicates that words that are *more* distinctive according to the KL score are *more* likely to be animate. This is the opposite of what we predicted: that restricted words were less likely to be animate. The reason for this effect could be that KL divergence is simply functioning to counteract the other predictors, which are all negative,

indicating that restricted words are less likely to be animate. The gains from using KL as a predictor are all in precision, which means that KL is functioning to cancel out the incorrect predictions of frequency and entropy score. It seems that unrestricted words tend to be animate, and words that are highly distinctive in context also tend to be animate.

4.3 Conclusion and Future directions

We have developed a promising metric for contextual restrictedness—the entropy of the dependecy distribution controlled for frequency—and shown that it captures our observations about which words are restricted, that it is a viable means of seeking out new restricted words, and that, when combined with with modified KL distinctiveness, it is predictive of a key lexical semantic property. In so doing, we have also reintroduced and formalized the notion of vector-space models based on syntactic depenencies for the measurement of lexical properties, and produced a corpus optimized for this purpose.

The most obvious continuation of this research would be the investigation of more potential metrics for contextual restrictedness. One promising direction in this line of work would be to develop metrics sensitive to the often metaphorical or frame-based nature of contextual restrictedness. For instance, suppose we observe phrase *strong wind* and we also observe the phrase *weak wind*, without finding other instances of the word *wind*. Then suppose we observe *strong wall* and *tall wall*. We should be able to infer that *wind* is more restricted than *wall*, because *wind* appears only with adjectives of strength, while *wall* appears with adjectives that are more semantically diverse. A measure that is sensitive to these patterns would be more robust to frequency, in that it would give different scores to these two hypothetical words, although they would receive the same score according to the metrics we have developed. The distributional similarity between $weak$ and $strong$ would allow the model to generalize beyond simple word-by-word co-occurence restrictions to the more complex restrictions based on metaphor.

Two possible kinds of metrics leap to mind to capture the semantic dimension of contextual restrictedness. A method applying similarity-based smoothing would result in a more restricted profile for *wind* than *wall* above, as measured the metrics we developed. Another kind of metric could locate each context word in distributional space and find the neighborhood density of the contexts of a word, perhaps using average pairwise distance. By taking semantics into account, these kinds of measures would yield more meaningful results than the current study.

It may also be worthwhile to investigate other possible lexical semantic correlates with restrictedness. For instance, the imageability of words—a subjective property shown McDonald and Shillcock show to be orthogonal CD—might correlate with contextual restrictedness. Contextual restrictedness will also be useful for comparing languages, and for discovering lists of words requiring special attention for foreign language learners.

Language modeling offers a promising application domain for dependency-based measures of contextual restrictedness. Popel and Mareček (2010) introduce and evaluate a novel class of language model based on syntactic dependencies, and show it to be extremely promising for domains where it can be realistically implemented. Their model conditions the probability of a word on its parent (and optionally, grandparent), the direction it looks towards that parent, and on any words that intervene between them. They linearly smooth all of the models they test, and find that the dependency language model provides much lower test set perplexities than do conventional models, with the most elaborate dependency model achieving a remarkable average of 65% of the perplexity of a standard trigram model (for which lower is better) across seven languages.

Though this approach has not yet been evaluated in an applied setting in any published literature, it is quite promising. Should it come in to use, it would provide opportunities for extensions based on dependency information, including, perhaps, a dependency-based adaptation of Modified Kneser-Ney smoothing (Chen and Goodman, 1999), a language model smoothing technique that already attempts to capture some information about the contextual restrictedness of words.

Appendix: Test words

Restricted nouns and adjectives: *bated, foggiest, foregone, beeline, beholder, caboodle, fruition, gamut, havoc, lucre, totem, umbrage*

High-freq controls: *restrained, obscure, predetermined, line, observer, bundle, success, range, damage, money, displeasure*

Low-freq controls: *bulging, untrimmed, ineffectual, sharpshooter, overseer, dockyard, mayhem, undershirt*

Restricted verbs and adverbs: *wreak, wreaks, wreaking, amok, roughshod, haywire*

High-freq controls: *cause, causing, crazily, harshly, run, ran, walk, sing, scream, do, does, understand, gasp, toss*

Low-freq controls: *tosses, masquerade, pollute, legalizing*

Acknowledgments

We owe thanks to Aaron Kalb for some useful ideas, and to Chris Potts and Bill MacCartney for a highly stimulating class!

References

R.H. Baayen. 2001. *Word frequency distributions*, volume 1. Springer.

R.H. Baayen. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3):436–461.

S. Calhoun, J. Carletta, J.M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. 2010. The nxt-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419.

S.F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

J.R. Curran and M. Moens. 2002. Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition-Volume 9*, pages 59–66. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

K. Erk, S. Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4).

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. *English Gigaword Third Edition*. Linguistic Data Consortium, Philadelphia.

G. Grefenstette. 1993. Evaluation techniques for automatic semantic extraction: comparing syntactic and window based approaches. In *Proc. of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text, Columbus Ohio*.

Dan Klein and Christopher D. Manning. 2003. Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*.

K. Lund and C. Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28(2).

S. McDonald and R. Shillcock. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3).

Martin Popel and David Mareček. 2010. Perplexity of n-gram and dependency language models. In *Proceedings of the 13th international conference on Text, speech and dialogue*.

P. Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington, DC.

P.D. Turney, P. Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

A. Zaenen, J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M.C. O'Connor, and T. Wasow. 2004. Animacy encoding in english: why and how. In *Proc. of the Association for Computational Linguistics Workshop on Discourse Annotation*, pages 118–125.