From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change

Ted Briscoe
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk
http://www.cl.cam.ac.uk/users/ejb

1. Introduction

I use the term valency in an extended sense as a relatively theory-neutral term to refer to lexical information concerning a predicate's realization as a single or multiword expression (such as a phrasal verb), the number and type of arguments that a particular predicate requires, and the mapping from these syntactic arguments to a semantic representation of predicate-argument structure which also encodes the semantic selectional preferences on these arguments. Thus, I use the term valency (frame) to subsume (syntactic) subcategorization and realization, argument structure, selectional preferences on arguments, and linking and/or mapping rules which relate the syntactic and semantic levels of representation.

For example, the verb, *believe*, in its primary sense and one (rare) realization takes a NP subject, NP object and infinitival VP complement where the subject is interpreted as the `believer', and therefore will typically denote the kind of entity capable of belief, the object NP is interpreted as the subject of the infinitival VP, and the proposition denoted by their composition is interpreted as `the belief', as in *Most voters believe the election to have been seriously mishandled*. Most grammatical frameworks treat valency almost entirely as a lexical property of predicates, although the inventory of valency frames, which varies between both major syntactic categories and between languages, can be described somewhat independently of individual words.

Some theories, such as Construction Grammar (e.g. Goldberg, 1995) argue that the valency frame itself, or `construction' contributes aspects of the overall meaning; for example, the dative frame is said to denote a `change of possession': hence, *The barman slid Fred the beer* but not *?The barman slid the end of the table the beer*. Such examples raise complications for a theory of the association between valency frames and predicates because it appears that the meaning of *slide* is coerced to entailing `change of possession' by virtue of insertion in the dative frame. Therefore a predicate's participation in alternant valency frames can result in predictable modification of meaning.

Abstracting over specific lexically-governed particles and prepositions and specific predicate selectional preferences, but including some `derived' / `alternant' semi-productive, and therefore only semi-predictable, bounded dependency constructions, such as particle or dative movement, there are at least 163 valency frames associated with verbal predicates in (current) English (Briscoe, 2000). In this paper, I will review the work that my colleagues and I have done to learn (semi-)automatically this very large number of associations between individual verbal predicates and valency frames.

Access to a comprehensive and accurate valency lexicon is critical for the development of robust and accurate parsing technology capable of recovering predicate-argument relations (and thus logical forms) from free text or transcribed speech. Without this information it is possible to `chunk' input into phrases but not to distinguish arguments from adjuncts or resolve most phrasal attachment ambiguities. Furthermore, for statistical parsers it is not enough to know the associations of predicates to valency frames, it is also critical to know the relative frequency of such associations given a specific predicate. Such information is a core component of that required to `lexicalize' a probabilistic parser, and it is now well–established that lexicalization is essential for accurate disambiguation (e.g. Collins, 1997, Carroll *et al*, 1998). While state-of-the-art wide-coverage grammars of English, capable of recovering predicate-argument structure and expressed as a unification-based phrase structure grammar, have on the order of 1000 rules, it is clear that the number of associations between valency frames and predicates needed in a lexicon for such a grammar will be much higher.

2. The dictionary-based approach

In the mid-eighties automated corpus analysis was not good enough to derive useful descriptions of the grammatical contexts in which predicates occurred, so we resorted to enhancing the valency information painstakingly gathered by lexicographers and available to us in the machine-readable versions of advanced learners' dictionaries, such as the *Longman Dictionary of Contemporary English* (LDOCE) (Boguraev and Briscoe, 1987; Boguraev et al, 1987). The resulting lexicon and subsequent similar efforts (e.g. Comlex syntax distributed by the LDC, Grishman *et al*, 1994) have high precision (approx. 95%) but disappointing recall (appox. 76% for ANLT, 84% for Comlex) which means that despite the large amount of lexicographical and linguistic resources deployed 24-16% of the required associations between predicates and valency frames were omitted for an open-class vocabulary of about 35,000 words (Briscoe and Carroll, 1997). However, given that there are more than 60,000 associations between words and valency frames (ignoring word sense differences) in the ANLT dictionary, perhaps the result is not so surprising. Many of the omitted associations are quite unremarkable; for example combining the associations from ANLT and Comlex would still leave the sentence types with *seem* in 1) (based on cited examples) unanalyzed:

1a) It seemed to Kim insane that Sandy should divorce.
1b) That Sandy should divorce seemed insane to Kim.
1c) It seemed as though Sandy would divorce.
1d) Kim seemed to me (to be) quite clever / a prodigy.
1e) (For Kim) to leave seemed to be silly.
1f) The issue now seems resolved.

In addition, neither of these projects yielded (public-domain) lexicons which associated predicates (i.e. word *senses* as opposed to word forms) with valency frames, nor did they record the relative frequency of a frame given a specific word, nor did they fully specify the mapping to argument structure or the selectional preferences on the arguments.

Though some of these limitations might be overcome by further work, inherently, the dictionary-based approach is untenable, because such associations vary, both absolutely and also in relative strength, depending on subject domain and genre (Roland and Jurafsky, 1998), and they change over time at a rate consonant with lexical sense change rather than grammatical change. For example, *swing* has a core sense of spatial motion with optional volitional / causative components, explaining its occurrence with locative and/or source/goal locational arguments: *Gay swung her into the saddle* (after the LOB corpus). However, in the financial domain, the extended sense of (share) price movements on a one-dimensional range is more common: *UAL's shares swung violently between an all time low of $6 and $36 in yesterday's trading* (after the WSJ). Sense specialization, such as that of *swinger* to mean sexually promiscuous, leads to new valency associations: *Kim swung with much of NY's literary society in the course of a long weekend* (after the Guardian). Therefore, at least the relative frequency of a valency frame varies depending on the relative frequency of the sense of a word, and in many cases valency frames are different under sense extensions. For example, with *swing* the selectional preferences and prepositional items are different for the financial usages, while with the example of *slide* above, or the following attested example (due to Annie Zaenen) *She smiled herself an upgrade*, the entire valency frame is only available under the extended sense.

3. The corpus-based approach

By the mid-nineties it was possible to reliably extract accurate phrasal analyses as well as part-of-speech (PoS) tags from corpora, and thus to identify the realization of predicate forms in phrasal contexts (e.g. Abney, 1996). That is, it became possible to `chunk' PoS-tagged input into verb groups, bare (unpostmodified) noun phrases, prepositional phrases, and so forth. Around this time, we developed a system to acquire the associations between predicate forms and valency frames based on on our own intermediate parsing technology (Briscoe and Carroll, 1997).

The system consists of :

- **A tagger**, a first-order HMM PoS and punctuation tag disambiguator, is used to assign and rank tags for each word and punctuation token in sequences of sentences (Elworthy, 1994).

- **A lemmatizer** is used to replace word-tag pairs with lemma-tag pairs, where a lemma is the morphological base or dictionary headword form appropriate for the word, given the PoS assignment made by the tagger. We use an enhanced version of the GATE project stemmer (Cunningham *et al*, 1995).

- **A probabilistic parser**, trained on a treebank, returns ranked analyses (Briscoe and Carroll, 1993) using a grammar written in a feature-based unification grammar formalism which assigns 'intermediate' phrase structure analyses to the PoS tag 'lattices' returned by the tagger (Briscoe and Carroll,1995; Carroll and Briscoe, 1996).

- **A pattern extractor** which extracts local syntactic frames, including the syntactic categories and head lemmas of constituents, from sentence subanalyses which begin/end at the boundaries of (specified) predicates.

- **A pattern classifier** which assigns patterns to valency frames or rejects patterns as unclassifiable on the basis of the feature values of syntactic categories and the head lemmas in each pattern.

- **A lexical filter** which evaluates sets of frames gathered for a (single) predicate, constructing putative lexical entries and filtering the latter on the basis of their reliability and likelihood.

For example, the sentence in 2a) is PoS-tagged as in 2b) and the parser returns the highest ranked analysis, shown as a labelled bracketing in 2c). Assuming that we are acquiring an entry for *attribute*, the extractor yields the pattern in 2d), which is classifiable as a known NP-NP-PSing valency frame with additional lexical information such as the preposition and heads of the NP arguments and of the NP and VP arguments of the PP, further parameterizing the frame hypothesized. In mapping from the analysis in 2b) to a valency frame, the classifier is able to repackage information from the fairly simple X-bar PSG used by the intermediate parser. In effect, the parser factors information between VSUBCAT and PSUBCAT features and the classifier looks at these, and sometimes lexical information, to deduce, in this case, that the appropriate frame is one in which the prepositional complement is a non-finite progressive clause.

2a) He attributed his failure, he said, to no-one buying his book.

2b) he_PPHS1 attribute_VVD his_APP$ failure_NN1 ,_, he_PPHS1 say_VVD ,_, to_II no-one_PN buy_VVG his_APP$ book_NN2

2c) (Tp
(V2 (N2 he_PPHS1)
(V1 (V0 attribute_VVD))
(N2 (DT his_APP$)
(N1
(N0 (N0 failure_NN1)
(Ta (Pu ,_,)
(V2 (N2 he_PPHS1)
(V1 (V0 say_VVD))) (Pu ,_,)))))
(P2
(P1 (P0 to_II)
(N2 no-one_PN)

(V1 (V0 buy_VVG)
(N2 (DT his_APP$) (N1 (N0 book_NN2)))))))))

2d) ((((he:1 PPHS1)) (VSUBCAT NP_PP) ((attribute:6 VVD) ((failure:8 NN1))
((PSUBCAT SING) ((to:9 II)) ((no-one:1 0 PN)) ((buy:11 VVG)))))

We call the level of analysis exemplified in 2c) `intermediate' because the parser finds singly rooted trees rather than simply chunking the input. However, many attachment decisions are `canonical'; for example, the attachment of the comma-delimited text adjunct (Ta) *he said* as a nominal postmodifier is incorrect but convenient for this application because it hides the text adjunct from the pattern extractor which computes the local syntactic context of *attribute*. The parser returns the highest ranked analysis using a purely structural probabilistic model for ranking alternative analyses. This makes training the parser on realistic amounts of data and using it in a domain-independent fashion feasible. However, it also means that patterns extracted from the highest ranked analysis are noisy because quite often the parser has no mechanism for choosing the correct analysis. For example, the correct analysis for 3a) is shown in 3c) and the correct analysis for 3b) in 3d).

3a) He looked up the word.
3b) He looked up the hill.
3c) (Tp (V2 (N2 he_ PPHS1) (V1 (V0 (V0 look_VVD) (P0 up_RP)) (N2 (DT the_AT) (N1 (N0 word_NN1)))))
3d) (Tp (V2 (N2 he_ PPHS1) (V1 (V0 look_VVD) (P2 (P1 (P0 up_RP) (N2 (DT the_AT) (N1 (N0 hill_NN1)))))))

However, the parser cannot reliably select between the derivations in 3c) and 3d) because it does not have access to any lexical information such as the likelihood of *look up* being a phrasal verb or the differing selectional restrictions on the NP as either PP or verbal argument.

The classifier rejects some noisy patterns which do not conform to the known valency frames for English. However, many classifiable patterns are still incorrect for the reasons noted above, so the filter accrues evidence for associations between the specified predicate form and specific valency frames. These are then filtered using a statistical confidence test which utilizes the overall error probability that a particular frame will be hypothesized, and the amount of evidence for an association of that frame with the predicate form in question. Specifically, we think of occurrences of predicate forms with putative frames as a sequence of independent (Bernoulli) trials and use the error probability, P_e^i, that a predicate form will be associated with an incorrect valency frame, *i*, to formulate the null hypothesis. The probability of an event with probability *p* happening exactly *m* times out of *n* such trials is given by the binomial distribution:

$$P(m,n,p) = n! \div m!(n-m)!\, P^m (1-p)^{n-m}$$

The probability of an event happening *m* or more times is:

$$P(m+,n,p) = \Sigma_{k=m}^{n} P(k,n,p)$$

So $P(m+,n\ P_e^i,)$ is the probability that *m* or more occurrences of valency frame *i* will be associated with a predicate form occurring *n* times. The threshold for rejecting this null hypothesis was set to 0.05 yielding a 95% or better confidence that a high enough proportion of frames has been seen, given the underlying error probability, to accept the hypothesis that the predicate form really is associated with the frame.

The error probability for a given frame *i* was estimated by:

$$P_e^i = (1 - (|pred\ in\ frame\ i| \div |preds|)) (|frames\ for\ i| \div |frames|)$$

where the counts for frames were obtained by running the parser and extractor software on the entire Susanne corpus (Sampson, 1995) and the estimates of the numbers of predicates associated with frame *i* is obtained by counting the number of predicates in the ANLT lexicon paired with that frame. Suppose that

the parser and extractor predict that verb tokens are associated with frame *i* ¾ of the time, and that only ¼ of the verb types in ANLT are associated with *i*, then the error probability for associating verbs with *i* will be slightly over half. If the frame is only hypothesized ¼ of the time but linked to verb types ¾ of the time, the error probability will be one eighth. Note, however, that the estimate of the true incidence of the association, being dictionary-based, doesn't take account of the relative frequency of tokens of the verb types.

The performance of the system can be evaluated by recording true positives (TP) and true negatives (TN), that is, cases where the filter correctly accepts or rejects an association between a predicate form and a frame, and false positives (FP) and false negatives (FN), where the filter incorrectly accepts or rejects such an association. The incidence of the four outcomes can be calculated either by comparing the lexical entries produced by the system to a set of accurate lexical entries, or to a manually compiled set of entries based on the same data that the system used to acquire the entries. The advantage of the former approach is that it is quicker – we can create accurate entries by intelligently merging the information in the ANLT and Comlex lexicons. However, the disadvantages are that we have no way of knowing whether the data the system used actually exemplified all the frames in the resulting lexical entries and, since these lack information about the relative frequency of frames for specific predicates, no way of knowing whether the system has acquired accurate frequencies. Therefore, we supplement this method with manual analysis of the data given to the system. This allows us to measure type recall and precision against the gold standard entries and the corpus analysis and token recall and ranking accuracy against the corpus analysis, as defined below:

Type recall: TP ÷ (TP + FN types in dict/corpus)
Type precision: TP ÷ (TP + FP types in dict/corpus)
Token recall: TP ÷ (TP + FN tokens in corpus)
Ranking accuracy: % pairs of TPs whose ranking by rel. Freq. is same as in corpus

For 14 pseudo-randomly chosen verbs and training data of between 200 and 1000 exemplars per verb, the system achieved the results shown in the table below:

	Dictionary (14 vbs)	Corpus (7 vbs)
Type precision	65.7%	76.6%
Type recall	35.5%	43.4%
Token recall		80.9%
Ranking Accuracy		81.4%

These results are considerably worse than the (type) precision and recall results for ANLT and Comlex, reported above, but nevertheless were promising. Analysis of the system behaviour showed that for associations seen less than 10 times, the binomial hypothesis test performed no better than chance. Therefore, much of our subsequent work has focused on how to improve the filtering and entry construction component of the system.

4. Better filtering

Briscoe, Carroll and Korhonen (1997) demonstrated that iteratively optimizing the error probabilities used in the above experiment on the basis of the correct entries resulted in a system yielding nearly 9% improvement in type precision, over 20% improvement in type recall, and a 10% improvement in ranking accuracy on a further 20 test verbs. This result shows that the dictionary-based estimation of error probabilities is far from optimal. However, the other critical problem with the binomial hypothesis test -- that it deals poorly with low frequency events -- remains. Korhonen, Gorrell and McCarthy (2000) compared results using the binomial test, the log likelihood ratio test and an empirically determined thresholding scheme. A log likelihood test has been proposed by Dunning (1993) as appropriate for non-normally distributed data with many rare events. However, on this task the log likelihood test performed significantly worse than the binomial, resulting in more FNs for high frequency frames, more FPs for medium frequency frames and no improvement in performance on low frequency frames. The thresholding

scheme simply involved converting the hypothesized counts for each frame for a given predicate form into a conditional probability for each frame given the predicate, $P(frame_i \mid predicate_j)$ using the maximum likelihood estimate (MLE, i.e. the ratio of the count for $predicate_j + frame_i$ over the count for $predicate_j$) and then rejecting any association with a probability lower than an empirically determined optimum on non-test data. This resulted in a 24% improvement in precision and a 2% improvement in recall. However, thresholding still results in many FNs for low frequency associations.

The reason why hypothesis testing does not work well for this task is that not only is the underlying distribution Zipfian, but also there is very little correlation between the unconditional distribution on valency frames independent of specific predicates and the conditional distribution of frames given a specific predicate. For example, *believe* occurs mostly with a sentential complement, but the sentential complement frame, in general, is rare. Therefore, any test that involves reference to the unconditional distribution, filtering hypotheses about the conditional distributions, will perform badly. Unfortunately, the same observation also undermines any attempt to use the unconditional distribution as the prior in a Bayesian scheme or to smooth the conditional distributions to compensate for the known poor performance of MLE on rare / unseen events. Korhonen (2000), compares the performance of the MLE and thresholding scheme smoothed with the conditional distributions for a predicate related to those whose associations the system is attempting to learn.

One simple approach to smoothing one distribution with another to infer the probabilities of unseen events is linear interpolation (e.g. Manning and Schutze, 1999:218f). The conditional probability $P(frame_i \mid predicate_j)$ is smoothed using the conditional probability of $P(frame_i \mid pre\ dicate_k)$ where $pre\ dicate_k$ stands in some specified relationship to $pre\ dicate_j$ according to the formula below:

$$P(frame_i \mid predicate_j) = \lambda_1 \ (P(frame_i \mid predicate_j)) \ + \ \lambda_2 \ (P(frame_i \mid predicate_k))$$

where the λ_i denote weights which sum to 1, and can be optimized with held out data so that most of the smoothed probability for $P(frame_i \mid predicate_j)$ is determined by its MLE estimate. Korhonen (2000) demonstrates that the distribution of frames given predicates is better correlated for hypernyms and synonyms of target predicates (derived from WordNet) than for the unconditional distribution. She, then, compares results for acquiring frames for 60 test verbs from 10 semantically-defined classes (based on Levin's (1993) classification), using as a baseline MLE thresholding with no smoothing and comparing this to linear interpolation of the MLE estimates with the unconditional distiribution for all verb frames and also with the merged conditional distributions for 3 other verbs from the same class. Her results are shown in the table below:

	Baseline	Unrelated smoothing	Sem.-related Smoothing
Type Precision	78.5%	71.4%	87.8%
Type Recall	63.3%	64.1%	68.7%
Ranking Accuracy	79.2%	67.6%	84.4%

The measures shown are based on comparison with the corpus data used by the system. The baseline performance is better than smoothing with the poorly correlated unconditional distribution. However, smoothing against the conditional distributions of semantically related predicates results in a significant improvement in performance. All but 3 of the 151 low frequency FNs rejected by MLE and thresholding exceed the threshold after semantically-related smoothing, suggesting that this approach provides an effective way of dealing with low frequency associations.

There is, of course, a cost to the smoothing approach, as it is necessary to obtain the smoothed distributions for all the different classes of predicates exemplified in the English lexicon. However, based on analysis and extension of Levin's (1993) classification, it seems likely that the total number of classes needed for good performance on this task is unlikely to exceed 50, so it should be possible to seed such a system by obtaining about 200 conditional distributions semi-automatically for smoothing purposes. The underlying reason why this approach is effective is that similar predicates have similar, though by no means identical, `paradigms' of associations to valency frames, because they tend to undergo the same types of frame alternation processes as other predicates in the same class.

This more semantically-driven approach to learning valency associations, seems to be getting us closer to the tested performance of existing manually-derived valency dictionaries such as ANLT or Comlex, with the added advantage that we recover the relative frequencies of these associations. However, it also raises the issue of polysemy and the what precisely the associations are between. I argued initially that associations hold between predicate *senses* and specific frames. However, our corpus-based work so far has resulted in associations between target predicate *forms* and frames. In order to classify predicate forms into semantic classes, Korhonen used the predominant sense of the form as defined by WordNet . However, it is clear that the results obtained with the technique would be improved if it were possible to classify occurrences of predicate forms into appropriate senses. Similarly, the corpus-based work so far has resulted in the recovery of lists of lemmas which occur as the heads of arguments in frames with given predicates, but the induction of selectional preferences from these lists requires their sense disambiguation too.

5. Predicate and argument sense disambiguation

The performance of word sense disambiguation systems (e.g. Wilks and Stevenson, 1998) has improved to the point where it is viable to integrate such technology into our system. McCarthy (1997) and McCarthy and Carroll (2000) develop a system which utilizes the WordNet semantic hierarchy on nouns and the lists of head lemmas in argument slots in valency frames returned by the pattern extractor to infer a probability distribution on semantic classes occurring in a given argument position in a given frame for specific predicates. This probability distribution characterizes the selectional preference(s) of the predicate on that argument.

The approach utilizes the minimum description length principle (MDL e.g. Li and Vitanyi, 1998) and aims to find the smallest model over semantic classes which describes the data (list of head lemmas) most succinctly. The head lemmas are assigned to WordNet semantic classes and counts of classes exemplified for a particular frame argument are maintained and propagated through the hierarchy. If a lemma is ambiguous between classes then the counts are evenly distributed between these classes. For example, for the direct object argument in the transitive frame for *eat*, lemmas such as *food* and *chicken* might be seen. In WordNet *food* is classified as a **substance** and is itself a semantic class **food**, while *chicken* is variously classified as **food, bird** and **person** with only the **food** class itself classified as a **substance**. Therefore, the counts would be divided and propagated so that **food** and **substance** each obtained 1 and 1/3 counts, while **bird** and **person** obtained 1/3 each. These counts are turned into MLE estimates, but the low probability classes are often filtered out using MDL since they add to model complexity without compressing the description of the lemmas that occur.

The resulting selectional preference distributions have been used for WSD on nouns occurring in frame slots with competitive results of around 70% precision (Kilgariff and Rosenzweig, 2000), and also to identify whether a specific predicate participates in a frame alternation. For example, many verbs with a causative component to their meaning participate in the so-called causative-inchoative alternation, exemplified in 4).

4a) Kim broke the window.
4b) The window broke.
4c) Kim galloped the horse.
4d) The horse galloped.

Identifying the precise class semantically is difficult if not impossible (e.g. Levin and Rappaport, 1995), probably because the alternation is semi-productive and partly driven by item familiarity (e.g. Goldberg, 1995). An alternative is to look for near identical selectional preference distributions on argument slots between valency frames putatively related by such alternations. In this case, we would expect the direct object slots in 4a and c) to share similar distributions to the subject slots in 4b) and 4d), respectively. This approach to such alternations has several potential advantages from the perspective of learning valency frames for predicates, such as more economic representation of valency lexicons, statistical estimation of the semi-productivity of alternation rules and thus using them as a further aid in the induction of low frequency frame associations (e.g. Briscoe and Copestake, 1999), and perhaps most importantly inferring

when the basic sense of a predicate will be systematically and predictably modified by association with a frame, as with the example of *slide* and dative movement discussed in the introduction. McCarthy (2000) reports results of an experiment with two alternation rules (causative-inchoative and conative) which yielded a 75% accuracy rate at classifying predicates as participants in these alternations. Most errors were FPs and it may be once again that polysemy of the predicate forms accounts for much of this noise. Most recently McCarthy in unpublished work has begun to disambiguate verb forms into WordNet defined senses using the distribution of nouns in argument slots which the verb form has been associated with. If this work is successful, it should be possible to associate associations with predicate senses directly, and thus to reduce some of the noise inherent in the work that relies on sorting predicates into semantic classes before identifying the frames they are associated with, the selection preferences on these frames, and so forth.

6. Related work

There seems to have been very little work done on corpus-based acquisition of valency frame associations prior to the nineties. Seminal work was done by Brent (1991) establishing the importance of precise cues and introducing the binomial hypothesis test. Manning (1993) extended this work by performing finite-state text chunking before extracting patterns for frames. Ushioda *et al* (1993) were the first to attempt to extract the relative frequency of different frames. These systems recognized a maximum of 23 frames and had performance rates of around 80% token recall, in line with the original system we developed which recognized 161 frames. More recent papers are Gahl (1998), Carroll and Rooth (1998), Lapata (1999), Stevenson and Merlo (1999) and Zeman and Sarkar (2000). None of these utilizes a set of frames as large as ours or reports results suggesting more accurate performance. Carroll and Rooth dispense with hypothesis testing and use estimation maximization without thresholding, prefiguring the results of Korhonen, Gorrell and McCarthy (2000) to some extent. Zeman and Sarkar also compared the log likelihood ratio to the binomial hypothesis test and preferred the latter, but start from preparsed data. Lapata and Stevenson and Merlo focus on learning which verbs participate in a small number of alternations.

7. The future: self-organizing dictionaries

The recent experiments we have undertaken are based on applying the pattern extractor to 20M words of the BNC (Leech, 1992). There is no reason, in principle, for us not to parse more data, however, the overall performance of the system would be improved if parse selection accuracy were better. Carroll, Minnen and Briscoe (1998) demonstrate that using the valency frame associations acquired by our system to rerank the derivations returned by the parser improves its accuracy by about 10%. Incrementally integrating this and other lexical information into the probabilistic ranking of derivations is a current avenue of research. Similarly, there are other incremental improvements that can and will be made to the different subsystems being developed for identifying predicate senses, selectional preferences and alternation behaviour.

Nevertheless, no matter how much data is analysed however accurately, this data will still be inadequate from a statistical perspective for the acquisition of an accurate and comprehensive valency lexicon. In the limit, the arguments that I made against the dictionary-based approach also apply to the corpus-based approach when it is applied to a finite corpus. Zipf (1949) demonstrated that several distributions derived from natural language approximate to power laws in that the probability mass is distributed non-linearly between types with a few of the most frequent types taking the bulk of the probability mass and a very long tail of rare types. Both the unconditional distribution of valency frames and the conditional distributions of frames given specific predicates are approximately Zipfian (e.g. Korhonen, Gorrell and McCarthy, 1998). Two conclusions that can be drawn from this are: 1) that, because the power law is scaling invariant, any finite sample will not be representative in the statistical sense, and 2) that power law distributions are very often a clue that we are not sampling from a stationary source but rather from a dynamical system (e.g. Casti, 1994). Baayen (1991) develops a dynamical model of word generation which predicts a Zipfian distributed vocabulary. Briscoe (2001a) develops a more general model of language as a dynamical system, the aggregate output of a changing population of partly heterogeneous generative grammars. From this perspective it is not surprising that classical statistical models of learning, which rely on representative

samples from stationary sources, do not perform optimally. A better model of a valency lexicon, given these observations, is of an adaptive self-organizing knowledge base which continually monitors data to update associations and the strengths of associations between predicates and valency frames. The techniques outlined in this paper will need to be improved before we will be able to meaningfully develop such a model. However, such models can be developed as an extension of the statistical techniques used here by augmenting these techniques with the ability to incrementally update and renormalize the conditional distributions acquired from data (see Briscoe 2001b) for an implementation of Bayesian incremental updating of parameter values.

References

Abney, S. 1996. Part-of-speech tagging and partial parsing. In Church, K., Young, S. and Bloothoft, G. *Corpus-based Methods in Speech and Language*, Dordrecht, Kluwer.

Baayen, H. 1991. A stochastic process for word frequency distributions. *Proc of 29th Assoc. For Comp. Ling.* 271-278.

Boguraev, B.K. and Briscoe, E.J. 1987. Large lexicons for natural language processing: utilising the grammar coding system of the *Longman Dictionary of Contemporary English, Computational Linguistics* 13.4, 219-240.

Boguraev, B.K., Briscoe, E.J., Carroll, J., Carter, D. and Grover, C. 1987. The derivation of a grammatically-indexed lexicon from the *Longman Dictionary of Contemporary English*, Proc. of 25th Assoc. for Comp. Ling., 193-200, Morgan Kaufmann, Palo Alto, CA.

Brent, M. 1991. Automatic acquisition of subcategorization frames from untagged text. *Proc. of 29th Assoc. For Comp.Ling.*, 209-214, Morgan Kaufmann, Palo Alto, CA.

Briscoe, E.J. 2000. Dictionary and System Subcategorisation Code Mappings, Ms. Computer Laboratory.

Briscoe, E.J. 2001a. Evolutionary perspectives on diachronic syntax, *Diachronic Syntax: Models and Mechanisms.* (eds) Pintzuk, S., Tsoulas, G. and Warner, A.}Oxford University Press, Oxford.

Briscoe, E.J. 2001b. Grammatical Acquisition and Linguistic Selection. In *Linguistic evolution through language acquisition: formal and computational models.* (ed.) Briscoe, E.J. Cambridge University Press, Cambridge.

Briscoe, E.J. and Carroll, J.A. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19.1, 25-60.

Briscoe, E.J. and Carroll, J.A. 1995. Developing and Evaluating a Probabilistic LR Parser of Part-of-Speech and Punctuation Labels, *4th Int. Workshop on Parsing Technologies (IWPT95)* Morgan Kaufmann, Palo Alto, CA.

Briscoe, E.J. and Carroll, J.A. 1997. Automatic extraction of subcategorisation from corpora. *Proc. of 5th Assoc. For Comp. Ling. Conf. on Applied Nat. Lg. Proc.*, Morgan-Kaufmann. Palo Alto, CA.

Briscoe, E.J. and Copestake, A.A. 1999. Lexical Rules in Constraint-based Grammars, *Computational Linguistics*, 25.4, 487-526.

Carroll, J.A. and Briscoe, E.J. 1996. Apportioning development effort in a probabilistic LR parsing system through evaluation. *2nd Conference on Empirical Methods in Natural Language Processing*, 92-100, Morgan-Kaufmann. Palo Alto, CA.

Carroll, J.A. and McCarthy, D. 2000. Word sense disambiguation using automatically acquired verbal preferences. In *Computers and the Humanities. Senseval Special Issue*, 34, 1-2.

Carroll, J.A., Minnen, G., and Briscoe, E.J. 1998. Can subcategorisation probabilities help a statistical parser? *6th Workshop on Very Large Corpora*, 35-41, Morgan-Kaufmann. Palo Alto, CA.

Carroll, G. And Rooth, M. 1998. Valence induction with a head-lexicalized PCFG. *3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

Casti, J.L. 1994. *Complexification*. Harper Collins, New York.

Collins, M. 1997. Three generative lexicalised models for statistical parsing, *Proc of 35th Assoc. for Comp. Ling.* 16-23, Morgan-Kaufmann. Palo Alto, CA.

Cunningham, H., Gaizauskas, R. \& Wilks, Y. 1995. A general architecture for text engineering (GATE) - a new approach to language R&D. Research memo CS-95-21, Department of Computer Science, University of
Sheffield, UK.

Elworthy, D. 1994. Does Baum-Welch re-estimation help taggers? *Proc. of 4th Conf. Applied Nat. Lang. Processing*, Morgan-Kaufmann. Palo Alto, CA.

Gahl, S. 1998. Automatic extraction of subcategorization frames from a part-of-speech tagged corpus. *Proc. of Assoc. For Comp. Ling.* Morgan-Kaufmann. Palo Alto, CA.

Goldberg, A. 1995. *A construction grammar approach to argument structure*, Chicago UP.

Grishman, R., Macleod, C. and Meyers, A. 1994. Comlex syntax: building a computational lexicon. *Int. Conf. on Computational Linguistics*, 268-272, Kyoto, Japan

Kilgariff, A. and Rosenzweig, J. 2000. English Senseval: report and results. Ms. ITRI, Brighton, UK (www.itri.bton.ac.uk)

Korhonen, A. 2000. Using semantically motivated estimates to help subcategorization acquisition. *Proc. of Jnt. Conf. on Empirical Methods in NLP and Very large Corpora*, Morgan-Kaufmann, Palo Alto, CA.

Korhonen, A., Gorrell, G. and McCarthy, D. 2000. Statistical Filtering and Subcategorization Frame Acquisition. *Proc. of Jnt. Conf. on Empirical Methods in NLP and Very large Corpora*, Morgan-Kaufmann, Palo Alto, CA.

Lapata, M. 1999. Acquiring lexical generalizations from corpora: a case study for diathesis alternations. *Proc. of 37th Assoc. For Comp. Ling.* 397-404, Morgan-Kaufmann, Palo Alto, CA.

Leech, G. 1992. 100 million words of English: the British National Corpus. *Language Research* 28.1, 1-13.

Levin, B. 1993. *English verb classes and alternations*. Chicago Univ. Press, Chicago.

Levin, B and Rappaport, H. 1995. *Unaccusativity*, MIT Press, Cambridge, MA.

Li, M. and Vitanyi, P.1998. *An Introduction to Kolmogoroff Complexity and Its Applications*, Springer-Verlag, Heidelberg.

Manning, C. 1993. Automatic acquisition of a large subcategorisation dictionary from corpora. *Proc. of 31st Assoc. for Comp. Ling.* 235-242, Morgan-Kaufmann, Palo Alto, CA.

Manning, C. and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA.

McCarthy, D. and Korhonen, A. 1998. Detecting verbal participation in diathesis alternations. *Proc. of 36th Assoc. For Comp. Ling.*, Morgan-Kaufmann. , Palo Alto, CA.

McCarthy, D. 2000. Using semantic preferences to identify verbal participation in role switching alternations. *Proc. of 1st Nth. Am. Assoc. For Comp. Ling.* Morgan-Kaufmann. , Palo Alto, CA.

Roland, D. and Jurafsky, D. 1998. How Verb Subcategorization Frequencies are Affected by Corpus Choice. *Proc. of 36th Assoc. for Comp. Ling.*, Morgan-Kaufmann, Palo Alto, CA.

Sampson. G. 1995. *English for the Computer*. Oxford University Press, Oxford.

Stevenson, S. and Merlo, P. 1999. Automatic verb classification using distributions of grammatical features. *Proc. of 9th Conf. Of Eur. Assoc. For Comp. Ling.* 45-52., Morgan-Kaufmann, Palo Alto, CA.

Ushioda, A., Evans, D., Gibson, T. and Waibel, A. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text*, Columbus, Ohio.

Wilks, Y. and Stevenson, M. 1998. Word sense disambiguation using optimised combinations of knowledge sources. *Proc. of 36th Assoc. for Comp. Ling.*, Morgan-Kaufmann, Palo Alto, CA.

Zeman, D. and Sarkar, A. 2000. Automatic extraction of subcategorization frames for Czech. *Proc. of Int. Conf. On Comp. Ling.*, 691-697, Saarbrucken, Germany.

Zipf, G. 1949. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA.