Controlling the application of lexical rules

Ted Briscoe
ejb@cl.cam.ac.uk
Computer Laboratory, University of Cambridge

Ann Copestake
aac@csli.stanford.edu
CSLI, Stanford University

Abstract

In this paper, we describe an item-familiarity account of the semi-productivity of morphological and lexical rules, and illustrate how it can be applied to practical issues which arise when building large scale lexical knowledge bases which utilize lexical rules. Our approach assumes that attested uses of derived words and senses are explicitly recorded, but that productive use of lexical rules is also possible, though controlled by probabilities associated with rule application. We discuss how the necessary probabilities and estimates of lexical rule productivity may be acquired from corpora.

1 Introduction

Lexicalist linguistic theories, such as HPSG, LFG and categorial grammar, rely heavily on lexical rules. Recently, techniques have been described which address the efficiency issues that this raises for fully productive rules, such as inflectional rules and 'syntactic rules' (such as the HPSG complement extraction lexical rule). For example, Bouma & van Noord (1994) and Johnson & Dorre (1995) propose techniques for delayed evaluation of lexical rules so that they apply 'on demand' at parse time. Meurers & Minnen (1995) present a covariation approach, in which a finite-state machine for the application of lexical rules is derived by computing possible follow relations between the set of rules and then pruned FSMs are associated with classes of actual lexical entries representing the restricted set of rules which can apply to those entries. Finally, entries themselves are extended with information common to all their derived variants. These techniques achieve most of the advantages of lexicon expansion in the face of recursive rules and cyclic rule interactions which preclude a full off-line expansion.

Although these treatments allow for the efficient use of productive lexical rules, they do not address the issue of semi-productivity of derivational morphological and sense extension rules, which causes considerable problems in construction of broad coverage lexical knowledge bases (LKBs) (see, for example, Climent and Martí (1995), Pirelli *et al.*, 1994). The standard formalization of lexical rules entails that derived entries will exist without exception for any basic entry which is compatible with the lexical rule input description. Formal accounts of some classes of exceptions, such as preemption by synonomy, have been developed (e.g. Briscoe et al, 1995), but these suffer from the disadvantage that detailed lexical semantic information must be available to detect potential synonyms. The search for a fully productive statement of verb alternations has led to an increasingly semantic perspective on such rules. Pinker (1989) argues that so-called broad semantic classes (e.g. creation or transfer verbs) provide necessary conditions for lexical rule application, but that narrow class lexical rules should be specified, breaking down such rules into a number of fully-productive subcases. But, in the attempt to define such subcases, Pinker is forced to make subtle and often unintuitive distinctions. Similarly, Levin (1992) delimits classes of verbs to which particular sets of alternations apply, but some of her classes are very small and do not have straightforward semantic criteria for membership. Thus, even if the narrow class approach is correct, its implementation is problematic.

From a computational perspective, an equally acute problem is the proliferation of senses that results when lexical rules are encoded as fully productive. For instance, the result of applying the `vehicle-name -> verb-of-motion` lexical rule can be input to several other lexical rules. The forms which would arise if the alternations given by Levin (1992:267) are applied to *helicopter* are illustrated in (1):

(1) a The pilot helicoptered

b The pilot helicoptered over the forests

c Mrs Clinton was helicoptered to the base

d The pilot helicoptered the forests

e The pilot helicoptered his passengers sick

Judgements of the grammaticality of such examples differ (though (1c) is an attested example) but even when such senses are plausible and attested, they are rare for the great majority of nouns which could in principle undergo the conversion.

Jackendoff (1975) and others have proposed that lexical rules be interpreted as redundancy statements which abbreviate the statement of the lexicon but which are not applied generatively. This conception of lexical rules has been utilized in computational lexical knowledge bases, for example by Sanfilippo (1993). However, this approach cannot account for the *semi*-productive nature of such rules, illustrated with respect to the dative alternation in (2):

(2) John faxed / xeroxed / emailed his colleagues a copy of the report

And for practical LKB building, there is a problem acquiring the information about which lexical entries a rule applies to. Machine readable dictionaries (MRDs) were used for this purpose by Sanfilippo, but the absence of a sense in an MRD does not mean it is unknown to the lexicographer: dictionaries have space limitations and senses may be omitted if they are rare or specialized, and also if they are 'obvious' — i.e. the result of a highly productive process (Kilgarriff, 1992). Furthermore, if broad-coverage is attempted, the polysemy problem is still acute. Finally, theories of the lexicon in which the consequences of lexical rules are precomputed, cannot be correct in the limit because of the presence of recursive lexical rules such as *re-*, *anti-* or *great-* prefixation (e.g. *rereprogram*, *anti-anti-missile* or *great-great-grandfather*).

Thus neither the interpretation of lexical rules as fully generative or as purely abbreviatory is adequate linguistically or as the basis for LKBs. Although many lexical rules are subject to exceptions, gaps and variable degrees of conventionalization, most are semi-productive in the sense that they play a role in the production and interpretation of nonce forms and errors. In the remainder of this paper, we illustrate how the linguistically-motivated probablistic framework for lexical rule application described in Copestake and Briscoe (1995) and Briscoe and Copestake (1995) might be utilized to address these practical problems.

2 Probabilistic lexical rules

Copestake and Briscoe (1995) and Briscoe and Copestake (1995) argue that lexical rules are sensitive to both type and token frequency effects which determine language users' assessments of the degree of acceptability of a given derived form and also their willingness to apply a rule in producing or interpreting a novel form. Arguments for a treatment of semi-productivity along these lines have been advanced by Goldberg (1995) and Bauer (1983) (though not with respect to lexical rules). We regard our use of probabilities as being consistent with Bauer's claim that accounting for semi-productivity is an issue of performance, not competence (Bauer 1983:71f).

The frequency with which a given word form is associated with a particular lexical entry (i.e. sense or grammatical realization) is often highly skewed; Church (1988) points out that a model of part-of-speech assignment in context will be 90% accurate (for English) if it simply chooses the lexically most frequent part-of-speech for a given word. Briscoe and Carroll (1995) found in one corpus that there were about 18 times as many instances of *believe* in the most common subcategorization class as in the 4 least common classes combined. In the absence of other factors, it seems very likely that language users utilize frequency information to resolve indeterminacies in both generation and interpretation. Such a strategy is compatible with and may well underlie the Gricean Maxim of Manner, in that ambiguities in language will be more easily interpretable if there is a tacit agreement not to utilize abnormal or rare means of conveying particular messages. We can model this aspect of language use as a conditional probability that a word form will be associated to a specific lexical entry:

$$\text{Prob}(\text{lexical-entry} \mid \text{word-form}) = \frac{freq(\text{lexical-entry with word-form})}{freq(\text{word-form})}$$

This proposal is not novel and is the analogue of proposals to associate probabilities with initial trees in a Lexicalized Tree Adjoining Grammar (Resnik, 1992; Schabes, 1992). The derivation probability which gives the probability of a particular sentence interpretation will depend on the product of the lexical probabilities (rule probabilities might also play a role, but can be ignored in the categorial framework we adopt here).

Lexical probabilities are acquired for both basic and derived lexical entries independently of the lexical rules used to create derived entries, so a derived entry might be more frequent than a basic one. Basic entries are augmented with a representation of the attested lexical rules which have applied to them and any such derived chains, where both the basic entry and these 'abbreviated' derived entries are associated with a probability. One way of implementing this approach is to adopt the covariation technique of Meurers & Minnen (1995) discussed above. If we assume a precompiled representation of this form, conditional probabilities that a word form will be associated with a particular (basic or derived) entry can be associated with states in the FSM, as illustrated in Figure 1. (The feature structure itself is based on the verb representation scheme developed by Sanfilippo (1993), though the details are unimportant for current purposes.)

In this representation, the states of the FSM, which have been given mnemonic names corresponding to their types, are each associated with a probability representing the relative likelihood that *fax* will be associated with the derived entry which results from applying the rule to the source entry (the probabilities shown here are purely for illustrative purposes). We call this representation the lexeme for a given word. Figure 2 shows part of the corresponding FSM explicitly. Note that there are states with no associated probabilities, reflecting possible but unattested usages. The topology of the FSM associated with a given word may be shared with other words, but the specific probabilities associated with the states representing lexical entries will be idiosyncratic so that the each lexeme representation must minimally encode the unique name of the relevant FSM and a probability for each attested state / lexical entry as shown in Figure 1. If the derived form is irregular in some way, then the exceptional information can be stipulated at the relevant state, and the feature structure calculated by default-unifying the specified information with the productive output of the lexical rule. For example, if *beggar* is treated as derived by the agentive *-er* rule (which

$$\begin{bmatrix} \textbf{trans-causative-verb} \\ \text{ORTH} : \textbf{fax} \\ \text{SYN} : \begin{bmatrix} \text{RESULT} : \begin{bmatrix}\text{RESULT} : \textbf{ssign} \\ \text{ACTIVE} : \textbf{npsign}\,\boxed{2}\end{bmatrix} \\ \text{ACTIVE} : \textbf{npsign}\,\boxed{3}\end{bmatrix} \\ \text{SEM} : < \begin{bmatrix}\textbf{fax-rel}\\ \text{EVENT} : \textbf{e}\end{bmatrix}, \begin{bmatrix}\textbf{p-agt-caus}\\ \text{EVENT} : \textbf{e}\\ \text{ARG} : \boxed{3}\end{bmatrix} \begin{bmatrix}\textbf{p-pat-aff-obj}\\ \text{EVENT} : \textbf{e}\\ \text{ARG} : \boxed{2}\end{bmatrix} > \end{bmatrix}$$

create/transfer-lexeme-fsm(trans(0.2), to-ditrans(0.3) ...)

Figure 1: Lexeme for *fax*

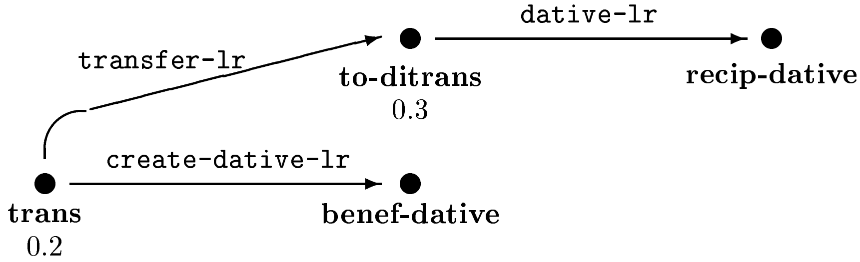

Figure 2: FSM for *fax*

is reasonable synchronically), then the irregular morphology can be stipulated and will override the predicted *begger*.

The resulting FSM is not a Markov model because probabilities on states represent output probabilities and not transition probabilities in the machine. In addition, since the probabilities encode the relative likelihood that a given word *form* will associate with a particular lexical entry, the set of probabilities on states of a FSM will not be globally normalized. One FSM will represent the application of both rules of conversion (zero affixation) and rules of derivation to a given lexeme and the latter will change the form of the word, and thus participate in a different distribution. See for example, Figure 3, which is intended to cover the noun and verb *lacquer*, plus the derived form, *lacquerer* (with agentive and instrument readings taken as distinct).

One problem with the acquisition of reliable estimates of such probabilities is that many possibilities will remain unseen and will, therefore, be unattested. There are a variety of well-known techniques for smoothing probability distributions which avoid assigning zero probability to unseen events. Church & Gale (1994) discuss the applicability of these to linguistic problems and emphasize the need for differential estimation of the probability of different unseen events in typical linguistic applications. For instance, one standard approach to smoothing involves assigning a hypothetical single observation to each unseen event in a distribution before normalizing frequencies to obtain probabilities. This captures the intuition that the more frequent the observation of some events in a distribution, the less likely it is that the unseen possibilities will occur. Thus, a rare word with only a few observations may be more likely to be seen in an alternative realization than a very frequent word which has been observed many times in some subset of the possible realizations licensed by the grammar. However, all unseen events will be assigned the same probability within each distinct distribution and this is at best a gross estimate of the actual distribution.

For unattested derived lexical entries for a given word form, the relative productivity of the lexical rule(s) required to produce the derived entry can be used to allow differential estimation

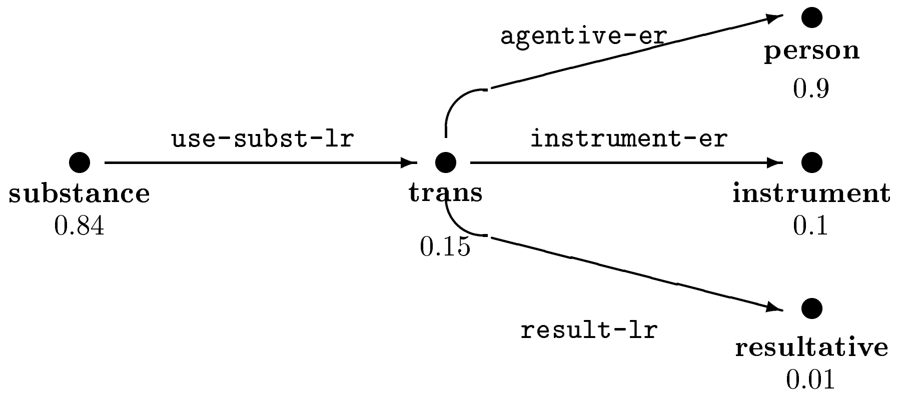

Figure 3: Lexeme for *lacquer*

the probability of an unattested derived entry given a word form. We can estimate the relative productivity of each lexical rule by calculating the ratio of possible to attested outputs for each rule (cf Aronoff, 1976):

$$\text{Prod(lexical-rule)} = \frac{M}{N}$$

(where N is the number of attested lexical entries which match the lexical rule input and M is the number of attested output entries). We discuss some more elaborate measurements for productivity in section 4.

This information concerning degree of productivity of a rule can be combined with a smoothing technique to obtain a variant enhanced smoothing method of the type discussed by Church & Gale (1994) capable of assigning distinct probabilities to unseen events within the same distribution. This can be achieved by estimating the held back probability mass to be distributed between the unseen entries using the basic smoothing method and then distributing this mass differentially by multiplying the total mass for unseen entries (expressed as a ratio of the total observations for a given word) by a different ratio for each lexical rule. This ratio is obtained by dividing the ratio representing the productivity of the lexical rule(s) by the sum of the ratios of the lexical rules required to construct all the unseen entries.

$$\text{Unseen-pr-mass(word-form)} = \frac{\text{number-of-unattested-entries(word-form)}}{freq\text{(word-form)} + \text{number-of-unattested-entries(word-form)}}$$

$$\text{Est-freq(lex-entry}_i \text{ with word-form}_j) = \text{Unseen-pr-mass(word-form}_j) \times \frac{Prod(lr_i)}{\sum Prod(lr_1), \ldots, Prod(lr_n)}$$

(where $lr_1 \ldots lr_n$ are the n lexical rules needed to derive the n unattested entries for word-form$_j$) This will yield revised ratios for each given word which can then be normalized to probabilities.

To make this clearer, consider the use of the probabilities to drive interpretation in the case of a nonce usage. Consider the lexical entry for the verb *fax* given in Figure 1 and assume the verb is unattested in a dative construction, such as *fax me the minutes of the last meeting.* But it may undergo either the benefactive-dative or recipient-dative rules to yield a dative realization. These rules would produce either a deputive reading where although the speaker is a beneficiary of the action the recipient is unspecified or a reading where the speaker is also the recipient of the transfer action. Choosing between these rules in the absence of clear contextual information could be achieved by choosing the derivation (and thus interpretation) with highest probability. This would depend solely on the relative probability of the unseen derived entries created by applying these two rules to *fax*. This would be (pre)computed by applying the formulae above to a representation of the lexeme for *fax* in which ratios represent the number of observations of an entry for a given word form over the total number of observations of that word form, and unattested entries are noted and assigned one observation each

$$\text{create/transfer-lexeme-fsm}(\text{trans}(\frac{20}{100}), \text{for-ditrans}(\frac{30}{100}), \text{recip-dative}(\frac{1}{100}), \text{benef-dative}(\frac{1}{100}), \ldots)$$

Now if we assume that the recipient dative rule can apply to 100 source entries and the resulting derived entries are attested in 60 cases, whilst the benefactive dative can apply to 1000 entries and the derived entries are attested in 100 cases, we can compute the revised estimates of the probabilities for the unseen entries for *fax* by instantiating the formula for estimated frequency as follows:

$$\text{Est-freq(fax with recipient-dative)} = \frac{2}{100} \times (\frac{1}{\sum(\frac{60}{100}, \frac{100}{1000})} \times \frac{60}{100})$$

and similarly for the benefactive-dative case. The resulting ratios can then be converted to probabilities by normalizing them along with those for the attested entries for *fax*. In this case, the recipient reading will be preferred as the recipient dative rule is more productive.

This general approach handles the possibility of specialized subcases of more general rules. For example, we could factor the computation of productivity between subtypes of the input type of a rule and derive more fine-grained measures of productivity for each narrow class a rule applies to. In the case of specialized subcases of lexical rules which apply to a narrower range of lexical items but yield a more specific interpretation (such as the rules of Meat or Fur grinding as opposed to Grinding proposed in Copestake & Briscoe, 1995), the relative productivity of each rule will be estimated in the manner described above, but the more specialized rule is likely to be more productive since it will apply to fewer entries than the more general rule. Similarly, in Figure 3, we assumed a `use-substance` lexical rule, but a more accurate estimation of probabilities is obtained by considering specialized subclasses, as we will see in the next section.

3 Acquiring probabilities

In order to implement the approach described, it is necessary to acquire probabilities for attested senses, and to derive appropriate estimates of lexical rule productivity. Probabilities of different word senses can be learned by a running analyzer, to the extent that lexical ambiguities are resolved either during processing or by an external oracle, and for limited domains this may well be the best approach. We are more interested in incorporating probabilities in a large, reusable, lexical knowledge base. Recent developments in corpus processing techniques have made this more feasible. For instance, work on word sense disambiguation in corpora (e.g. Resnik 1995), could lead to an estimate of frequencies for word senses in general, with rule-derived senses simply being a special case. Many lexical rules involve changes in subcategorization, and automatic techniques for extracting subcategorization from corpora (e.g. Briscoe and Carroll, 1995; Manning, 1993) could eventually be exploited to give frequency information.

In some cases, a combination of large corpora and sense taxonomies can be used to provide a rough estimate of lexical rule productivity suitable for instantiating the formulae given in the previous section. For example, we examined verbs derived from several classes of noun from the 90 million word written portion of the British National Corpus, using the wordlists compiled by Adam Kilgarriff. We looked at four classes of nouns: vehicles, dances, hitting weapons (e.g. *club*, *whip*) and decorative coatings (e.g. *lacquer*, *varnish*). For the sake of this experiment, we assumed that these undergo four different lexical rules[1]:

- vehicle -> go using vehicle (Levin, 1992 : 51.4.1)
- dance -> perform dance ((Levin : 51.5)
- hitting weapon -> hit with weapon (subclass of Levin : 18.3)
- paint-like substance -> apply paint-like substance (Levin : 24)

The first problem is isolating the nouns which can be input to the lexical rules. For the purposes of deriving a productivity measurement for the rule as a whole, it does not matter much if the set is incomplete, as long as there are no systematic differences in productivity between the included and the excluded cases. There are several potential sources for semantically coherent noun classes. The list of vehicle nouns was derived from a taxonomy constructed semi-automatically from Longman Dictionary of Contemporary English (Procter, 1978), as described by Copestake (1990). This taxonomy only included land vehicles, not boats or airplanes. The other three classes were derived manually from a combination of Roget's and WordNet, since the relevant taxonomies were not available. For the 'hitting weapon' and 'paint-like substance' classes, this involved combining several Roget categories and WordNet synsets. We excluded entries made up of more than one word, such as *square dance* and also pruned the set of nouns to exclude cases where a non-derived verb form would confuse the results (e.g. *taxi*).

Initially we used the automatically assigned part of speech tags to identify verbs, but these gave a large number of false positives, because of errors in the tagging process. Therefore we looked instead for forms ending in *-ed* and *-ing* which had been tagged as verbs. There is still the potential for false positives if an adjectival *-ed* form (e.g. *bearded*) was mistakenly tagged as a verb, but this did not appear to cause a problem for these experiments. Only considering inflected forms means that we are systematically underestimating frequencies, but since the main aim is to acquire the correct relative ordering of lexical rules, this is not too problematic. Figure 4 shows some raw frequencies of noun and verb (*-ed*, and *-ing* form) from the BNC. We also show frequencies of the *-er* nominal, which we assume is derived from the verb form. For comparison, we show whether the word is found in the Cambridge International Dictionary of English (CIDE), a modern learner's dictionary. A more sophisticated system for acquisition of accurate frequencies for each word would have to be capable of sense-disambiguation. For example, according to Figure 4, *distemper* was found as a noun 37 times, but many of these uses actually referred to the disease, rather than the paint.

We assumed that a unique conversion rule applied to each noun and calculated the productivities of the lexical rules as the ratio of the number of words for which verbs were found over the total number of words in the class which were found in the corpus. The results are summarized in Figure 5. The results for vehicle nouns were manually checked to ensure that the unusual verb forms were genuine. This resulted in one putative example of the conversion rule being discarded: *trailered* and *trailering* were found in one section of the corpus, but turned out to refer to getting

[1] It is irrelevant here exactly how these rules are to be formalized, though see references in Levin (1992) and also Kiparsky (1996). It is also not essential to our approach that these rules be treated as distinct, from the viewpoint of their representation as typed feature structures, since it would be possible to attach probabilities to subrules which only differed in the semantic type of their input.

word	BNC			CIDE		
	noun	verb	-er	noun	verb	-er
acrylic	104	0	0	+	-	-
calcimine	0	0	0	-	-	-
distemper	37	4	0	+	-	-
dye	291	153	19	+	+	-
emulsion	117	2	0	+	-	-
enamel	287	25	1	+	+	-
gouache	85	0	0	+	-	-
japan	17	1	0	-	-	-
lacquer	132	25	1	+	+	-
paint	1783	2170	965	+	+	+
primer	236	0	0	+	-	-
semigloss	0	0	0	-	-	-
shellac	35	1	0	-	-	-
tempera	29	0	0	+	-	-
tint	60	46	0	+	+	-
undercoat	44	0	0	+	-	-
varnish	231	37	1	+	+	-
veneer	156	12	0	+	-	-
watercolour	272	0	0	+	-	-

Figure 4: Raw frequencies for some *paint* nouns.

class	total	alternating	productivity
vehicle	75	12	0.16
dance	41	9	0.22
paint	23	15	0.65
club	35	18	0.51

Figure 5: Productivity estimates

a horse into a trailer, rather than transporting by trailer. In other words, *trailer* here is being regarded as a container or location, rather than as a vehicle. Manual checking of the rare derived forms is not particularly time-consuming, so a semi-automatic approach, where high frequency forms which are found in an MRD are assumed to be genuine, but where low frequency examples are manually checked, should be adequate.

As expected, some very frequent nouns such as *car* and *vehicle* had no corresponding verbs. Of course we could hypothesize that verb formation is preempted by synonymy (e.g. by *drive*). But, whatever the cause, blocking is allowed for automatically by the approach proposed here, since the probability calculated for unseen entries of high frequency words will be very low (see section 4). Similarly, it should not be necessary to explicitly encode the fact that the conversion rule does not apply to an already derived form such as *primer*.

Even with a 90 million word corpus, some words occurred very infrequently, and others which were found in Roget's and/or WordNet were absent completely. For example *calcimine* is defined in WordNet as a type of water-based paint, and is also found in Roget's, but does not occur in the BNC. Even the relative estimates for productivity of rules will be inaccurate if there is a systematic difference between the frequency of words in one input class as compared to another, since infrequently occurring words are less likely to have attested derived forms. We discuss modifications to the formulae which would allow for this in the next section. This effect might have accounted for the relatively low productivity observed for the dance rule. However, there might also be phonological effects since many dance names are taken from languages other than English. The results for productivity are only strictly comparable within a particular corpus. It should be apparent from the frequencies that large corpora are needed to find instances of some words.

4 Utilizing probabilistic lexical rules

The majority of implemented NLP systems have either simply listed derived forms and extended senses, or treated them using lexical rules as redundancy statements. In the introductory section, we argued that this approach cannot be correct in principle, because of the problem of nonce senses. But it is also demonstrably inadequate, at least for systems which are not limited to a narrow domain. In an experiment with a wide-coverage parsing system (Alvey NL Tools, ANLT) Briscoe and Carroll (1993) observed that half of the parse failures were caused by inaccurate subcategorization information in the lexicon. The ANLT lexicon was derived semi-automatically from a machine readable dictionary (LDOCE), and although the COMLEX syntax dictionary (Grishman *et al.*, 1994), which was derived with much greater amounts of human effort, has a slightly better performance, the difference is not great. Automatic acquisition of information from corpora is a partial answer to this problem, and one which is in many respects complementary to the approach assumed here, but successful acquisition of a broad-coverage lexicon from a really large corpus would lead to a similar problem of massive ambiguity as we see in the case of productive lexical rules. Control of syntactic ambiguity by the use of lexical and other probabilities has been demonstrated by several authors (e.g. Black *et al.*, 1993; Schabes, 1992; Resnik, 1992), but the difficulty of acquisition means that the validity of utilizing lexical probabilities of the type assumed here has not yet been demonstrated on a large scale.

This approach fits in most naturally with systems where probabilistic information is incorporated systematically. However it could be useful with more traditional systems. Different applications could utilize probabilistic information in different ways. For word choice in generation, it would be appropriate to take the highest-probability suitable entry, and, if none are attested, to construct a phrase, rather than apply a semi-productive lexical rule to produce a nonce form. For analysis, the most likely rules can be applied first, in the case of known senses, and since nonce senses are (by definition) rarer, rules will be applied productively only when this fails. This improves on the control principle suggested in Copestake (1992), that lexical rules should only be applied if no interpretation was applicable which did not involve a lexical rule, since it allows for cases such as *turkey*, where the derived (meat) use is more frequent than the non-derived (animal) use in the corpora which we have examined. The two other control effects suggested in Copestake (1992) are both also superseded by the current proposal. One of these was to allow for blocking, which is discussed below. The other was that more specific lexical rules should be preferred over more general ones. We would expect that, in general, the more specialized rule will be more productive, as a natural consequence of applying to a smaller class, but the earlier proposal would have had the undesirable consequence that this was a fixed consequence, which could not be adjusted for cases where the generalization did not hold. Thus the grammar writer was, in effect, required to consider both competence and performance when stipulating a rule.

The general claim we make here is that if we assume that speakers choose well-attested high-frequency forms to realize particular senses and listeners choose well-attested high-frequency senses when faced with ambiguity, then much of the 'semi-productivity' of lexical rules is predicted. Blocking can be treated as a special case of this principle: if speakers use higher frequency forms to convey a given meaning, an extended meaning will not become conventionalized if a common synonym exists. This means that we do not have to stipulate a separate blocking principle in interpretation, since the blocked senses will not be attested or will have a very low frequency. And in generation, we assume that higher probability forms are preferred as a way of conveying a given meaning. Practically, this has considerable advantages over the earlier proposal, that blocking should be detected by looking for synonyms, since the the state of the art in acquisition and representation of lexical semantic information makes it difficult to detect synonymy accurately. We can assume, for example, that a verbal use of *car* will not be postulated by a generator, because it is unattested, and will only be possible for an analyzer when forced by context. It is necessary to allow for the possibility of *unblocking*, because of examples such as the following:

(3) a There were five thousand extremely loud people on the floor eager to tear into roast cow with both hands and wash it down with bourbon whiskey.
(Tom Wolfe, 1979. *The Right Stuff*)

b In the case of at least one county primary school ... they were offered (with perfect timing) saute potatoes, carrots, runner beans and roast cow.
(Guardian newspaper, May 16th 1990, in a story about mad cow disease.)

However, this is not the complete story, since we have not accounted formally for the extra implicatures that the use of a blocked form conveys, nor have we allowed for the generation of blocked forms (apart from in the circumstances where the generator's lexicon omits the synonym). Both these problems require an account of the interface with pragmatics, though the latter is perhaps not serious for computational applications, since we are unlikely to want to generate blocked forms.

The treatment proposed here is one of many possible schemes for estimating the productivity of lexical rules and integrating these estimates with the estimation of the probabilities of unseen entries for given word forms. Other more complex schemes could be developed, which, for example, took account of the average probability of the output of a lexical rule. This might be necessary, for example, to model the relative frequencies of *-er* vs *-ee* suffixation, since although the latter is more productive (by Baayen and Lieber's (1991) definition), tokens of the former are more frequent overall (Barker, 1996). However, we will assume the simple approach here, since acquiring the average probability of lexical rule output raises some additional difficulties, and we currently have no evidence that the more complex approach is justified, given that our main aim is to rank unseen senses by plausibility. Another problem, mentioned above, is the need to ensure that classes have comparable frequency distributions. This could matter if there were competing lexical rules, defined on different but overlapping classes, and if one class has a very high percentage of low frequency words compared to the other, the estimate of its productivity will tend to be lower. The productivity figure could be adjusted to allow for item frequency within classes, but we will not discuss this further here.

5 Conclusion

In this paper, we have described a possible approach to application-independent technique for controlling lexical rule application. We have concentrated on sense-extension, but the same machinery could be used for derivational morphology, with the advantage that acquiring frequencies from corpora is easier, at least for unambiguous affixes. Our approach requires some lexical semantic information, to identify possible inputs to rules, but the need for detailed definitions of narrow classes for which rules can be treated as fully productive is reduced (since failure to identify a narrow class will lead to less accurate prediction of probabilities, rather than over-generation) as is the requirement to identify synonyms to predict blocking. The probabilistic approach to lexical rules integrates neatly with existing proposals to control application of lexical rules efficiently within a constraint-based framework, such as those of Meurers & Minnen (1995). Thus we believe that this approach could provide a linguistically motivated and practical solution to the problem of semi-productivity. However, further work remains to be done on acquiring sense frequencies and productivity measurements, before evaluation in a full system is feasible.

References

Aronoff, M. (1976) *Word Formation in Generative Grammar,* Linguistic Inquiry Monograph 1. MIT Press, Cambridge, Mass.

Baayen, H. and R. Lieber (1991) 'Productivity and English Derivation: A Corpus-Based Study', *Linguistics, vol.29,* 801–843.

Barker, C. (1996) 'Episodic *-ee* in English: Thematic relations and new word formation' in M. Simons and T. Galloway (eds.), *Semantics and Linguistic Theory V,* Cornell University, Ithaca, NY, pp. 1–18.

Bauer, L. (1983) *English word-formation,* Cambridge University Press, Cambridge, England.

Black, E., F. Jelinek, J. Lafferty, D.M. Magerman, R. Mercer, S. Roukos (1993) 'Towards history-based grammars: using richer models for probabilistic parsers', *Proceedings of the 31st Annual Meeting of the Association of Computational Linguistics,* Ohio State University, Columbus, Ohio, pp. 31–37.

Bouma, G. and G. van Noord (1994) 'Constraint-based categorial grammar', *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics,* Las Cruces, NM.

Briscoe, E.J. and J. Carroll (1993) 'Generalized probabilistic LR parsing for unification-based grammars', *Computational Linguistics, vol.19.1,* 25–60.

Briscoe, E.J. and J. Carroll (1995) *Towards automatic extraction of argument structure from corpora,* ACQUILEX II Working Paper.

Briscoe, E.J., A. Copestake and A. Lascarides (1995) 'Blocking' in P. St. Dizier and E. Viegas (eds.), *Computational lexical semantics,* Cambridge University Press, pp. 273–302.

Briscoe, E.J. and A. Copestake (1995) *Dative constructions as lexical rules in the TDFS framework,* ACQUILEX II Working Paper.

Church, K. (1988) 'A stochastic parts program and noun phrase parser for unrestricted text', *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-88),* Austin, Texas, pp. 136–143.

Church, K. and W. Gale (1994) 'Enhanced Good-Turing and Cat-Cal estimation', *Speech and Language Processing, vol.10,*

Climent, S. and M. A. Martí (1995) *On using lexical rules with semantic effects,* Paper presented at ACQUILEX Workshop on lexical rules.

Copestake, A. (1990) 'An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary', *Proceedings of the First International Workshop on Inheritance in Natural Language Processing,* Tilburg, The Netherlands, pp. 19–29.

Copestake, A. and E.J. Briscoe (1995) 'Semi-Productive Polysemy and Sense Extension', *Journal of Semantics, vol.12,* 15–67.

Goldberg, A. (1995) *Constructions,* Chicago University Press.

Grishman, R., Macleod, C. & Meyers, A. (1994) 'Comlex syntax: building a computational lexicon', *Proceedings of the International Conference on Computational Linguistics, COLING-94,* Kyoto, Japan, pp. 268–272.

Jackendoff, R. (1975) 'Morphological and semantic regularities in the lexicon', *Language, vol.51(3),* 639–71.

Johnson, M. and J. Dorre (1995) 'Memoization of coroutined constraints', *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics,* Cambridge, MA, pp. 100–107.

Kilgarriff, A. (1992) *Polysemy,* D.Phil. thesis, Cognitive Science Research Paper 261, University of Sussex, UK.

Kiparsky, P. (1996) *Remarks on Denominal Verbs,* Stanford University.

Levin, Beth (1992) *Towards a lexical organization of English verbs,* Chicago University Press, Chicago.

Manning, C. (1993) 'Automatic acquisition of a large subcategorisation dictionary from corpora', *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-93),* Columbus, Ohio.

Meurers, D. and G. Minnen (1995) 'A computational treatment of HPSG lexical rules as covariation lexical entries', *Proceedings of the 5th Int. Workshop on Natural Language Understanding and Logic Programming,* Lisbon.

Pinker, S (1989) *Learnability and Cognition: The Acquisition of Argument Structure,* MIT Press, Cambridge, MA.

Pirelli, V., N. Ruimy and S. Montemagni (1994) *Lexical regularities and lexicon compilation,* Acquilex-II Working Paper 36.

Procter, P. (ed) (1978) *Longman Dictionary of Contemporary English,* Longman, London.

Resnik, P. (1992) 'Probabilistic Lexicalised Tree Adjoining Grammar', *Proceedings of the Coling92,* Nantes, France.

Resnik, Philip (1995) 'Disambiguating noun groupings with respect to WordNet senses', *Proceedings of the 3rd workshop on very large corpora,* MIT, Cambridge, Mass..

Sanfilippo, A. (1993) 'LKB encoding of lexical knowledge from machine-readable dictionaries' in E. J. Briscoe, A. Copestake and V. de Paiva (eds.), *Inheritance, defaults and the lexicon,* Cambridge University Press, Cambridge, England, pp. 190–222.

Schabes, Y. (1992) 'Stochastic Lexicalized Tree Adjoining Grammar', *Proceedings of the Coling92,* Nantes, France.