ENCODING FREQUENCY INFORMATION IN LEXICALIZED GRAMMARS

John Carroll **David Weir**
School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton, BN1 9QH, UK

{johnca,davidw}@cogs.susx.ac.uk

Abstract

We address the issue of how to associate frequency information with lexicalized grammar formalisms, using Lexicalized Tree Adjoining Grammar as a representative framework. We consider systematically a number of alternative probabilistic frameworks, evaluating their adequacy from both a theoretical and empirical perspective using data from existing large treebanks. We also propose three orthogonal approaches for backing off probability estimates to cope with the large number of parameters involved.

1 Introduction

When performing a derivation with a grammar it is usually the case that, at certain points in the derivation process, the grammar licenses several alternative ways of continuing with the derivation. In the case of context-free grammar (CFG) such nondeterminism arises when there are several productions for the nonterminal that is being rewritten. Frequency information associated with the grammar may be used to assign a probability to each of the alternatives. In general, it must always be the case that at every point where a choice is available the probabilities of all the alternatives sum to 1. This frequency information provides a parser with a way of dealing with the problem of ambiguity: the parser can use the information either to preferentially explore possibilities that are more likely, or to assign probabilities to the alternative parses.

There can be many ways of associating frequency information with the components making up a grammar formalism. For example, just two of the options in the case of CFG are: (1) associating a single probability with each production that determines the probability of its use wherever it is applicable (i.e. Stochastic CFG; SCFG (Booth and Thompson, 1973)); or (2) associating different probabilities with a production depending on the particular nonterminal occurrence (on the RHS of a production) that is being rewritten (Chitrao and Grishman, 1990). In the latter case probabilities depend on the context (within a production) of the nonterminal being rewritten. In general, while there may be alternative ways of associating frequency information with grammars, the aim is always to provide a way of associating probabilities with alternatives that arise during derivations.

This paper is concerned with how the kind of frequency information that would be useful to a parser can be associated with lexicalized grammar formalisms. To properly ground the discussion we will use Lexicalized Tree Adjoining Grammar (LTAG) as a representative framework, although our remarks can be applied to lexicalized grammar formalisms more generally. We begin by considering the derivation process, and, in particular, the nature of derivation steps. At the heart of a TAG is a finite set of trees (the elementary trees of the grammar). In an LTAG these trees are 'anchored' with lexical items and the tree gives a possible context for its anchor by providing a structure into which its complements and modifiers can be attached. For example, Figure 1 shows four elementary trees—one *auxiliary* tree β and three *initial* trees α_1, α_2 and α_3. Nodes marked with asterisks and downarrows are foot and substitution nodes, respectively. In a derivation these trees are combined using the operations of substitution and adjunction to produce a derived tree for a complete sentence. Figure 2 shows a single derivation step in which α_2 and α_3 are substituted at frontier nodes (with addresses 1 and $2 \cdot 2$, respectively) of α_1 and β is adjoined at an internal node of α_1 (with address 2)[1].

[1] The root of a tree has the address ϵ. The ith daughter (where siblings are ordered from left to right) of a node with address a has address $a \cdot i$

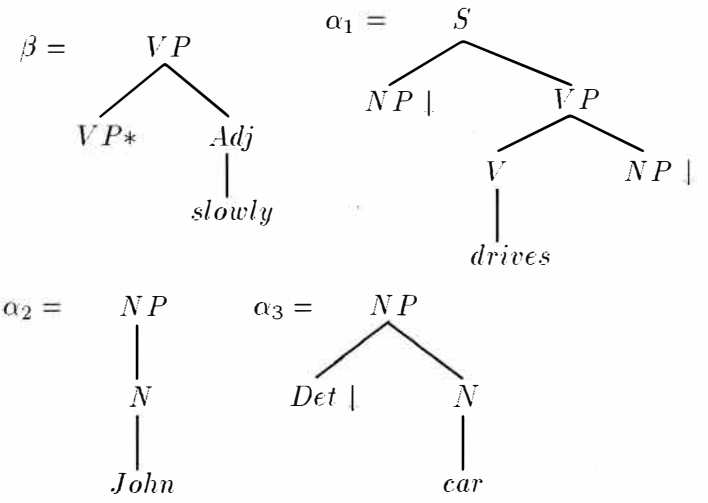

Figure 1: An Example Grammar

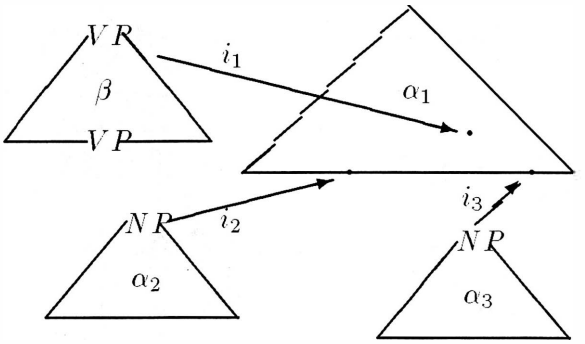

Figure 2: A Derivation Step

When formalizing LTAG derivations, a distinction must be made between the (object-level) trees that are derived in a derivation and the (meta-level) trees that are used to fully encode what happens in derivations. These trees are referred to as derived and derivation trees, respectively. A scheme for encoding TAG derivations was proposed by Vijay-Shanker (1987) and later modified by Schabes and Shieber (1994). Derivation trees show, in a very direct way, how the elementary trees are combined in derivations. Nodes of the derivation trees are labeled by the names of elementary trees, and edge labels identify tree addresses (i.e. node locations) in elementary trees. Figure 3 shows the derivation tree resulting from the derivation step in Figure 2. The nodes identified in the derivation tree encode that when the elementary tree α_1 was used, the elementary trees $\beta, \alpha_2, \alpha_3$ were chosen to fit into the various complement and modifier positions. These positions are identified by the tree addresses i_1, i_2, i_3 labeling the respective edges, where in this example $i_1 = 2$, $i_2 = 1$ and $i_3 = 2 \cdot 2^2$. In other words, this derivation tree indicates which choice was made as to how the node α_1 should be *expanded*. In general, there may have been many alternatives since modification is usually optional and different complements can be selected.

By identifying the nature of nondeterminism in LTAG derivations we have determined the role that frequency information plays. For each elementary tree of the grammar, frequency information must somehow determine how the probability mass is to be distributed among all the alternative ways of expanding that tree. In section 2 we consider a number of ways in which this frequency information can be associated with a grammar. We then go on to evaluate the degree to which each scheme can, in principle, distinguish the probability of certain kinds of derivational phenomena, using data from existing large treebanks (section 3). We discuss in section 4 how

[2] As Schabes and Shieber (1994) point out matters are somewhat more complex that this. What we describe here more closely follows the approach taken by Rambow, Vijay-Shanker, and Weir (1995) in connection with D-Tree Grammar.

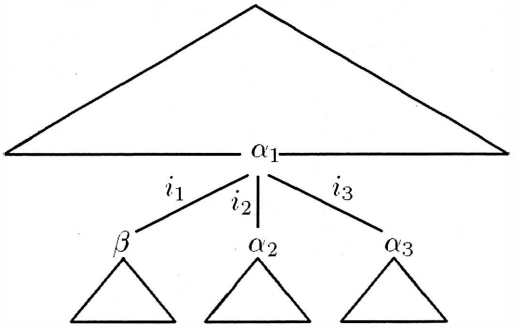

Figure 3: A Derivation Tree

to estimate the large number of probabilistic parameters involved, and propose three orthogonal approaches for smoothing the probability estimates obtained. The paper concludes (section 5) with comparisons with other related work.

2 Frequency Information in Lexicalized Grammars

In this section we consider four ways of associating frequency information with lexicalized grammars. Using the LTAG framework outlined in section 1 as a basis we define four Stochastic Lexicalized Grammar formalisms which we will refer to as SLG(1), SLG(2), SLG(3) and SLG(4). The differences between them lie in how fine-grained the frequency information is, which in turn determines the extent to which the resulting probabilities can be dependent on derivational context.

2.1 Context-Free Frequencies

The first approach we consider is the simplest and will be referred to as **SLG(1)**. A single probability is associated with each elementary tree. This is the probability that that tree is used in a derivation in preference to another tree with the same nonterminal at its root. A grammar is therefore well-formed if, for each nonterminal symbol that can be at the root of a substitutable (adjoinable) tree, the sum of probabilities associated with all substitutable (adjoinable) trees with the same root nonterminal is 1. When nondeterminism arises in a derivation nothing about the derivational context can influence the way that a tree is expanded, since the probability that the various possible trees are adjoined or substituted at each node depends only on the identity of the nonterminal at that node. As a result we say the frequency information in an SLG(1) is *context-free*.

2.2 Node-Dependent Frequencies

The second approach considered here, which we will call **SLG(2)**, has been described before by both Schabes (1992) and Resnik (1992). We describe the scheme of Schabes here, though the approach taken by Resnik is equivalent. In defining his scheme Schabes uses a stochastic version of a context-free-like grammar formalism called Linear Indexed Grammar (LIG). Based on the construction used to show the weak equivalence of TAG and LIG (Vijay-Shanker and Weir, 1994), a LIG is constructed from a given LTAG such that derivation trees of the LIG encode the derived trees of the associated LTAG. Compiling LTAG to LIG involves decomposing the elementary trees into single-level trees and introducing additional productions explicitly encoding every possible adjunction and substitution possibility[3]. It is the LIG productions encoding adjunction and substitution possibilities that are assigned probabilities[4]. The probabilities associated with all the productions that encode possible adjunctions (substitutions) at a node must sum to 1. The key feature of these probability-bearing LIG productions, in the context of the current discussion, is that they encode the adjunction or substitution of a specific elementary tree at a specific place in another elementary tree. This means that the frequency

[3] This scheme has proved useful in the study of LTAG parsing (Schabes, 1990; Vijay-Shanker and Weir, 1993; Boullier, 1996) since this pre-compilation process alleviates the need to do what amounts to the same decomposition process during parsing.

[4] The other productions (that decompose the tree structure) are assigned a probability of 1 since they are deterministic.

information can to some extent be dependent on context. In particular, when faced with nondeterminism in the way that some elementary tree is expanded during a derivation, the probability distribution associated with the alternative adjunctions or substitutions at a given node can depend on which elementary tree that node comes from. As a result we call the frequency information in SLG(2) **node-dependent**. This makes SLG(2) more expressive than SLG(1). As both Schabes and Resnik point out, by leveraging LTAG's extended domain of locality this approach allows probabilities to model both lexical and structural co-occurrence preferences.

The head automata of Alshawi (1996) also fit into the SLG(2) formalism since they involve a dependency parameter which gives the probability that a head has a given word as a particular dependent.

2.3 Locally-Dependent Frequencies

The third approach is **SLG(3)** which falls out quite naturally from consideration of the TAG derivation process. As we discussed in the introduction, LTAG derivations can be encoded with derivation trees in which nodes are labeled by the names of elementary trees and edges labeled by the addresses of substitution and adjunction nodes. The tree addresses can be omitted from derivation trees if a fixed linear order is established on *all* of the adjunction and substitution nodes in each elementary tree and this ordering is used to order siblings in the derivation tree. Given this possibility, Vijay-Shanker, Weir, and Joshi (1987) have shown that the set of derivation trees associated with a TAG forms a local set and can therefore be generated by a context-free grammar (CFG)[5]. The productions of this **meta-grammar** encode possible derivation steps of the grammar. In other words, each meta-production encodes one way of (fully) expanding an elementary tree[6]. In SLG(3) a probability is associated with each of these meta-productions. A SLG(3) is well-formed if for each elementary tree the sum of the probabilities associated with the meta-productions for that tree is 1.

In contrast to SLG(2)—which is limited to giving the probability that a tree anchored with a given lexical item is substituted or adjoined into a tree anchored with a second lexical item—SLG(3) specifies the probability that a particular *set* of lexical items is combined in a derivation step. It is the elementary trees of the underlying LTAG that determine the (extended local) domains over which these dependencies can be expressed since it is the structure of an elementary tree that determines the possible daughters in a meta-production. Although the types of elementary tree structures licensed are specific to a particular LTAG, it might be expected that a SLG(3) meta-grammar, for example, could encode the probability that a given verb takes a particular (type of) subject and combination of complements, including cases where the complements had been moved from their canonical positions, for example by extraction. A meta-grammar would also be likely to be able to differentiate the probabilities of particular co-occurrences of adverbial and prepositional phrase modifiers, and would moreover be able to distinguish between different orderings of the modifiers.

The approach described by Lafferty, Sleator, and Temperley (1992) of associating probabilities with Link Grammars—taken to its logical conclusion—corresponds to SLG(3), since in that approach separate probabilities are associated with each way of linking a word up with a combination of other words[7].

2.4 Globally-Dependent Frequencies

The fourth and final approach we consider is Bod's Data-Oriented Parsing (DOP) framework (Bod, 1995). In this paper we call it **SLG(4)** for uniformity and ease of reference. Bod formalizes DOP in terms of a *stochastic tree-substitution grammar*, which consists of a finite set of elementary trees, each with an associated probability such that the probabilities of all the trees with the same non-terminal symbol sum to 1, with an operation of substitution to combine the trees. In DOP, or SLG(4), the elementary trees are arbitrarily large subtrees anchored at terminal nodes by words/part-of-speech labels, and acquired automatically from pre-parsed training data. This is in contrast to SLG(3), in which the size of individual meta-productions is bounded, since the structure of the meta-productions is wholly determined by the form of the elementary trees in the grammar.

[5] In such context-free grammars, the terminal and nonterminal alphabets are not necessarily disjoint, and only the trees generated by the grammar (not their frontier strings) are of any interest.

[6] In the formulation of TAG derivations given by Schabes and Shieber (1994) an arbitrary number of modifications can take place at a single node. This means that there are an infinite number of productions in the meta-grammar, i.e., an infinite number of ways of expanding trees. This means that a pure version of SLG(3) is not possible. See Section 4.2 for ways to deal with this issue.

[7] Lafferty, Sleator, and Temperley (1992) appear to consider only cases where a word has at most one right and one left link, i.e., probabilities are associated with at most triples. However, the formalism as defined by Sleator and Temperley (1993) allows a more general case with multiple links in each direction, as would be required to deal with, for example, modifiers.

3 Empirical Evaluation

We have described four ways in which frequency information can be associated with a lexicalized grammar. Directly comparing the performance of the alternative schemes by training a wide-coverage grammar on an appropriate annotated corpus and then parsing further, unseen data using each scheme in turn would be a large undertaking outside the scope of this paper. However, each scheme varies in terms of the degree to which it can, in principle, distinguish the probability of certain kinds of derivational phenomena. This can be tested without the need to develop and run a parsing system, since each scheme can be seen as making verifiable predictions about the absence of certain dependencies in derivations of sentences in corpus data.

SLG(1), with only context-free frequency information, predicts that the relative frequency of use of the trees for a given nonterminal is not sensitive to where the trees are used in a derivation. For example, there should be no significant difference between the likelihood that a given NP tree is chosen for substitution at the subject position and the likelihood that it is chosen for the object position. SLG(2) (using so-called node-dependent frequency information) is able to cater for such differences but predicts that the likelihood of substituting or adjoining a tree at a given node in another tree is not dependent on what else is adjoined or substituted into that tree. With SLG(3) (which uses what we call locally-dependent frequency information) it is possible to encode such sensitivity, but more complex contextual dependencies cannot be expressed: for example, it is not possible for the probability associated with the substitution or adjunction of a tree γ into another tree γ' to be sensitive to where the tree γ' itself is adjoined or substituted. Only SLG(4) (in which frequency information can be globally-dependent) can do this.

In the remainder of this section we present a number of empirical phenomena that support or refute predictions made by each of the versions of SLG.

3.1 SLG(1) vs. SLG(2–4)

Magerman and Marcus (1991) report that, empirically, a noun phrase is more likely to be realized as a pronoun in subject position than elsewhere. To capture this fact it is necessary to have two different sets of probabilities associated with the different possible NP trees: one for substitution in subject position, and another for substitution in other positions. This cannot be done in SLG(1) since frequency information in SLG(1) is context-free. This phenomenon therefore violates the predictions of SLG(1), but it can be captured by the other SLG models.

Individual lexemes also exhibit these types of distributional irregularities. For example, in the Wall Street Journal (WSJ) portion of the Penn Treebank 2 (Marcus et al., 1994), around 38% of subjects of verbs used intransitively (i.e., without an object NP) in active, ungapped constructions are either pronouns or proper name phrases[8]. However, for the verbs *believe*, *agree*, and *understand*, there is a significantly higher proportion (in statistical terms) of proper name/pronoun subjects (in the case of *believe* 57%; χ^2, 40.53, 1 *df*, $p < 0.001$) [9]. This bias would, in semantic terms, be accounted for by a preference for subject types that can be coerced to *human*. SLG(2–4) can capture this distinction whereas SLG(1) cannot since it is not sensitive to where a given tree is used.

3.2 SLG(2) vs. SLG(3–4)

The Penn Treebank can also allow us to probe the differences between the predictions made by SLG(2) and SLG(3–4). From an analysis of verb phrases in active, ungapped constructions with only pronominal and/or proper name subjects and NP direct objects, it is the case that there is a (statistically) highly significant dependency between the type of the subject and the type of the object (χ^2, 29.79, 1 *df*, $p < 0.001$), the bias being towards the subject and direct object being either (a) both pronouns, or (b) both proper names. Thus the choice of which type of NP tree to fill subject position in a verbal tree can be dependent on the choice of NP type for object position. Assuming that the subject and object are substituted/adjoined into trees anchored by the verbs, this phenomenon violates the predictions of SLG(2)—hence also SLG(1)—but can still be modeled by SLG(3–4).

[8] Subjects were identified as the NP-SBJ immediately preceding a VP bracketing introduced by a verb labeled VBD/VBP/VBZ; pronouns, words labeled PRP/PRP$; and proper noun phrases, sequences of words all labeled NNP/NNPS.

[9] A value for p of 5 corresponds to statistical significance at the standard 95% confidence level; smaller values of p indicate higher confidence.

A similar sort of asymmetry occurs when considering the distribution of pronoun and proper name phrases against other NP types in subject and direct object positions. There is again a significant bias towards the subject and object either both being a pronoun/proper name phrase, or neither being of this type (χ^2, 8.77, 1 *df*, $p = 0.3$). This again violates the predictions of SLG(2), but not SLG(3–4).

Moving on now to modifiers, specifically prepositional phrase (PP) modifiers in verb phrases, the Penn Treebank distinguishes several kinds including PPs expressing manner (PP-MNR), time (PP-TMP), and purpose (PP-PRP). Where these occur in combination there is a significant ordering effect: PP-MNR modifiers tend to precede PP-TMP (χ^2, 4.12, 1 *df*, $p = 4.2$), and PP-TMP modifiers in their turn have a very strong tendency to precede PP-PRP ($p < 0.001$). Adopting Schabes and Shieber's (1994) formulation of the adjunction operation in TAG, multiple PP modifier trees would be adjoined independently at the same node in the parent VP tree, their surface order being reflected by their ordering in the derivation tree. Therefore, in SLG(3) multiple modifying PPs would appear within a single meta-production in the order in which they occurred, and the particular ordering would be assigned an appropriate probability by virtue of this. In contrast, SLG(2) treats multiple adjunctions separately and so would not be able to model the ordering preference.

Significant effects involving multiple modification of particular lexical items are also evident in the treebank. For example, the verb *rise* occurs 83 times with a single PP-TMP modifier—e.g. (1a)—and 12 times with two (1b), accounting in total for 6% of all PPs annotated in this way as temporal.

(1) a *Payouts on the S&P 500 stocks rose 10 % [PP-TMP in 1988] , according to Standard & Poor 's Corp. ...*
b *It rose largely [PP-TMP throughout the session] [PP-TMP after posting an intraday low of 2141.7 in the first 40 minutes of trading] .*

The proportion of instances of two PP-TMP modifiers with *rise* is significantly more than would be expected given the total number of instances occurring in the treebank (χ^2, 25.99, 1 *df*, $p < 0.001$). The verb *jump* follows the same pattern ($p = 1.0$), but other synonyms and antonyms of *rise* (e.g. *fall*) do not. This idiosyncratic behavior of *rise* and *jump* cannot be captured by SLG(2), since each adjunction is effectively considered to be a separate independent event. In SLG(3), though, the two-adjunction case would appear in a single meta-production associated with *rise/jump* and be accorded a higher probability than similar meta-productions associated with other lexical items.

There is another, more direct but somewhat less extensive, source of evidence that we can use to investigate the differences between SLG(2) and (3–4). B. Srinivas at the University of Pennsylvania has recently created a substantial parsed corpus[10] by analyzing text from the Penn Treebank using the XTAG system (Group, 1995). Some of the text has been manually disambiguated, although we focus here on the most substantial set—of some 9900 sentences from the WSJ portion—which has not been disambiguated, as yet. For each sentence we extracted the set of meta-level productions that would generate the XTAG derivation. To obtain reliable data from ambiguous sentences, we retained only the (approximately 37500) productions that were common across all derivations. In this set of productions we have found that with the elementary tree licensing subject–transitive-verb–object constructions, the likelihood that the object NP is expanded with a tree anchored in *shares* is much higher if the subject is expanded with with a tree anchored in *volume*, corresponding to sentences such as (2a) and (2b).

(2) a *Volume totaled 14,890,000 shares .*
b *Overall Nasdaq volume was 151,197,400 shares .*

Indeed, in all 11 cases where *volume* is the anchor of the subject, an NP anchored in *shares* is analyzed as the object, whereas more generally *shares* is object in only 18 of the 1079 applications of the tree. This difference in proportions is statistically highly significant ($p < 0.001$). Correlation between each of *volume* and *shares* and the verbs that appear is much weaker. There is of course potential for bias in the frequencies since this data is based purely on unambiguous productions. We therefore computed the same proportions from productions derived from all sentences in the XTAG WSJ data; this also resulted in a highly significant difference. SLG(2) models the substitution of the subject and of the object as two independent events, whereas the data show that they can exhibit a strong inter-dependency.

[10] We wish to thank B. Srinivas for giving us access to this resource.

3.3 SLG(3) vs. SLG(4)

Bod (1995) observes that there can be significant inter-dependencies between two or more linguistic units, for example words or phrases, that cut across the standard structural organization of a grammar. For example, in the Air Travel Information System (ATIS) corpus (Hemphill, Godfrey, and Doddington, 1990) the generic noun phrase (NP) *flights from X to Y* (as in sentences like *Show me flights from Dallas to Atlanta*) occurs very frequently. In this domain the dependencies between the words in the NP—but without X and Y filled in—are so strong that in ambiguity resolution it should arguably form a single statistical unit. Bod argues that Resnik and Schabes' schemes (i.e. SLG(2)) cannot model this; however it appears that SLG(3) can since the NP would give rise to a single meta-production (under the reasonable assumption that the *from* and *to* PPs would be adjoined into the NP tree anchored by *flights*).

An example given by Bod that does demonstrate the difference between SLG(3) and SLG(4) concerns sentences like *the emaciated man starved*. Bod argues that there is a strong (semantic) dependence between *emaciated* and *starved*, which would be captured in DOP—or SLG(4)—in the form of a single elementary tree in which *emaciated* and *starved* were the only lexical items. This dependence cannot be captured by SLG(3) since *emaciated* and *starved* would anchor separate elementary trees, and the associations made would merely be between (1) the S tree anchored by *starved* and the substitution of the NP anchored by *man* in subject position, and (2) the modification of *man* by *emaciated*.

3.4 Discussion

The empirical phenomena discussed above mainly concern interdependencies within specific constructions between the types or heads of either complements or modifiers. The phenomena fall clearly into two groups:

- ones relating to distributional biases that are independent of particular lexical items, and
- others that are associated with specific open class vocabulary.

Token frequencies—with respect to treebank data—of phenomena in the former group are relatively high, partly because they are not keyed off the presence of a particular lexical item: for example in the case study into the complement distributions of pronoun/proper name phrases versus other NP types (section 3.2) there are 13800 data items (averaging one for every four treebank sentences). However, there appears to be a tendency for the phenomena in this group to exhibit smaller statistical biases than are evident in the latter, lexically-dependent group (although all biases reported here are significant at least to the 95% confidence level). In the latter group, although token frequencies for each lexical item are not large (for example, the forms of *rise* under consideration make up only 1% of comparable verbs in the treebank), the biases are in general very strong, in what are otherwise an unremarkable set of verbs and nouns (*believe*, *agree*, *understand*, *rise*, *jump*, *volume*, and *shares*). We might therefore infer that although individually token frequencies are not great, *type* frequencies are (i.e. there are a large number of lexical items that display idiosyncratic behavior of some form or other), and so lexicalized interdependencies are as widespread as non-lexical ones.

4 Parameter Estimation

4.1 Training Regime

Schabes (1992) describes an iterative re-estimation procedure (based on the Inside-Ouside Algorithm (Baker, 1979)) for refining the parameters of an SLG(2) grammar given a corpus of in-coverage sentences; the algorithm is also able to simultaneously acquire the grammar itself. The aim of the algorithm is to distribute the probability mass within the grammar in such as way that the *probability of the training corpus* is maximized, i.e. model as closely as possible the language in that corpus. However, when the goal is to return as accurately as possible the *correct analysis* for each sentence using a pre-existing grammar, estimating grammar probabilities directly from normalized frequency counts derived from a pre-parsed training corpus can result in accuracy that is comparable or better to that obtained using re-estimation (Charniak, 1996). Direct estimation would mesh well with the SLG formalisms described in this paper.

4.2 Smoothing

The huge number of parameters required for a wide-coverage SLG(2) (and even more so for SLG(3–4)) means that not only would the amount of frequency information be unmanageable, but data sparseness would make useful probabilities hard to obtain. We briefly present three (essentially orthogonal and independent) backing-off techniques that could be used to address this problem.

Unanchored Trees

It is the size of a wide-coverage lexicon that makes pure SLG(2–4) unmanageable. However, without lexical anchors a wide-coverage SLG would have only a few hundred trees (Group, 1995). Backup frequency values could therefore be associated with unanchored trees and used when data for the anchored case was absent.

Lexical Rules

In a lexicalized grammar, elementary trees may be grouped into families which are related by lexical rules—such as *wh* extraction, and passivization. (For example, the XTAG grammar contains of the order of 500 rules grouped into around 20 families). In the absence of specific frequency values, approximate (backup) values could be obtained from a tree that was related by some lexical rule.

SLG(i) to SLG($i-1$)

Section 3 indicated informally how, when moving from SLG(1) through to SLG(4), the statistical model becomes successively more fine-grained, with each SLG(i) model subsuming the previous ones, in the sense that SLG(i) is able to differentiate probabilistically all structures that previous ones can. Thus, when there is insufficient training data, sub-parts of a finer-grained SLG model could be backed off to a model that is less detailed. For example, within a SLG(3) model, in cases where a particular set of meta-productions all with the same mother had a low collective probability, the set could be reduced to a single meta-production with unspecified daughters (i.e. giving the effect of SLG(1)).

5 Comparison with Other Work

The treatment of stochastic lexicalized grammar in this paper has much in common with recent approaches to statistical language modeling outside the TAG tradition. Firstly, SLG integrates statistical preferences acquired from training data with an underlying wide-coverage grammar, following an established line of research, for example (Chitrao and Grishman, 1990; Charniak and Carroll, 1994; Briscoe and Carroll, 1995). The paper discusses techniques for making preferences sensitive to context to avoid known shortcomings of the context-independent probabilities of SCFG (see e.g. Briscoe and Carroll (1993)).

Secondly, SLG is *lexical*, since elementary trees specify lexical anchors. Considering the anchor of each elementary tree as the head of the construction analyzed, successive daughters for example of a single SLG(3) meta-grammar production can in many cases correspond to a combination of Magerman's (1995) mother/daughter and daughter/daughter head statistics (although it would appear that Collins' (1996) head-modifier configuration statistics are equivalent only to SLG(2) in power). However, due to its extended domain of locality, SLG(3) is not limited to modeling local dependencies such as these, and it can express dependencies between heads separated by other, intervening material. For example, it can deal directly and naturally with dependencies between subject and any verbal complement without requiring mediation via the verb itself: c.f. the example of section 3.2.

Thirdly, the SLG family has the ability to model explicitly syntactic *structural* phenomena, in the sense that the atomic structures to which statistical measures are attached can span multiple levels of derived parse tree structure, thus relating constituents that are widely-separated—structurally as well as sequentially—in a sentence. Bod's DOP model (Bod, 1995) shares this characteristic, and indeed (as discussed in section 2.4) it fits naturally into this family, as what we have called SLG(4).

Srinivas et al. (1996) (see also Joshi and Srinivas (1994)) have recently described a novel approach to parsing with LTAG, in which each word in a sentence is first assigned the most probable elementary tree—or 'supertag'—given the context in which the word appears, according to an trigram model of supertags. The rest of the parsing

process then reduces to finding a way of combining the supertags to form a complete analysis. In this approach statistical information is associated simply with linear sub-sequences of elementary trees, rather than with trees within derivational contexts as in SLG(2–4). Although Srinivas' approach is in principle efficient, mistaggings mean that it is not guaranteed to return an analysis for every in-coverage sentence, in contrast to SLG. Also, its relatively impoverished probabilistic model would not be able to capture many of the phenomena reported in section 3.

Acknowledgement

This work was supported by UK EPSRC project GR/K97400 'Analysis of Naturally-occurring English Text with Stochastic Lexicalized Grammars' (<http://www.cogs.susx.ac.uk/lab/nlp/dtg/details.html>), and by an EPSRC Advanced Fellowship to the first author. We would like to thank Nicolas Nicolov and Miles Osborne for useful comments on previous drafts.

References

Alshawi, Hiyan. 1996. Head automata and bilingual tilings: Translation with minimal representations. In *34th Meeting of the Association for Computational Linguistics (ACL'96)*, pages 167–176.

Baker, J. 1979. Trainable grammars for speech recognition. In D. Klatt and J. Wolf, editors, *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*. MIT, Cambridge, MA, pages 547–550.

Bod, R. 1995. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. Ph.D. thesis, University of Amsterdam, ILLC dissertation series 95-14.

Booth, T. and R. Thompson. 1973. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5):442–450.

Boullier, P. 1996. Another facet of LIG parsing. In *34th Meeting of the Association for Computational Linguistics (ACL'96)*, pages 87–94.

Briscoe, E. and J. Carroll. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. *Computational Linguistics*, 19(1):25–60.

Briscoe, E. and J. Carroll. 1995. Developing and evaluating a probabilistic LR parser of part-of-speech and punctuation labels. In *4th International Workshop on Parsing Technologies (IWPT'4)*, pages 48–58.

Charniak, E. 1996. Tree-bank grammars. Technical Report CS-96-02, Brown University, Department of Computer Science.

Charniak, E. and G. Carroll. 1994. Context-sensitive statistics for improved grammatical language models. In *12th National Conference on Artificial Intelligence (AAAI'94)*, pages 728–733.

Chitrao, M. and R. Grishman. 1990. Statistical parsing of messages. In *DARPA Speech and Natural Language Workshop*, pages 263–266.

Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *34th Meeting of the Association for Computational Linguistics (ACL'96)*, pages 184–191.

Group, The XTAG Research. 1995. A lexicalized tree adjoining grammar for English. Technical Report IRCS 95-03, The Institute for Research in Cognitive Science, University of Pennsylvania.

Hemphill, C., J. Godfrey, and G. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*.

Joshi, A. and B. Srinivas. 1994. Disambiguation of super parts of speech (or supertags): almost parsing. In *15th International Conference on Computational Linguistics (COLING'94)*.

Lafferty, J., D. Sleator, and D. Temperley. 1992. Grammatical trigrams: a probabilistic model of link grammar. In *AAAI Fall Symposium on Probabilistic Approaches to Natural Language*.

Magerman, D. 1995. Statistical decision-tree models for parsing. In *33rd Meeting of the Association for Computational Linguistics (ACL'95)*, pages 276–283.

Magerman, D. and M. Marcus. 1991. Pearl: a probabilistic chart parser. In *2nd International Workshop on Parsing Technologies (IWPT'2)*, pages 193–199.

Marcus, M., G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*.

Rambow, O., K. Vijay-Shanker, and D. Weir. 1995. D-Tree Grammars. In *33rd Meeting of the Association for Computational Linguistics (ACL'95)*, pages 151–158.

Resnik, P. 1992. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *14th International Conference on Computational Linguistics (COLING'92)*, pages 418–424.

Schabes, Y. 1990. *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Schabes, Y. 1992. Stochastic lexicalized tree-adjoining grammars. In *14th International Conference on Computational Linguistics (COLING'92)*, pages 426–432.

Schabes, Y. and S. Shieber. 1994. An alternative conception of tree-adjoining derivation. *Computational Linguistics*, 20(1):91–124.

Sleator, D. and D. Temperley. 1993. Parsing English with a link grammar. In *3rd International Workshop on Parsing Technologies (IWPT'3)*, pages 277–292.

Srinivas, B., C. Doran, B. Hockey, and A. Joshi. 1996. An approach to robust partial parsing and evaluation metrics. In J. Carroll, editor, *Proceedings of the Workshop on Robust Parsing*. 8th European Summer School in Logic, Language and Information, pages 70–82.

Vijay-Shanker, K. 1987. *A Study of Tree Adjoining Grammars*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA.

Vijay-Shanker, K. and D. Weir. 1993. Parsing some constrained grammar formalisms. *Computational Linguistics*, 19(4):591–636.

Vijay-Shanker, K. and D. Weir. 1994. The equivalence of four extensions of context-free grammars. *Math. Syst. Theory*, 27:511–546.

Vijay-Shanker, K., D. Weir, and A. Joshi. 1987. Characterizing structural descriptions produced by various grammatical formalisms. In *25th Meeting of the Association for Computational Linguistics (ACL'87)*, pages 104–111.