

Corpus Induction of Lexicons for Treebank PCFGs by Inside-Outside Estimation and Frequency Transformations

Tejaswini Deoskar
Department of Linguistics
Cornell University
td72@cornell.edu

Mats Rooth
Department of Linguistics
Cornell University
rooth@cornell.edu

Abstract

We describe procedures which pool lexical information from a treebank with frequency information estimated from an unannotated corpus with the inside-outside algorithm. PCFG parameters for non-lexical productions are obtained purely from the treebank. The procedures produce substantial improvements (upto 20.34%) on the task of determining valences of tokens of novel verbs, relative to a smoothed treebank model.

1 Introduction

Earlier research on inside-outside estimation of PCFG or lexicalized PCFG probability models for natural language grammars has reported positive results, including positive results in estimating features related to verbal valence (Carroll and Rooth 1998; Beil et. al. 1999; Schulte 2002). But a wealth of questions remain open, including questions of the relative efficacy of supervised estimation using a treebank and unsupervised inside-outside estimation on a much larger corpus, the possibility of a disjuncture between the numerical optimization of corpus probability performed by inside-outside with the intended interpretation of tree-structural markup, and finally, the issue of the efficacy of inside-outside estimation as measured by standardized evaluation criteria. To address these questions, it is in some cases necessary and in all cases helpful to work with grammars aligned with standard treebank databases,

rather than hand-built grammars which are not. For instance, since Carroll and Rooth (1998) used a hand-built grammar of English (and not a treebank grammar), a maximal-probability tree according to the model they induced has a non-standard shape, and it is not possible to evaluate parsing performance by standardized criteria.

The research reported here addresses these issues, working with a treebank PCFG. Like the authors cited earlier, the specific task that we focus on is the detection of verbal valence. Valence¹ is the phenomenon of an individual verb or other head selecting complements of particular types, e.g. nominal, prepositional, or clausal complements or combinations of them. The valence preference of most verbs is represented inadequately (or not at all) in PCFGs trained solely by supervised techniques, due to their low or zero occurrence frequency in even large annotated corpora like the Penn Treebank. The learning of verb valences from unannotated data is important for creating deep lexical resources useful for tasks like parsing (see, e.g. Manning 1993; Briscoe and Carroll 1997; Korhonen 2002) and in addition, also a good test candidate for evaluation of the procedures we describe, which combine lexical parameters obtained by unsupervised inside-outside estimation with grammatical parameters from a treebank. Other semi-supervised methods like self-training which combine supervised and unsupervised training of treebank PCFGs have obtained improvement in parsing results (McClosky et.al. 2006).

We work with a grammar obtained from the Penn Treebank (Marcus et.al., 1993), where sym-

¹Also called subcategorization.

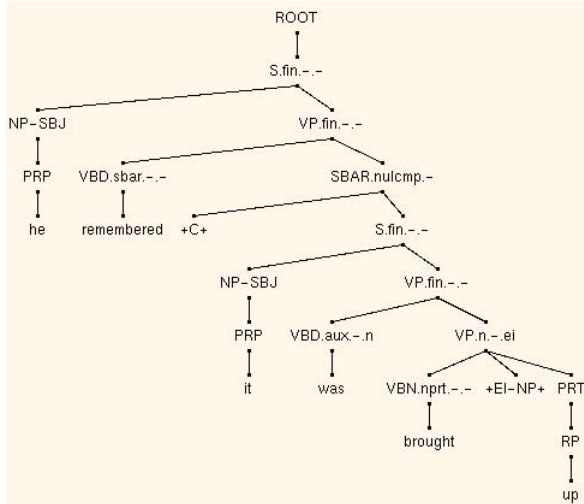


Figure 1: A Treebank-style tree with incorporated features, including a valence feature on verb tags.

bolds and empty categories have been transformed by a method described elsewhere (self-reference omitted). Figure 1 illustrates a tree licensed by our grammar. Preterminal symbols have an incorporated *valence* feature. Labels such as *VP.fin.-.-* are labels in Penn Treebank II style, up to the first period. The subsequent substring (for e.g. *fin.-.-*) represents three incorporated feature values. The first one is a GPSG-style *Vform* feature indicating valence, with an extended range of values. Figure 2 lists some of the most common valence types, with their frequencies in sections 0-15 of the Penn Treebank. A feature *n* indicates a transitive valence (with an NP object and no other complements), *zero* indicates an intransitive valence, *nn* a ditransitive valence, *sbar* a *that*-clause or other clausal complement with a complementizer, *p* a prepositional complement, and *np* a combination of an NP object and a prepositional complement. *Aux* indicates an auxiliary verb, with a VP complement, while *prd* indicates a predicative complement.

The last field in the incorporation sequence subclassifies valence, with *-* being a default value. For example, the value *g* or *to* indicates a *ing*-participle (present participle) or an infinitival in a VP complement, and *sc* indicates a small clause complement (as in [*consider [her nice]*]). The middle field subclassifies the subject of a clausal complement, with *ei* indicating an indexed PRO or trace of rais-

24945	n.-.-	3783	np.-.-	7724	zero.-.-
3506	p.-.-	7208	sbar.-.-	3168	s.ei.to
7129	prd.-.-	1420	s.-.-	4758	aux.-.h
1352	s.-.to	4372	aux.-.n	889	s.-.sc

Figure 2: Frequencies of verbal incorporations (Penn Treebank sections 0-15).

ing. Thus the incorporation sequence for the example [*try [to stay awake]*] is *s.ei.to* where *s* describes the clausal complement of *try*, *ei* indicates the null subject of *to stay awake*, and *to* indicates that the complement is infinitival.

Tree shape and labels without incorporations are in Penn Treebank II style.²

2 Inside-outside estimation

As a basic unsupervised estimation method, we use standard inside-outside estimation of PCFGs, which realizes EM estimation (Pereira and Schabes 1992; Prescher 2003). Let $I(C, f)$ designate the new frequency model computed from the corpus C using a probability model based on the frequency model f . Iterative inside-outside estimation has the following simple form, where each successive frequency model e_{i+1} is estimated from C using a probability model determined by the previous frequency model e_i .

$$\begin{aligned}
 e_1 &= I(C, e_0) \\
 e_2 &= I(C, e_1) \\
 &\dots \\
 e_{i+1} &= I(C, e_i)
 \end{aligned}
 \tag{1}$$

Our notation always refers to frequency models such as e_i , rather than the relative-frequency probability models they determine. We use t_0 to designate the frequency model obtained from counting local tree configurations from the Penn Treebank augmented with incorporations on symbols. As described in §5, this treebank model is based on treebank sections 0-15, with about 1000 sentences held out for testing. Figures 3 and 4 illustrate syntactic and lexical frequencies in this model.³ Where r is

²The exception are empty categories which are flanked by plus signs: +C+ is an empty complementizer, +EI-NP+ is an indexed NP A-trace, for instance the trace of passive, etc.

³Productions can have non-integral frequencies, because our

21643.84	ROOT S.fin.-.-
14133.0	PP.of IN.of NP
13835.0	NP DT NN
12690.24	S.fin.-.- NP-SBJ VP.fin.-.-
11484.16	S.fin.-.- NP-SBJ VP.fin.-.- -PER-
9685.0	NP-SBJ PRP

Figure 3: Syntactic rule frequencies in the treebank frequency model t_0

a syntactic rule such as $PP.of \rightarrow IN.of NP$, $f(r)$ is its frequency in the model f . For lexical frequencies we use the notation $f(w, \tau, \iota)$, where w is a word form, τ is a part of speech tag, and ι is a feature incorporation. Thus $t_0(PP.of \rightarrow IN.of NP)=14133$, and $t_0(\text{attaches}, \text{VBZ}, \text{np}, \text{-}) = 1.0$.

§4 will define a smoothed version t of the raw treebank model t_0 . The notation in the next section will already refer to t .

3 Frequency transformations

Our guiding hypothesis is that lexical parameters $t(w, \tau, \iota)$ determined from the treebank are poorly estimated because of the sparseness of treebank data for a particular word, while syntactic parameters $t(r)$ are comparatively well estimated, because they are not keyed to particular words. If this is so, it might be beneficial to use supervised estimation for syntactic parameters, while using a combination of supervised and unsupervised estimation for lexical parameters.

To this end, we define a frequency transformation $T(c, t)$ which pools information from a frequency model c obtained from a corpus C (separate from the treebank corpus) with information from the treebank model t . Our definitions re-allocate frequencies in a way which preserves marginal frequencies seen in the treebank model. A marginal tag-incorporation frequency is defined by summation:

$$f(\tau, \iota) = \sum_w f(w, \tau, \iota). \quad (2)$$

T will be defined to preserve these marginal frequencies.

algorithm for adding features to a treebank tree can produce multiple solutions. In such cases a unit frequency is split among the solutions, and counts for local trees are scaled by the split frequency.

attaches	VBZ.np.-.- 1.0
attaching	VBG.n.-.- 1.0
attack	NN 22.0 VBP.-.-.- 1.0
	VB.n.-.- 3.0 VB.zero.-.- 1.0
attacked	VBD.zero.-.- 1.0 VBN.np.-.- 1.0
	VBN.n.-.- 5.0

Figure 4: Lexical frequencies in the treebank frequency model t_0 , illustrating sparsity of the lexical distribution. Each line has a word, followed by pairs of preterminal symbols and frequencies. Valences exhibited for one inflectional form are absent for others.

Let c and t be two frequency models, called the ‘‘corpus’’ and ‘‘treebank’’ models. We define a new frequency model d from c and t in two steps. First, corpus frequencies are scaled by the ratio of treebank and corpus marginal frequencies (3). The resulting quantity preserves marginal frequencies (4).

$$\bar{c}(w, \tau, \iota) = \frac{t(\tau, \iota)}{c(\tau, \iota)} c(w, \tau, \iota). \quad (3)$$

$$\begin{aligned} \bar{c}(\tau, \iota) &= \sum_w \bar{c}(w, \tau, \iota) \\ &= \sum_w \frac{t(\tau, \iota)}{c(\tau, \iota)} c(w, \tau, \iota) \\ &= \frac{t(\tau, \iota)}{c(\tau, \iota)} \sum_w c(w, \tau, \iota) \\ &= \frac{t(\tau, \iota)}{c(\tau, \iota)} c(\tau, \iota) \\ &= t(\tau, \iota). \end{aligned} \quad (4)$$

Second, for a given tag and incorporation, d is an interpolation of t and \bar{c} for lexical parameters.

$$d(w, \tau, \iota) = (1 - \lambda_{\tau, \iota}) t(w, \tau, \iota) + \lambda_{\tau, \iota} \bar{c}(w, \tau, \iota), \quad (5)$$

where $\lambda_{\tau, \iota}$ is a parameter with $0 < \lambda_{\tau, \iota} < 1$ which may depend on the tag and incorporation. T is the transformation of t and c into d which is defined in (5) for lexical entries w, τ, ι , and by setting $T(c, t)$ to equal t for syntactic parameters. The defined quantity again preserves marginals:

$$\begin{aligned} d(\tau, \iota) &= \sum_w d(w, \tau, \iota) \\ &= \sum_w (1 - \lambda_{\tau, \iota}) t(w, \tau, \iota) + \lambda_{\tau, \iota} \bar{c}(w, \tau, \iota) \\ &= (1 - \lambda_{\tau, \iota}) \sum_w t(w, \tau, \iota) \\ &\quad + \lambda_{\tau, \iota} \sum_w \bar{c}(w, \tau, \iota) \\ &= (1 - \lambda_{\tau, \iota}) t(\tau, \iota) + \lambda_{\tau, \iota} \bar{c}(\tau, \iota) \\ &= (1 - \lambda_{\tau, \iota}) t(\tau, \iota) + \lambda_{\tau, \iota} t(\tau, \iota) \\ &= t(\tau, \iota). \end{aligned} \quad (6)$$

We use T in conjunction with inside-outside estimation in two ways, leading to two procedures.

In one procedure, straight inside-outside estimation as in (1) is run on an unsupervised training corpus C to obtain a sequence of frequency models $e_1, e_2, \dots, e_i, \dots$. Then, one of the inside-outside models is merged with the treebank model as $T(e_i, t)$. The resulting model uses t for syntactic parameters, and merges information from t and e_i for lexical parameters without affecting the iterative inside-outside procedure.

In an alternative procedure, T is interleaved into iterative inside-outside estimation as follows.

$$\begin{aligned}
 d_0 &= t \\
 c_1 &= I(C, d_0) \\
 d_1 &= T(c_1, t) \\
 c_2 &= I(C, d_1) \\
 d_2 &= T(c_2, t) \\
 &\dots \\
 c_{i+1} &= I(C, d_i) \\
 d_{i+1} &= T(c_{i+1}, t) \\
 &\dots
 \end{aligned} \tag{7}$$

In the iteration with index $i+1$, a frequency model $c_{i+1} = I(C, d_i)$ is obtained using inside-outside from the corpus C and a merged model d_i from the previous iteration. The next merged model d_{i+1} is obtained as $T(c_{i+1}, t)$. The models d_i have syntactic frequencies from the treebank, and lexical frequencies pooled from the treebank and the corpus, but in a way which is interleaved into iterative estimation.

T can be motivated as preserving a certain property of the original treebank lexicon, namely the marginal frequencies. This would be significant if the preterminals in the lexicon were also used as left hand sides of syntactic rules, something which does not happen in our grammar. In terms of the probability model determined by the frequency model, the effect of T is to allocate a fixed proportion of the probability mass for each τ, ι to the corpus, and share it out among words w in proportion to relative frequencies

$$\frac{c_i(w, \tau, \iota)}{c_i(\tau, \iota)}$$

in the inside-outside estimate c_i . Stating this in terms of frequencies retains the scaling of frequencies found in the treebank.

4 Smoothing the treebank model

To initialize the iterative procedures, a smoothing scheme is required which allocates frequency to all possible incorporations, and also allocates frequency to combinations of words w and part of speech tags τ which are not present in the treebank. Otherwise, if t_0 has zero frequency for some lexical parameter, the inside-outside estimate $I(C, t_0)$ for that parameter would also be zero, and new lexical entries would never be induced.

The smoothed treebank model t is obtained as follows. First a part of speech tagger is run on the unsupervised corpus C , and tokens of words and part of speech tags are tabulated to obtain a frequency table $g(w, \tau)$. Each frequency $g(w, \tau)$ is split among possible incorporations ι in proportion to a ratio of marginal frequencies in t :

$$g(w, \tau, \iota) = \frac{t(\tau, \iota)}{t(\tau)} g(w, \tau) \tag{8}$$

Then t is defined as an interpolation of g and t for lexical parameters as below, with syntactic parameters copied from t .

$$t(w, \tau, \iota) = (1 - \lambda_{\tau, \iota})t(w, \tau, i) + \lambda_{\tau, \iota}g(w, \tau, i) \tag{9}$$

5 Hundred verb experiment

To test the iterative procedures on a moderate scale, we selected 117 test words whose frequency in Penn Treebank Wall Street Journal sections 0-15 as verbs was 8-10.⁴ Although no criteria other than frequency range were applied in selecting the test words, they have suitably varied valence patterns.

⁴The words are: *Having accomplish acquires admit aims aired arguing asserts auctioned betting boasts buoyed combining completing concede converting cooperate coordinate coupled decides declaring defeated defended demanded demanding denying deserve disagree discontinued eating evaluate examine exceeding exceeds execute explore export fails fare favors firmed gathering grows guarantee hearing hinted hiring hitting honor imposing inched indicted influenced integrated interpreted justify labeled lending leveraged locked loved lowering marketed matching meaning mention miss missed monitored name observed obtaining occurs opens owes pegged plead proposes pulling rally receives recommend recommending relied resign reviewing ride rushed satisfy scuttle searching signaled slash slipping spoke spur spurred stick supports switched taped tendered tends threw track transform treated trim troubled understands upset voting waive wear wins withdrawn yielding.*

All sentences containing these word forms were removed from the treebank training sample and reserved for testing, making the test words novel. For each test word, we extracted 100 sentences containing the word from a separate corpus of New York Times stories. This resulted in an unsupervised training corpus C with about 235,200 word tokens.

We initialized the frequency model by computing t from t_0 and C as described above, using a tagger. The interpolation factor λ was $\frac{1}{1000}$ for all τ, ι , so that most of the frequency in the initial model comes from treebank. Because sentences containing test words w were removed from the treebank, the test words have frequency zero in t_0 . In consequence, for any τ, ι , the test words have a scaled version of the average distribution for that tag-incorporation pair in the treebank. Thus t , which is the initial model used in the iterative procedures, has no information specific to the test words coming from the treebank.

For inside-outside estimation, we used Bitpar (Schmid 2004), which inputs and outputs frequency grammars and lexicons in a simple text format. The frequency transformations were implemented in Perl. The resulting frequency model c_i was transformed with T , using interpolation parameters $\lambda_{\tau, \iota}$ of 0.5 for all τ, ι . This results in the frequency model d_i used in the next iteration. In our setup of five 3.2GHz Pentiums, the run time for one iteration is about 1 1/2 days.

We ran two versions of iterative estimation. The straight inside-outside procedure (1) starting an initial model $e_0 = t$ results in a sequence of models $e_1 \dots e_5$. Each of these was merged with t using T to obtain a sequence of models $T(e_1, t) \dots T(e_5, t)$. Second, the interleaved procedure (7) results in a sequence of models $d_1 \dots d_5$.

In the straight inside-outside procedure, we found that marginal frequencies $e_i(\tau, \iota)$ occasionally dropped to zero, causing division by zero in the definition of a probability model by relative frequency. In this case, the combination τ, ι was eliminated from the lexicon.

6 Valence evaluation

We tested the frequency lexicons produced by the iterative procedures in a model parsing task, the de-

Iteration i	e_i	e_i, t	$T(e_i, t)$	d_i
0	48.810	48.810	48.810	48.810
1	39.921	36.743	36.210	37.698
2	38.928	33.466	33.036	30.159
3	37.041	31.877	30.357	29.167
4	37.239	30.686	30.258	28.869
5	37.835	30.983	30.555	28.472

Figure 5: Valence error percentages for four conditions

tection of valences for token occurrences of the test words in the held-out Treebank sample (1008 tokens). For a given test item, the treebank annotation, including the preterminal tag, is stripped away, and a Viterbi (i.e. maximal probability) parse is computed for the terminal string using Bitpar and one of the estimated frequency models. The preterminal symbol for the token test word, which consists of a part of speech and an incorporation (encoding the valence if the tag is verbal) is extracted from the maximal probability tree.

The gold standard is obtained from the incorporation markup in the transformed Treebank tree. However, our transformation procedure occasionally fails to disambiguate some features. In the 84 of the test items where this was true of the valence feature, we disambiguated the valence by examining the original treebank tree. We also spot checked other test items, and found only isolated problems in the valences, usually due to what in our opinion were errors in Treebank markup. Since the errors were infrequent, and since we wanted to perform an evaluation using the Treebank database, we did not undertake to correct them.

We scored the tag-incorporation pair τ, ι found in the maximal probability parse as correct if it matched the pair obtained from the treebank in both components. Thus a preterminal VBD.n.-.- indicating a transitive past tense verb is called incorrect if the correct markup is VBD.nn.-.- with a ditransitive valence, but also if the correct markup is VBN.n.-.-, with a past participle part of speech VBN. In short, part of speech errors are scored as incorrect, even if the incorporation is correct.

Results are stated as an error rate, the fraction of test items which receive incorrect tag-incorporation

pairs in the maximal probability parses. In Figure 5, the column labeled e_i gives error rates for the sequence of models obtained with the straight inside-outside procedure. These models have re-estimated frequencies for both lexical and syntactic parameters. The column labeled e_i, t uses e_i for lexical parameters, and t for syntactic parameters. The switch to the treebank model for syntactic parameters produces a substantial drop in error rate, a drop of more than five points in the third iteration. In the column $T(e_i, t)$, each model e_i was merged with t after inside-outside, using the transformation T . This pools lexical information from the treebank with the iterative inside-outside estimate, without affecting the sequence of inside-outside estimates e_i . Finally, the column labeled d_i gives results for the interleaved procedure.

The baseline error rate is nearly 49%. Recall that the baseline model has for each test verb a valence frequency distribution characteristic proportional to the overall valence distribution in Treebank for verbs of the same tag, i.e. no information specific to the test verb. So with this PCFG probability model on trees, the surrounding sentence context has enough information to correctly assign the verb valence in the Viterbi tree only about half the time. In all conditions, there is a substantial drop in error rate from the baseline, with the pure inside-outside models e_i looking substantially worse than the others. Lexically smoothing e_i with transformation T is in each case better than just substituting the treebank frequencies for syntactic parameters. In iterations 2 to 5, the interleaved models d_i are better than the models $T(e_i, t)$ where the treebank model is merged only after the last step of inside-outside. In iteration 5, d_i is 2.08 error points better than $T(e_i, t)$ and still improving, while the other conditions get worse in the final iteration.

7 Other evaluations and observations

In order to measure how good a model the probability distributions determined by a frequency model d_i are of the word-tag-incorporation tuples in the test sample, Carroll and Rooth (1998) evaluated valence distributions obtained by inside-outside using *cross entropy*. They did this by looking at the probability of valence given word, and computing a cross

Iteration	$T(e_i, t)$	d_i
0	16.39	16.39
1	6.73	6.73
2	6.52	6.46
3	6.46	6.44
4	6.49	6.46

Figure 6: Mean of $-\log_2 p(w; \tau, \iota, f)$ for test items w, τ, ι , computed for frequency models f in the model sequences $T(e_i, t)$ and d_i .

acquires	197.42	VBZ.n.-.-
	21.44	VBZ.prd.-.-
	10.70	VBZ.-.-.-
	4.43	VBZ.s.-.-
favors	46.41	VBZ.n.-.-
	24.09	VBZ.aux.-.g
	12.65	VBZ.prd.-.-
	10.44	VBZ.np.-.-
	3.94	VBZ.s.ei.g
owes	35.22	VBZ.s.-.-
	18.48	VBZ.s.-.sc
	16.98	VBZ.prd.-.-
	13.63	VBZ.np.-.-

Figure 7: Three grossly incorrect entries in d_3

entropy $H(p, q)$, where p is the distribution exhibited in a test sample, and q is a distribution based on a frequency model obtained with inside-outside. While this provides a concrete numerical measure, it is somewhat ad-hoc in the PCFG context because the probability of valence given word has no direct role in the PCFG probability model. We will instead base a measure on probability of word given tag and incorporation, which is a parameter of the probability model.

Let $p(w; \tau, \iota, f)$ be the probability of w as a realization of the tag-incorporation pair τ, ι in the probability model determined from the frequency model f . For a test item with word w and correct markup τ, ι , the negative log probability $-\log_2 p(w; \tau, \iota, f)$ is relatively low if w is relatively probable as a realization of τ, ι in the probability model determined by f , and in this sense measures how good a model f is of the observed datum. We computed the mean of this figure for all the test items. This mean has

a simple interpretation in PCFG probability model for trees, namely the mean contribution that a token test word makes to the negative log probability of the correct tree in which it occurs.⁵ Figure 6 gives this figure for the model sequences $T(e_i, t)$ and d_i . It was not possible to compute this quantity for the straight inside-outside models e_i , because there were test items with zero probability. Note that both $T(e_i, t)$ and d_i have the advantage of smoothness from contribution from the smoothed treebank model t .

The big drop from model 0 to model 1 in both conditions of Figure 6 does not have a simple interpretation, because it conflates the first step in learning good distributions with concentrating probability mass on the test verbs, because of their high frequency in C . There is an increase in iteration 4 for both conditions, suggestive of overfitting or convergence to incorrect solutions for some test words.

We looked at frequency distributions for individual verbs to try to catch obvious errors. These were not common, but Figure 7 gives three of them. The verb *acquires* has a high frequency for the incorrect predicative valence *prd*(*prd* was appropriate for *is* or *becomes*). The verb *favors* has the auxiliary valence which is appropriate for a progressive use of *is*, as in *is walking*. However, the treebank analysis of *favors walking* is with an S complement and controlled empty subject corresponding to the valence *s.ei.g* which has relatively low frequency in this lexicon. The verb *owes* has a relatively low frequency for the ditransitive valence, and instead there are incorrect valences specifying an S complement. This lexicon tends to misanalyze a ditransitive such as *give* as in [*give him a gerbil*] with a small clause, comparable to [*consider* [_s *him a gerbil*]].

Figure 8 tracks the evolution of the frequency models d_1 , d_2 , and d_3 for three verbs. Note that most high-frequency valences get higher frequencies in successive models, though the transitive *n.-* valence for *combining* is falling, perhaps because of competition with the NP-PP valence *np.-.*, which we presume is based on *combining NP with*. Here and elsewhere, there is substantial movement in the frequencies, indicating that the frequency models

⁵Note that it is not, however, a cross entropy because it averages information from different distributions indexed by tags and incorporations.

have not stabilized yet.

8 Discussion

The principal positive result of the hundred-verb experiment is the large drop in valence error rate from the baseline with no word-specific information. This was seen for all the conditions which use the treebank syntactic parameters. The drop from the baseline model d_0 to the model d_5 is 20.34 error points. A second positive result is that interleaved estimation has an advantage. For the interleaved scheme, error rate drops 11.11 points in the first step to d_1 , and a further 9.23 points in the subsequent steps to d_5 .

A couple of points of interpretation are obscured by the fact that our unsupervised corpus was a New York Times corpus, while the testing data was a held out portion of the Wall Street Journal treebank. For instance, the straight inside-outside procedure could have an advantage of tuning its syntax model to the NYT. It seems likely that all the training methods would do better if one substituted an unsupervised Wall Street Journal (WSJ) corpus.

The hundred-verb experiment should be regarded as a test of concept. Given the positive results, it is appropriate to run the re-estimation algorithms with a larger and unselected WSJ corpus. While our unsupervised corpus C has a lot of data for the hundred test words, it is in fact only about 1/3 the size of our Treebank sample. This could have the effect that lexical parameters unrelated to the test verbs are poorly estimated. There could be more specific problems with iterative re-estimation of parameters for words which are infrequent in C and absent in t_0 , with a negative impact on both straight inside-outside estimation and the interleaved scheme. In a more comprehensive experiment, we have scaled up our corpus sample by an order of magnitude (about 7.5 million words, as compared to our treebank sample of 755,860 words) in order to obtain estimates of lexical parameters based on significantly more data than there is in the treebank sample. We are running an experiment of this magnitude using WSJ data on a larger computing cluster.

While our evaluation focused on identifying verbal incorporations and the valences they encode, it can be noted that, given our grammar, all or nearly

combining	42.97	VBG.n.-.-	35.88	VBG.n.-.-	29.98	VBG.n.-.-
	19.29	VBG.-.-.-	17.28	VBG.np.-.-	19.53	VBG.np.-.-
	13.41	VBG.np.-.-	15.93	VBG.-.-.-	15.44	VBG.-.-.-
	4.49	VBG.zero.-.-	3.67	VBG.p.-.-	3.26	VBG.p.-.-
completing	65.52	VBG.n.-.-	71.70	VBG.n.-.-	73.10	VBG.n.-.-
	8.92	VBG.-.-.-	7.54	VBG.np.-.-	6.58	VBG.np.-.-
	7.50	VBG.np.-.-	4.01	VBG.-.-.-	3.25	VBG.-.-.-
	1.52	VBG.zero.-.-	0.27	VBG.sbar.-.-	0.16	VBG.zero.-.-
concede	62.14	VBP.sbar.-.-	75.26	VBP.sbar.-.-	80.61	VBP.sbar.-.-
	10.59	VB.sbar.-.-	11.94	VB.sbar.-.-	15.71	VB.n.-.-
	8.36	VBP.zero.-.-	11.77	VB.n.-.-	12.48	VB.sbar.-.-
	6.64	VB.n.-.-	5.10	VB.zero.-.-	4.14	VB.zero.-.-

Figure 8: Frequency of lexical entries for three verbs in d_1 , d_2 , and d_3

all valence errors are associated with errors in tree shape, usually in the tree shape of the VP headed by the verb. So one can expect that improvements in valence identification would be associated with improvements on labeled bracketing measures.

The grammar we worked with has very few incorporated features compared to the PCFG grammars with incorporated features used by Johnson (1998) and Klein and Manning (2003). It would make sense to experiment with grammars with much richer sets of incorporated features, including the mainly tree-geometric features employed by Johnson and Klein and Manning.⁶ This should improve the probability model on trees, and consequently improve the estimated frequency models. Particularly interesting is the possibility of learning features distinct from verbal valence, but which are similar to valence in that treebank data is sparse, and that there is a strong correlation between an open class of lexical items and tree shape. For instance, it might make sense to learn a feature on adverbs which identifies a role as sentence adverbs, VP adverbs, or an element of nominal construction. There are probably many features which have this character.

References

F. Beil, G. Carroll, D. Prescher, S. Riezler and M. Rooth. 1999. Inside-outside estimation of a lexicalized PCFG

⁶Some issues of grammar design arise in applying Johnson and Klein and Manning’s ideas, particularly related to how our valence feature would interact with factoring long right hand sides of VP rules.

for German. ACL 1999.

- T. Briscoe and J. Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Applied NLP 1997*, pages 356-363
- G. Carroll and M. Rooth. 1998. Valence induction with a head-lexicalized PCFG. *EMNLP 1998*.
- M. Johnson. 1998. PCFG models of linguistic tree representations. In *Computational Linguistics 24*(4).
- D. Klein and C. Manning. 2003. Accurate unlexicalized parsing. *ACL 2003*.
- A. Korhonen. 1992. Subcategorization Acquisition. Ph.D. Dissertation, Univ. of Cambridge.
- C. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. *ACL 1993*.
- M. Marcus, B. Santorini and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. In *Computational Linguistics 19*(2).
- D. McCloskey, E. Charniak and M. Johnson. 2006. Effective Self-Training for Parsing. *HLT-NAACL 2006*.
- F. Pereira and Y. Schabes. 1992. Inside-Outside Reestimation from Partially bracketed Corpora. *ACL 1992*.
- D. Prescher. 2003. A Tutorial on the Expectation-Maximization Algorithm Including Maximum-Likelihood Estimation and EM Training of Probabilistic Context-Free Grammars. *ESSLLI 2003*.
- H. Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. *COLING 2004*.
- S. Schulte im Walde. 2002. A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. *LREC 2002*.