Parsing with Local Context

Thesis

By

John Kenton Pate

Graduate Program in Linguistics

Presented in Partial Fulfillment of the Requirements for the Degree Master of Arts
in the Graduate School of The Ohio State University

Thesis Committee:

Chris Brew, Advisor

Laura Wagner, Advisor

Shari Speer

June 2009
Abstract

Treebanks, as a quantitative extension of decades of syntactic theorizing, typically use annotation schemes with a small set of well-motivated phrasal categories. For constituency-based treebanks, these phrasal categories are selected to describe distributional regularities. These treebanks are often used as a data set for estimating Probabilistic Context Free Grammars (PCFGs) for parsing, but the phrasal category sets which are best for constituency description may be suboptimal for constituency parsing. Specifically, phrasal categories may exhibit a probabilistic bias towards different expansions in different parts of the overall tree, and there may be unanticipated but useful correlations between constituency annotation and other levels of linguistic annotation.

In this thesis, the symbol-splitting technique of Johnson (1998) is extended to enrich syntactic categories with information about local syntactic context on the English Penn Treebank and the German Verbmobil II Treebank. The split symbols are then subjected to two different clustering techniques to preserve only relevant category distinctions, forming linguistically-motivated generalizations and assuaging data sparsity. The symbol-splitting and clustering techniques are then employed, on the Verbmobil treebank, to enrich syntactic categories with information about implicit prosodic break strength alone and then together with information about local context.

Local syntactic context is found to be helpful on both treebanks examined. Experiments on the German Verbmobil II Treebank then show that information about
implicit prosodic break strength presents slightly larger gain over information about local syntactic context, and that combining both sorts of information leads to the largest increase in parse accuracy. This research shows that implicit prosody, as imposed by the annotators of the Verbmobil project, does vary with syntactic structure in a useful way outside of a laboratory setting. It is moreover suggestive of exploring prosody as a cue to grammar learning in children.
Acknowledgements

Thanks go first to Detmar Meurers, who introduced me to computational linguistics and spent countless hours helping me clarify my ideas. Most importantly, he demonstrated a patient trust in me that gave me both the confidence and freedom to articulate and implement those ideas.

Chris Brew has also proven an invaluable source of encouragement and insight, and moreover has impressed upon me the importance of simplicity and clarity in both experimental design and presentation.

Laura Wagner has helped me not only see the deeper questions relating to grammar representation and learning, but to appreciate the merits and deficits of conflicting approaches. Her willingness to forcefully advocate for viewpoints which I initially (and naively) dismiss has led me to a much more considered theoretical alignment on issues of language acquisition.

I also owe a debt to Cynthia Clopper instilling in me an appreciation for the staggering amount of variation that real language use exhibits, and showing me that increased variation can make tasks easier.
Vita

September 1986 ................................................................. Born

June 2005 ......................... Cincinnati Hills Christian Academy Highschool

Publications


Field of Study

Major Field: Linguistics
List of Tables

2.1 Information about the two treebanks ........................................... 8

3.1 Parsing results for categories contextualized with local tree information. 14

3.2 Parsing results for unclustered and clustered contextualized categories 21

3.3 Breakdown by original category for the mother-both-sisters grammar 25

4.1 Parsing results for categories split according to implicit prosodic breaks with baseline. .......................................................... 31

4.2 Parsing results for unclustered and clustered categories with information about implicit prosodic breaks, with baseline and grammars with structural information. .............................................. 32

5.1 Parsing results for categories contextualized with information about implicit prosodic breaks and local syntactic information. ................. 36
5.2 Parsing results for grammars integrating prosodic and syntactic information after clustering, with other grammars for comparison. . . . . . 37
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Encoding the local mother context in the category (à la Johnson)</td>
<td>11</td>
</tr>
<tr>
<td>3.2</td>
<td>Result of encoding mother and right sister context in the category</td>
<td>12</td>
</tr>
<tr>
<td>3.3</td>
<td>Dendrogram for English particle (PRT) contextualized with mother-right-sister</td>
<td>18</td>
</tr>
<tr>
<td>3.4</td>
<td>Tree with mother and right sister context categories after clustering</td>
<td>20</td>
</tr>
<tr>
<td>4.1</td>
<td>Encoding rightmost implicit prosodic breaks into phrasal categories</td>
<td>29</td>
</tr>
<tr>
<td>5.1</td>
<td>Encoding mother context and rightmost implicit prosodic breaks into phrasal categories</td>
<td>35</td>
</tr>
</tbody>
</table>
Contents

Abstract ................................................................. ii

Acknowledgements ....................................................... iv

Vita ................................................................. v

List of Tables .......................................................... vi

List of Figures .......................................................... viii

Chapter 1   Introduction .................................................. 1

Chapter 2   Materials & Setup ............................................ 5

2.1 English .............................................................. 5

2.2 German ............................................................. 6

2.3 The Treebanks Compared ............................................ 8
2.4 Experimental Schema ........................................ 9

Chapter 3  Syntactic Context  ...................................... 10

3.1 Raw Contextualization ...................................... 10
3.2 Experiment 1: The space of locally contextualized categories ........ 13
3.3 Experiments 2 & 3: Collapsing irrelevant distinctions .............. 15
3.4 Discussion ......................................................... 25

Chapter 4  Incorporating Prosodic Context  .................. 28

4.1 Materials ......................................................... 28
4.2 Raw Contextualization ...................................... 29
4.3 Experiment 1: Introducing Prosodic Context ..................... 31
4.4 Experiments 2 & 3: Collapsing irrelevant distinctions .............. 31

Chapter 5  Integrating Prosodic and Syntactic Information ........ 34

5.1 Raw Contextualization ...................................... 34
5.2 Experiment 1: Splitting according to Context and Prosody ......... 35
5.3 Experiments 2 & 3: Collapsing irrelevant distinctions .............. 36
Chapter 1

Introduction

The syntactic annotation schemes used in treebanks typically build on long traditions of linguistic investigation and analysis. For constituency-based annotation schemes such as the Penn Treebank (Taylor et al., 2003), the syntactic structure and the constituency labels are rooted in the distributional analysis of structural linguistics and generative theory, which established a range of constituency tests for determining and labelling syntactic structure. While the exact nature and reliability of constituency tests are a recurring subject of debate, the essential idea is readily apparent in the so-called substitution test: a string that is labelled with a syntactic category can be substituted with other strings bearing the same label. The result is another grammatical sentence. In the context of this chapter, the relevant aspect is that substitutability does not make reference to the context in which a category occurs. A string of a given syntactic category can be replaced with another string of that category, independent of syntactic or prosodic context. For example, an NP occurring as sister of a VP (i.e.,
a subject), can be realized in all the same ways as an NP occurring within the VP (i.e., an object). Where this turns out to be empirically incorrect, distinct categories need to be postulated for the different environments to capture the difference.

A Context Free Grammar (CFG) can be made to appreciate statistical regularities by attaching to each rule a probability, forming a Probabilistic Context Free Grammar (PCFG). These rule-associated probabilities are standardly estimated from treebanks, usually by computing the frequency of mother-expansion pairs relative to the frequency of the mother node. In PCFG parsing, the above independence assumption is taken one step further by assuming that the different ways of realizing a given category have the same likelihood independent of the context, whether syntactic or prosodic, in which that category occurs. There is a single probability distribution over the different ways to rewrite a given category, independent of the context in which it occurs. As before, when this assumption turns out to be wrong for a specific category occurring in different contexts, the conclusion is that one needs to assume distinct categories, one for each context in which its realization differs.

Sidestepping the question for which categories in which syntactic contexts additional category distinctions are warranted, Johnson (1998) explored enriching all nonterminal categories in a treebank with information about their local syntactic environment. He enriched each category with the category label of the mother of the local tree that it occurs in and found that this significantly improves the performance of parsing with a PCFG extracted from such a treebank. Klein and Manning (2003) subsequently
pursued manually distinguishing linguistically motivated distinctions, and automatic methods for dividing the traditional linguistic categories into a richer category set have resulted in some of the best PCFG parsing results for the Penn Treebank (cf. Petrov et al., 2006, and references within). However, these methods have grown progressively more complex, typically requiring estimations over entire parse trees and specialized parsing algorithms.

Moreover, endeavors to improve parsing accuracy using local prosodic context have so far met with little success. Gregory et al. (2004) gather acoustic correlates to prosodic tones and breaks from the Switchboard speech corpus, and report only decreases in parsing accuracy when these measures are incorporated as terminals in an otherwise vanilla PCFG. Other work (Kompe et al., 1997; Nöth et al., 2000, e.g.) has successfully exploited prosody to increase parsing speed and to shrink the overall hypothesis space, but these reports do not include parse accuracy figures. Interestingly, Gregory et al. (2004) attempt to incorporate a variety of acoustic cues into their parser, but do not find any improvement.

The experiments outlined in this work seek to investigate and measure the impact of prosodic and syntactic information readily available in the local tree or local prosodic events, using a plain PCFG parser, and working with two dramatically different corpora. In so doing, I seek to isolate the impact of new distinctions in syntactic categories per se. I map out the maximal category space that can result from including attested local context distinctions. Then I return to the original linguistic intuition of
only keeping those new distinctions which differ in their distribution, i.e., I investigate which of the categories resulting from local contextualization are distinctive enough to warrant distinguishing them categorically.

I begin by measuring the gain in parse performance obtained by contextualizing non-preterminals according not only to mother context, as in Johnson (1998), but also to local left sister and local right sister contexts, individually and in combination. Such contextualization results in a large number of syntactic categories and rules, which, as motivated above, are not necessarily useful and for which data sparsity and overfitting to the training data become an issue. I thus proceed by exploring two methods of clustering the newly created contextualized categories. Both are based on the distributional similarity of the contextualized categories as measured by the probability distribution over the local trees dominated by the contextualized categories created for a given category. The first method identifies clusters solely on the basis of the similarity of the probability distributions, while the second method identifies clusters on the basis of expected information gain.

Both clustering methods result in grammars with a dramatically reduced number of categories and rules compared to raw contextualization and, in some cases, a small further improvement in parsing accuracy.
Chapter 2

Materials & Setup

Two treebanks are used in these experiments. One, the Penn Treebank, is used for its size and to enable a straightforward extension of previous work. The other, the Verbmobil II Treebank, is used to replicate the work on a Treebank in a different language and domain, and also for its prosodic annotation (which the Penn Treebank lacks).

2.1 English

Using the standard setup for English PCFG research, the English experiments were run on the Wall Street Journal portion of the third edition of the Penn Treebank Taylor et al. (2003). Sections 2–21 were used for training, section 22 served as a development set, and final parse performance numbers were obtained from section 23. I removed all grammatical function and coindexation tags, so that every nonterminal node contained only the core syntactic category information. All nodes dominating
only the empty string were removed from the trees. No other transformations were performed on the corpus except for the use of contextualized categories explored in Chapter 3.

2.2 German

In the interest of both replicating the results from our English treebank on a very different corpus and language, and extending the technique to the incorporation of prosodic cues, the experiments have also been carried out on a subset of the German portion of the Verbmobil II corpus from the Verbmobil Speech-to-Speech translation system. This is a corpus of telephone speech between individuals arranging business meetings, and contains in total thirty layers of annotation, including six layers of prosodic annotation. However, not every dialogue has all thirty layers, so for these experiments a special subcorpus has been created, consisting of every turn which has both syntactic and prosodic annotation. Of the six layers of prosodic annotation, only one layer (the “Syntactic-Prosodic” level) co-occurs with syntactic annotation frequently. This subcorpus has been divided into 80% train, 10% development, and 10% test. The turns which were used along with their assignment will be provided online at my website:

http://ling.osu.edu/~jpate/

The Lingua::Treebank Perl module was modified to read the relevant tiers from BAS Partitur files into a flexible object-oriented interface, and the modified module
together with a script exemplifying its use will be released shortly on my website, and soon on the Comprehensive Perl Archive Network (CPAN, http://cpan.org/).

Similarly to our Penn Treebank corpus, all grammatical function and coindexation labels were removed, along with all nodes dominating only the empty string. As a spontaneous speech corpus, many turns also contained non-lexical filler vocalizations (which could be easily identified since they were not provided with a Part of Speech annotation). Nodes dominating only non-lexical filler vocalizations were also discarded.

2.2.1 “Syntactic-Prosodic” Annotation

The “Syntactic-Prosodic” annotation is provided on the tier labeled PRO in the BAS Partitur edition of the Verbmobil II corpus. As described in Batliner et al. (1998), this is an annotation scheme designed to provide a fast and coarse description of the prosodic content of utterances without listening to the speech. Each label provides three levels of classification:

1. One of nine types (e.g: sentence, internal constituent, discourse particle).
2. Hierarchical location (one of Main, Subordinate, or Coordinate).
3. Strength (one of 0, 1, 2, or 3; higher number indicates stronger break).

As the labels have been produced only from reading each sentence, they correspond most closely to implicit prosody (Fodor, e.g.), the default prosodic contour that silent
Table 2.1: Information about the two treebanks

(a) Treebank Sizes

<table>
<thead>
<tr>
<th></th>
<th>English (Penn Treebank)</th>
<th>German (Verbmobil II)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Dev</td>
</tr>
<tr>
<td>Sentences</td>
<td>39,832</td>
<td>1,565</td>
</tr>
<tr>
<td>Words</td>
<td>950,009</td>
<td>33,851</td>
</tr>
</tbody>
</table>

(b) Annotation Information

Table 2.1: Information about the two treebanks

readers impose on sentences. The experiments incorporating prosodic information will use only the boundary strength annotation, as the other levels contain more detail than may be specified by implicit prosody.

2.3 The Treebanks Compared

Table 2.1a presents figures on the number of sentences and words for each treebank, and we see that, although the German treebank has roughly the same number of sentences as the English treebank, the German sentences are much shorter. Since the English treebank is edited newspaper text but the Verbmobil treebank consists of spontaneous speech, this is not surprising.

Table 2.1b presents figures on the number of Part of Speech tags, Phrasal categories, and unique subtrees in each corpus (prior to the transformations described in the upcoming chapters). Although the two treebanks, with their different annotation schemes, have a comparable number of Part of Speech (POS) tags, it should be
stressed that the character of these tags is very different. A number of the English Penn Treebank tags encode different classes of punctuation, while none of the Verb-mobil tags do. The German Verbmobil tags, on the other hand, encode numerous morphosyntactic distinctions that are not relevant for the English Penn Treebank.

2.4 Experimental Schema

All experiments begin by estimating a grammar from the training portion of the relevant treebank, after whichever transformation is in question has been applied. A PCFG is estimated by counting the number of times each subtree appears in the corpus, and dividing that by the number of times the mother of the subtree appears in the corpus. Thus, we are using maximum likelihood PCFGs.

Performance figures are then obtained by producing the Viterbi parse of gold-standard POS tags for each sentence of the test corpus, mapping the category labels of the transformed grammar to their original labels, and then producing labeled Precision, Recall, and Balanced F-measure scores of the Viterbi parses against the gold standard.

\[
\text{Precision} = \frac{\# \text{ Correctly labeled constituents}}{\# \text{ Hypothesized Constituents}} \quad \text{Recall} = \frac{\# \text{ Correctly labeled constituents}}{\# \text{ Gold Standard Constituents}}
\]

\[
\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
CHAPTER 3

Syntactic Context\(^1\)

First, syntactic categories will be split according to local structural context. As noted in Chapter 1, the intention is to reveal latent distributional distinctions in a semi-supervised fashion. This chapter also sets out the basic format for Chapters 4 and 5, beginning with an experiment that introduces a large number of distinctions, and then exploring ways to collapse these distinctions using automated clustering techniques.

3.1 Raw Contextualization

As a starting point, all non-terminal and non-preterminal nodes are contextualized according to local mother context, local left sister context, and local right sister context, and in all the combinations thereof. This was done by transforming each training corpus (separately) to enrich each such node with the information from the local tree it occurs in. Figure 3.1 exemplifies the transformation that is performed to obtain

\(^1\)This chapter contains material previously published in Pate and Meurers (2007)
the contextualized categories encoding the mother context. Note that although the examples of this Chapter are all in English, the same transformations are applied to the German data.

The special symbol $\emptyset$ is used to encode that a particular context does not exist, such as the mother context of the VP root node in this example. Note also that the preterminal V is not contextualized, i.e., I do not enrich the set of preterminal categories since otherwise I would also need to obtain the contextualized categories for the input that is to be parsed.\footnote{Such richer lexical categories could possibly be obtained by supertagging, in the spirit of Clark and Curran (2004). Preliminary experiments including gold-standard contextualized preterminals with the input obtained F-scores in the low 90's, suggesting that a supertagging approach would be worthwhile exploring in future work.}

When contextualizing according to both mother and local right sister context, I obtain atomic categories of the form $Cat\_Mother\_RightSister$. Transforming the input of Figure 3.1 using mother and local right sister information results in the tree shown in Figure 3.2.

![Tree Diagram](image)
Finally, when contextualizing a category with both the local mother, local left sister, and local right sister information, I make use of categories of the form

Cat_LeftSister_Mother_RightSister. Throughout, I will refer to the categories as found in the original treebank as the “original categories” and to the categories after the above-described transformation as “contextualized categories”.

Grammars were extracted from the transformed training corpus in the usual way by counting the number of times each local tree appears.\(^3\) Note that one grammar is extracted for English, and a completely separate grammar is extracted for German, and that these grammars are evaluated independently of each other.

I ran our parsing experiments using BitPar Schmid (2004), a freely available, efficient PCFG parser implementation, without lexicalization or additional components. As input, I followed the common practice of using the POS tags from the original corpus.

\[^3\]I performed no smoothing whatsoever. In particular, I did not assign non-zero counts to rules which could be constructed by inserting all potential contextualized categories for each category in each local tree. Only local trees actually occurring in the transformed training corpus were counted.
as input, which prevents confounding the comparative contribution of local context with the accuracy of a separate part-of-speech analysis.\textsuperscript{4}

For evaluation, I mapped the output of the parser, bearing the contextualized categories, back to the original categories simply by stripping off the contextualization suffix. I evaluated parse performance on sentences with 40 or fewer words from each original corpus in terms of labelled bracketing precision, recall, and F-score using the standard EVALB program (Sekine and Collins, 1997). Failed parses were handled by assigning each word in the failed sentence the tag ‘FAILED’, so that the precision, recall, and F-score numbers produced by EVALB correctly reflect the performance on the complete corpus section.

3.2 Experiment 1: The space of locally contextualized categories

I first explored the contribution of the local left daughter, the local right daughter, and the local mother context, as well as combinations thereof. The parsing performance resulting for the locally-contextualized grammars is shown in Table 3.1.

The table separately lists the percentage of failed parses, but these are taken into account in the precision and recall figures. The table also includes the number of categories in the training set, the number of categories occurring only once in the training set, and the number of rules, i.e., distinct local trees in the training set.

\textsuperscript{4}For BitPar I thus used a dummy lexicon that pairs each part-of-speech tag with itself.
The parsing performance figures in Table 3.1 show that, for both treebanks, every contextualization scheme outperforms the baseline grammar (i.e. with the original uncontextualized categories). On the Penn Treebank data, the two-context grammars outperform the single-context grammars, and the mother-right-sister grammar, the best case, outperforms the baseline grammar in terms of F-score by 10 points. On the Verbmobil data, the best performance is obtained by mother context alone, but the two- and three- context grammars obtain comparable scores. Interestingly, the German grammars exhibit a much smaller improvement in overall F-score (about 2-3 points) than do the English grammars.
With both data sets, it is surprising that the three-context mother-both-sisters grammar does not achieve the best performance. As contextualized grammars always easily outperform the uncontextualized baseline, it seems that the decrease in performance is due to the data sparsity concomitant with the exploding number of both rules and categories. The English mother-both-sisters grammar contains almost twice as many categories as the English mother-right-sister, as does the German mother-both-sisters grammar compared to the German mother grammar. For both data sets, fully half of the categories in the mother-both-sisters grammar are observed only once in the appropriate training set, and these grammars also exhibit the most failed parses. The nonce categories are particularly alarming because they are essentially descriptions of single data instances without evidence for generalization.

3.3 Experiments 2 & 3: Collapsing irrelevant distinctions

While it is possible to introduce all category distinctions derivable from the local context, as in the results reported in the previous section, there is a clear price for introducing spurious distinctions. There only is a limited amount of training data, so the more category distinctions are introduced, the less empirical evidence is available to characterize the distributional properties of a given distinction. Rather than handle this data sparsity problem with clever smoothing techniques, I am interested in whether unmotivated distinctions can be automatically collapsed to recover a smaller, more general category set.
Returning to the original motivation for postulating contextualized categories from the introduction, I should only use new category distinctions when a category in a particular context is not realized in the same way as in other contexts. As a measure of how different two contextualized variants of the same category are, I use the relative frequencies over the possible right-hand sides for those categories, i.e., I count in the training data how often each contextualized category immediately dominates which string of daughters, divided by the number of total occurrences of the contextualized category. I represent these relative frequencies as a vector with one dimension for every distinct expansion the original category can take. Note that the vectors of the contextualized categories have the same dimensionality as the vector of the original category they are derived from.

To keep only the relevant contextualized categories, I collapse distinctions between contextualized categories which have similar expansion vectors, which I assess using hierarchical clustering. Agglomerative hierarchical clustering produces a dendrogram, such as the one displayed in Figure 3.3 below, which expresses the similarity not only between individual items but also between groups of items.

Lee (1999) demonstrates that distributional similarity measures differ from each other in terms of their attention to the support of each distribution being compared, and finds that those which focus on the intersection of the supports perform best. The measure must be symmetric not only due to the constraints of hierarchical clustering

---

5The support of a distribution is the set of dimensions with non-zero values.
but also because I aim to collapse distinctions between contextualized categories which may be freely substituted for each other. Indeed, this very intersubstitutability forms the basic notion of a syntactic category. On the Penn Treebank dataset, I tried both symmetric similarity measures examined by Lee, the Jensen-Shannon Divergence and the Manhattan distance, and obtained very similar results from each. This is not surprising, given that they consider the same information and performed similarly on Lee’s own assessment of the measures. For felicity of exposition I present only results for the somewhat simpler Manhattan distance:

\[
\text{Manhattan} (p, q) = \sum_{i=0}^{n} |p_i - q_i|
\]

To obtain the hierarchically clustered dendrograms, I used the \texttt{hclust} routine from the R statistical package (R Development Core Team, 2007). It proceeds using a recursive bottom-up algorithm, each step of which calculates pairwise distances between the clusters of probability vectors under consideration, and assigns the two least distant clusters to a new cluster. I use ‘complete link’ clustering, where the distance between two clusters is the distance between the two members most distant from each other.\(^6\) The algorithm proceeds until only one cluster remains.

Hierarchical clustering can be an effective method for capturing linguistic generalizations, as exemplified by the dendrogram for English particle (‘PRT’) contextualized according to mother and right sister displayed in Figure 3.3.

\(^6\)This marginally outperformed (on the Penn Treebank development set) calculating cluster distance in terms of average vectors and vectors calculated directly from observed cluster member counts.
There are two clear groupings that can be identified in the dendrogram. The smaller grouping, on the right, consists of particles with a variety of mothers but with no local right sister, meaning that the smaller grouping corresponds to phrase-final particles, i.e., particles displaced to the end of the verb phrase or particles to verbs with no locally realized complement. The larger grouping consists of particles with a right sister (in all but one of twenty-one items) and it corresponds roughly to particles between a verb and its complements. In fact, when the left sister context is added, all three distinctions (particle with no complement, non-displaced particle, displaced particle) become evident in the larger dendrogram.

Dendrograms portray how contextualized categories and groups of contextualized categories relate to each other generally. As I want to collapse distinctions between
similar contextualized categories while maintaining distinctions between different con-
textualized categories, I will essentially prune the dendrogram. Not contextualizing at
all is equivalent to cutting the dendrogram at the root node, whereas contextualizing
fully is equivalent to cutting the dendrogram at the leaf nodes. As we have seen, the
former loses useful distributional information, whereas the latter gives rise to data
sparsity problems compromising the reliability of the information. The intention is
to prune somewhere in the middle – where exactly, and based on which criterion, is
discussed further below.

3.3.1 Setup

I ran our clustering methods on two contextualization schemes per dataset. The first
scheme was the one that achieved the best unclustered performance according to Ta-
ble 3.1: mother for the Verbmobil data, mother-right-sister for the Penn Treebank
data. Second, the mother-both-sisters contextualization scheme was used for each
data set in the expectation that it would most clearly demonstrate the benefit of
clustering. Since mother-both-sisters contextualization introduces the most distinc-
tions, I expect it to provide both the most latent relevant distinctions and the most
data sparsity to be assuaged by clustering approaches.

For the two experiments in this section, the training corpus undergoes a second trans-
formation after the contextualization step, in which the contextualization annotation
is replaced with the label of the appropriate cluster. Contextualized categories assigned to singleton clusters naturally can keep their names as these names are straightforward atomic symbols. For example, assuming that the clustering procedure assigns NP\_NP\_∅ and NP\_NP\_PP to a cluster NP\_3, and that the other contextualized categories in Figure 3.2 are assigned to singleton clusters, the tree of Figure 3.2 would be transformed to the one in Figure 3.4.

Figure 3.4: Tree with mother and right sister context categories after clustering

### 3.3.2 Experiment 2: Clustering according to Dendrogram Height

Remember that each iteration of the hierarchical clustering algorithm calculates pairwise distances between all the vectors to be clustered and merges the two least distant vectors. The dendrogram records the distance at which each pair of vectors was merged in the height of the node representing the merge, displayed in the y-axis of the example dendrogram in Figure 3.3. The smaller the height, the more similar the probability distribution over the expansions for the two (sets of) contextualized categories – and the less important the category distinction.

The first method thus simply defines clusters to be the largest sub-dendrograms whose merge height is less than some cut-off value. On the Penn Treebank development set,
I found that a cut-off value of 0.7 worked best for mother-right-sister context and that a cut-off value of 1 worked best for mother-both-sisters context. On the Verbmobil dataset, I found that a cut-off value of 0.1 worked best for mother context, and a cut-off value of 0.5 worked best for mother-both-sisters context. The performance of these height-clustered grammars on the test set is shown in Table 3.2.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>74.79</td>
<td>69.94</td>
<td>72.28</td>
<td>0%</td>
<td>28</td>
<td>1</td>
<td>14,974</td>
</tr>
</tbody>
</table>

### Mother and Right Sister

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered</td>
<td>82.37</td>
<td>82.07</td>
<td>82.22</td>
<td>0%</td>
<td>1,592</td>
<td>640</td>
<td>34,390</td>
</tr>
<tr>
<td>Height-Clust.</td>
<td>82.13</td>
<td>81.83</td>
<td>81.98</td>
<td>0%</td>
<td>849</td>
<td>224</td>
<td>29,259</td>
</tr>
<tr>
<td>KLD-Clust.</td>
<td>82.35</td>
<td>81.41</td>
<td>81.88</td>
<td>0%</td>
<td>215</td>
<td>21</td>
<td>24,392</td>
</tr>
</tbody>
</table>

### Mother and Both Sisters

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered</td>
<td>80.99</td>
<td>81.34</td>
<td>81.16</td>
<td>0.8%</td>
<td>5,177</td>
<td>2,627</td>
<td>52,756</td>
</tr>
<tr>
<td>Height-Clust.</td>
<td>82.17</td>
<td>82.57</td>
<td>82.37</td>
<td>0%</td>
<td>1,672</td>
<td>556</td>
<td>33,628</td>
</tr>
<tr>
<td>KLD-Clust.</td>
<td>82.32</td>
<td>82.24</td>
<td>82.28</td>
<td>0%</td>
<td>495</td>
<td>87</td>
<td>28,781</td>
</tr>
</tbody>
</table>

(a) English Penn Treebank

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>56.78</td>
<td>55.12</td>
<td>55.94</td>
<td>0%</td>
<td>32</td>
<td>3</td>
<td>8,907</td>
</tr>
</tbody>
</table>

### Mother

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered</td>
<td>54.55</td>
<td>63.08</td>
<td>58.51</td>
<td>0%</td>
<td>187</td>
<td>33</td>
<td>11,292</td>
</tr>
<tr>
<td>Height-Clust.</td>
<td>54.63</td>
<td>62.83</td>
<td>58.44</td>
<td>0%</td>
<td>116</td>
<td>15</td>
<td>10,907</td>
</tr>
<tr>
<td>KLD-Clust.</td>
<td>54.62</td>
<td>63.03</td>
<td>58.52</td>
<td>0%</td>
<td>133</td>
<td>15</td>
<td>11,144</td>
</tr>
</tbody>
</table>

### Mother and Both Sisters

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered</td>
<td>53.73</td>
<td>63.57</td>
<td>58.24</td>
<td>0.02%</td>
<td>2,388</td>
<td>1,453</td>
<td>17,112</td>
</tr>
<tr>
<td>Height-Clust.</td>
<td>55.18</td>
<td>61.69</td>
<td>58.25</td>
<td>0%</td>
<td>625</td>
<td>116</td>
<td>12,984</td>
</tr>
<tr>
<td>KLD-Clust.</td>
<td>53.96</td>
<td>63.42</td>
<td>58.31</td>
<td>0%</td>
<td>945</td>
<td>301</td>
<td>15,335</td>
</tr>
</tbody>
</table>

(b) German Verbmobil Treebank

Table 3.2: Parsing results for unclustered and clustered contextualized categories

We see across the board that clustering according to dendrogram height effectively decreases the number of syntactic categories, nonce categories, and distinct rules.
However, we see a benefit in parse performance only in the case of English mother-
both-sisters context. The improvement confirms the conjecture that the earlier under-
performance of mother-both-sisters’ context was due to data sparsity for the English
case. For the German data, performance is essentially the same for both clustered
and unclustered grammars aside from a single failed parse in the unclustered case.
Discussion of why this may be the case is reserved for the final section of this chapter.

3.3.3 Experiment 3: Clustering according to expected information gain

The dendrogram expresses distance relationships between the contextualized expa-
sion vectors without reference to the expansion vector of the original category. The
above clustering method, then, identifies clusters consisting of contextualized cate-
gories which are similar to each other – but not necessarily different from the origi-
nal category. I thus tried a second clustering method, which instead identifies sub-
dendrograms which diverge from the original category. I evaluate the divergence in
terms of the Kullback-Leibler Divergence (KLD)\(^7\) of the expansion distribution from
that of the original category:

\[
\text{KLD}(p, q) = \sum_{i=0}^{n} p_i \cdot \log_2 \left( \frac{p_i}{q_i} \right)
\]

The Kullback-Leibler divergence of one probability distribution \(p\) from another \(q\)
expresses the amount of information lost, in bits, by using \(q\) to encode the behavior

\(^7\)The German data was clustered using Lee (1999)’s \(\alpha\)-skew divergence, which is essentially a
smoothed version of the Kullback-Leibler divergence, with very little smoothing (\(\alpha = 0.95\))
of $p$. In our terms, then, this sum represents the information lost by ignoring that a particular syntactic node appears with a particular class of contexts.$^8$

The KLD-based method described here involves pruning the same dendrogram as in experiment 2. But this time the cut-off values are in terms of the Kullback-Leibler Divergence of the expansion vector at the dendrogram node from that of the overall category. Any sub-dendrogram which exceeds the KLD cut-off value is assigned to its own cluster.

The underlying dendrogram is derived using the same pairwise Manhattan-distance computation as before since the Kullback-Leibler Divergence is not a replacement for the pairwise distance measure. It is not symmetric and it does not satisfy the triangle inequality.$^9$ Note that this method evaluates only the subset of the powerset of the contextualized categories provided by the dendrogram. There may be a member of that powerset which would obtain a large KLD but is not itself made available by the dendrogram because it does not consist of contextualized categories deemed maximally intersubstitutable by the Manhattan distance. This particular method creates the largest possible clusters given that dendrogram by beginning at the root node of the dendrogram and descending until an excess of the cut-off value is encountered. The KLD of successive sub-dendrograms does not necessarily increase monotonically, and beginning from the leaves and moving up would likely result in more clusters each with fewer members. The representative vector used to compute the KLD for

$^8$Since many dimensions of the category vectors are 0, one frequently relies on $0 \cdot \log(0) = 0$.

$^9$In other words, $\text{KLD}(a, b) + \text{KLD}(b, c) \geq \text{KLD}(a, c)$ does not necessarily hold.
a sub-dendrogram is recalculated directly from the counts observed for each member of the sub-dendrogram.

Proceeding top-down in the dendrogram, it is possible to descend all the way to the leaves without obtaining a large Kullback-Leibler divergence from the original category’s expansion vector. As these leaves have been determined to be similar to each other by the dendrogram and non-divergent from the original category, I assign each hitherto unclustered sub-dendrogram to be its own cluster.

In sum, the method used in this experiment collapses distinctions between expansion vectors which are both similar to each other and indistinct from the norm.

Using the development sets, for the English Penn Treebank data I found that a cut-off of 1 worked best for mother-right-sister context and that a cut-off of 2 worked best for mother-both-sisters context. For the German Verbmobil data, I found that a cut-off of 1.5 worked best for mother context, and that a cutoff of 1.25 worked best for mother-both-sisters context. The parse performance results for the test set are included in Table 3.2, displayed in the previous section. The KLD-clustering grammars exhibit performance that is almost identical to that of the height-clustering grammars. However, the KLD-clustering grammars exhibit dramatically fewer categories, nonce categories, and unique rules than do the height-clustering grammars.
### 3.4 Discussion

One way to assess and compare these clustering methods is to look at the number of clusters produced for each original category. Table 3.3 displays such data for the grammar contextualized with mother and both sisters.

<table>
<thead>
<tr>
<th>Original</th>
<th>Ctxt’d</th>
<th>Clustered Height</th>
<th>KLD Height</th>
<th>Original</th>
<th>Ctxt’d</th>
<th>Clustered Height</th>
<th>KLD Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJP</td>
<td>465</td>
<td>178 (38%)</td>
<td>52 (11%)</td>
<td>S</td>
<td>457</td>
<td>52 (11%)</td>
<td>66 (14%)</td>
</tr>
<tr>
<td>ADVP</td>
<td>601</td>
<td>107 (18%)</td>
<td>21 (3%)</td>
<td>SBAR</td>
<td>360</td>
<td>48 (13%)</td>
<td>28 (8%)</td>
</tr>
<tr>
<td>CONJP</td>
<td>46</td>
<td>9 (20%)</td>
<td>2 (4%)</td>
<td>SBARQ</td>
<td>37</td>
<td>18 (49%)</td>
<td>9 (24%)</td>
</tr>
<tr>
<td>FRAG</td>
<td>78</td>
<td>59 (76%)</td>
<td>23 (29%)</td>
<td>SINV</td>
<td>46</td>
<td>32 (70%)</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>INTJ</td>
<td>44</td>
<td>14 (32%)</td>
<td>4 (9%)</td>
<td>SQ</td>
<td>52</td>
<td>39 (75%)</td>
<td>7 (13%)</td>
</tr>
<tr>
<td>LST</td>
<td>13</td>
<td>6 (46%)</td>
<td>1 (8%)</td>
<td>UCP</td>
<td>72</td>
<td>63 (88%)</td>
<td>27 (38%)</td>
</tr>
<tr>
<td>NAC</td>
<td>51</td>
<td>19 (37%)</td>
<td>7 (14%)</td>
<td>VP</td>
<td>319</td>
<td>202 (63%)</td>
<td>24 (8%)</td>
</tr>
<tr>
<td>NP</td>
<td>1043</td>
<td>474 (45%)</td>
<td>76 (7%)</td>
<td>WHADJP</td>
<td>8</td>
<td>3 (38%)</td>
<td>1 (13%)</td>
</tr>
<tr>
<td>NX</td>
<td>59</td>
<td>37 (63%)</td>
<td>18 (31%)</td>
<td>WHADVP</td>
<td>43</td>
<td>37 (7%)</td>
<td>3 (7%)</td>
</tr>
<tr>
<td>PP</td>
<td>734</td>
<td>51 (7%)</td>
<td>23 (3%)</td>
<td>WHNP</td>
<td>54</td>
<td>16 (30%)</td>
<td>8 (15%)</td>
</tr>
<tr>
<td>PRN</td>
<td>325</td>
<td>148 (46%)</td>
<td>66 (20%)</td>
<td>WHPP</td>
<td>13</td>
<td>3 (23%)</td>
<td>1 (8%)</td>
</tr>
<tr>
<td>PRN</td>
<td>93</td>
<td>5 (5%)</td>
<td>2 (2%)</td>
<td>X</td>
<td>57</td>
<td>32 (56%)</td>
<td>16 (28%)</td>
</tr>
<tr>
<td>QP</td>
<td>94</td>
<td>45 (48%)</td>
<td>4 (4%)</td>
<td>ROOT</td>
<td>1</td>
<td>1 (100%)</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>RRC</td>
<td>11</td>
<td>7 (64%)</td>
<td>2 (18%)</td>
<td>PRU</td>
<td>ADVP</td>
<td>1</td>
<td>1 (100%)</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>5,177</td>
<td>1,672</td>
<td>495</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3: Breakdown by original category for the mother-both-sisters grammar

For each original category, it lists the number of subcategories created by raw contextualization, after height-clustering, and after KLD-clustering along with the ratio of clustered subcategories to contextualized subcategories within each clustering method as a percent. I observe great variability in the number of distinct contexts that original categories appear in, ranging from just eight contexts in the case of WHADJP to 1,043 in the case of NP. Wh-phrases appear in a substantially smaller number of contexts than do their non-wh-counterparts. Although the number of contextualized subcategories obviously limits the number of clustered subcategories, the number of
contextualized categories does not necessarily predict the number of clustered categories. For example, PRT (Particle) appears in 93 distinct contexts whereas UCP (Unlike Coordinate Phrase) appears in only 72, but PRT is consolidated into 5 clusters (height) or 2 clusters (KLD), whereas UCP appears significantly more diverse with 63 (height) and 27 (KLD) clusters.

Both clustering methods rely on pulling out sub-dendrograms on the basis of some cut-off value, which I crudely optimized by trying values to maximize the F-score on the development set. However, the significant variability apparent from the sub-cluster counts in Table 3.3 suggests that a single cut-off value, while optimized for the contextualization scheme as a whole, is not necessarily optimal for each original category. Indeed, in the discussion of PRT dendrograms in section 3.3 I mentioned that three theoretically-appealing subclusters are readily apparent from the mother-both-sisters dendrogram, but neither method successfully identifies exactly three. Height-clustering creates 5 clusters, forming one cluster each for two of the groups, but splitting the third into three singleton subcategories. KLD-clustering obtains two clusters, conflating into one subcategory the first two groups distinguished by height-clustering, but appealingly clustering into one subcategory the remaining three ‘missed’ by height-clustering. An approach which optimizes the cut-off values for each original category might improve clustering further.

It remains curious that the German grammars did not improve in performance upon clustering. One possibility has to do with the difference in character, noted in Sec-
tion 2.3, between the tagsets for each Treebank. The German tagset encodes more morphosyntactic contrasts and contains no punctuation tags, while the English tagset contains several punctuation tags but (as is appropriate for English) maintains few morphosyntactic contrasts. It may be the case that the original German tags were already largely saturated with information about local context, and data sparsity simply presented less of a problem.
Chapter 3 sought to incorporate information about local structural context into enriched sets of syntactic categories. Raw contextualization proved helpful for both English newspaper editorials and German situated dialogs, and clustering on top of contextualization served to reduce overall grammar size without harming performance (even producing a slight increase in performance for the English data).

In this section, the German grammars are manipulated to incorporate break strength from the pro tier of the BAS Partitur edition of the Verbmobil II treebank. The experiments take the same overall form as those in Chapter 3, beginning with raw contextualization and proceeding with the same two clustering schemes.

4.1 Materials

These experiments are run on the same 80%/10%/10% train/dev/test division of the German Verbmobil corpus as those in Chapter 3. No experiments were run on the
Penn Treebank, as it lacks relevant annotation.

4.2 Raw Contextualization

Prosodic information was introduced in a similar symbol splitting manner. Figure 4.1 provides an example of this contextualization process, where each non-terminal node is annotated with the strength of the implicit prosodic break (the number following the ".") of its rightmost terminal daughter.

Figure 4.1: Encoding rightmost implicit prosodic breaks into phrasal categories

Note that in these experiments, the introduced annotation is made available on the preterminals (which is what the parser sees as input at test runtime) as well, whereas the structural contextualization was not made available on the preterminals. Structural context was omitted on preterminals before because it is difficult to see how to obtain it without simply providing the parser with an additional level of gold-standard structural information, and it was against precisely that gold-standard structure that the grammars were being evaluated. The prosodic annotation is being included on preterminals to evaluate the utility of implicit prosody as reported by human annotators.
Due to the relatively small training corpus, many of the development sentences became un-parseable upon introduction of the prosodic annotation. To counter this, a very simple back-off strategy was implemented: if a sequence of preterminals was unparseable, the parser was free to change the break strength annotation of the preterminals until the sentence became parseable. For example, if the sequence:

\textbf{VAFIN.1 CARD.1 PPER.0 VVINF.2}

is unparseable, the parser may try, for example, changing the break strength annotation of PPER from 0 to 2:

\textbf{VAFIN.1 CARD.1 PPER.2 VVINF.2}

This back-off strategy was implemented in the weighted lexicon provided to the grammar. In Chapter 3, the lexicon simply mapped each Part Of Speech (POS) tag to itself with a weight of 1. For example, the entry for PPER looked like:

\textbf{PPER PPER 1}

For the experiments in this chapter and the next, each POS tag is mapped to itself with a large weight, and mapped to other versions of itself (with different strength annotations) with a small weight. For example, the entry for PPER.1 looks like:

\textbf{PPER.1 PPER.0 1 PPER.1 10000 PPER.2 1 PPER.3 1}
Table 4.1: Parsing results for categories split according to implicit prosodic breaks with baseline.

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>56.78</td>
<td>55.12</td>
<td>55.94</td>
<td>0%</td>
<td>32</td>
<td>3</td>
<td>8,907</td>
</tr>
<tr>
<td>Right Break (R)</td>
<td>56.37</td>
<td>60.84</td>
<td>58.52</td>
<td>0%</td>
<td>265</td>
<td>147</td>
<td>14,313</td>
</tr>
<tr>
<td>Left Break (L)</td>
<td>56.92</td>
<td>62.11</td>
<td>59.40</td>
<td>0%</td>
<td>270</td>
<td>156</td>
<td>14,028</td>
</tr>
<tr>
<td>R &amp; L</td>
<td>50.09</td>
<td>47.12</td>
<td>48.56</td>
<td>0%</td>
<td>298</td>
<td>38</td>
<td>16,313</td>
</tr>
</tbody>
</table>

4.3 Experiment 1: Introducing Prosodic Context

Table 4.1 presents the parsing results for grammars manipulated to contain the rightmost break strength, the leftmost break strength, and both together, along with the results for the baseline grammar which has no prosodic (or structural) information whatsoever; this is the same baseline as was used for Chapter 3. The table shows that the rightmost break provides the greatest improvement over the baseline, and that the leftmost break is helpful. To the author’s knowledge, this is the first reported use of prosodic information towards the improvement of parsing accuracy. Surprisingly, incorporating both breaks together leads the grammar to perform well below baseline, suggesting that the both-breaks grammar is especially hurt by data sparsity.

4.4 Experiments 2 & 3: Collapsing irrelevant distinctions

The experiments in this section apply the same clustering techniques that were explored in Chapter 3. Experiment 2 generates a dendrogram using the same hierarchical clustering algorithm, and defines clusters by a simple cut. Experiment 3 takes those dendrograms, and defines clusters in terms of subdendrograms which produce
clusters that are more divergent, according to the Kullback-Leibler Divergence based on \( \alpha \)-skew divergence \( (\alpha = 0.95) \), than some value. As before, the precise cut-offs were crudely optimized by simply trying different values on the development set, and using the grammar which performs best on the development set to parse the test set.

Since there are only three contextualization schemes (L, R, and L&R) and one data set, clustering results are presented for all schemes.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline –</td>
<td>56.78</td>
<td>55.12</td>
<td>55.94</td>
<td>0%</td>
<td>32</td>
<td>3</td>
<td>8,907</td>
</tr>
</tbody>
</table>

Mother (from Chapter 3, for comparison)

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered –</td>
<td>54.55</td>
<td>63.08</td>
<td>58.51</td>
<td>0%</td>
<td>187</td>
<td>33</td>
<td>11,292</td>
</tr>
<tr>
<td>KLD-Clust. 1.5</td>
<td>54.62</td>
<td>63.03</td>
<td>58.52</td>
<td>0%</td>
<td>133</td>
<td>15</td>
<td>11,144</td>
</tr>
</tbody>
</table>

Right break

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered –</td>
<td>56.37</td>
<td>60.84</td>
<td>58.52</td>
<td>0%</td>
<td>265</td>
<td>147</td>
<td>14,313</td>
</tr>
<tr>
<td>Height-Clust. 0.1</td>
<td>56.92</td>
<td>62.11</td>
<td>59.40</td>
<td>0%</td>
<td>104</td>
<td>7</td>
<td>14,027</td>
</tr>
<tr>
<td>KLD-Clust. 2.75</td>
<td>56.46</td>
<td>60.99</td>
<td>58.64</td>
<td>0%</td>
<td>107</td>
<td>8</td>
<td>14,314</td>
</tr>
</tbody>
</table>

Left break

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered –</td>
<td>56.92</td>
<td>62.11</td>
<td>59.40</td>
<td>0%</td>
<td>270</td>
<td>156</td>
<td>14,028</td>
</tr>
<tr>
<td>Height-Clust. 0.1</td>
<td>57.03</td>
<td>62.15</td>
<td>59.48</td>
<td>0%</td>
<td>97</td>
<td>6</td>
<td>14,006</td>
</tr>
<tr>
<td>KLD-Clust. 2</td>
<td>56.88</td>
<td>61.94</td>
<td>59.30</td>
<td>0%</td>
<td>97</td>
<td>9</td>
<td>13,924</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unclustered –</td>
<td>50.09</td>
<td>47.12</td>
<td>48.56</td>
<td>0%</td>
<td>298</td>
<td>38</td>
<td>16,313</td>
</tr>
<tr>
<td>Height-Clust. 1.25</td>
<td>52.11</td>
<td>46.87</td>
<td>49.39</td>
<td>0%</td>
<td>54</td>
<td>11</td>
<td>13,688</td>
</tr>
<tr>
<td>KLD-Clust. 0.1</td>
<td>50.97</td>
<td>46.31</td>
<td>48.53</td>
<td>0%</td>
<td>79</td>
<td>11</td>
<td>14,645</td>
</tr>
</tbody>
</table>

Table 4.2: Parsing results for unclustered and clustered categories with information about implicit prosodic breaks, with baseline and grammars with structural information.

Table 4.2 presents parsing results for the prosodically-informed grammars once they have been subjected to both clustering approaches, and presents both the original baseline and the best performing clustered grammar from Chapter 3 for comparison. Table 4.2 shows that the clustered grammars are substantially smaller than the unclustered grammars in terms of the total number of categories and the number of
nonce categories, and somewhat smaller than the unclustered grammars in terms of the total number of rules. Table 4.2 also shows that while the clustering approaches are somewhat helpful for improving the performance of the grammar incorporating both Leftmost and Rightmost breaks, clustering is not nearly enough to return performance even to baseline. Notably, although the clustering approaches do not produce much of a difference in performance for grammars incorporating the leftmost prosodic break, clustering according to dendrogram height improves F-score by nearly a full point (with most of the improvement in Recall). Finally, Table 4.2 shows that splitting according to prosodic annotation outperforms splitting according to local structural context.
Chapter 3 showed that splitting syntactic categories according to local structural context was useful for improving parse performance in both English Wall Street Journal editorials and German situated dialogs. Chapter 4 followed this up by showing that splitting according to the prosodic annotation of the pro tier of the Verbmobil II Treebank is even more useful for improving parse performance. In this chapter, syntactic categories are split to encode both local structural context and implicit prosodic break context.

5.1 Raw Contextualization

Figure 5.1 presents an example (mother context with rightmost break strength) of the sort of transformation that is applied in this chapter. Note that the structural context itself does not contain prosodic annotation; that is, the mother of the MF
node is \textit{SIMPX} with rightmost break strength of 2, but the local structural context of \textit{MF} is only \textit{SIMPX}, not \textit{SIMPX.2}.

Figure 5.1: Encoding mother context and rightmost implicit prosodic breaks into phrasal categories

5.2 Experiment 1: Splitting according to Context and Prosody

Following the experimental format laid out in previous chapters, syntactic categories are first simply split on several factors without clustering. Rather than split on every combination of factors, this experiment will be restricted to Mother and Mother-both-sisters structural context. All three sorts of prosodic information are explored (leftmost break strength, rightmost break strength, and the strength of both breaks together).

Table 5.1 presents the parsing results for these six grammars, together with the original baseline from Chapter 3. All combinations of structural and prosodic information explored exhibit an improvement over the baseline, except for those employing both leftmost and rightmost prosodic break strength. These grammars exhibit a
<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>56.78</td>
<td>55.12</td>
<td>55.94</td>
<td>0%</td>
<td>32</td>
<td>3</td>
<td>8,907</td>
</tr>
<tr>
<td><strong>Mother</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Break (R)</td>
<td>53.87</td>
<td>66.26</td>
<td>59.29</td>
<td>0.02%</td>
<td>524</td>
<td>93</td>
<td>18,352</td>
</tr>
<tr>
<td>Left Break (L)</td>
<td>54.28</td>
<td>67.06</td>
<td>60.00</td>
<td>0.02%</td>
<td>518</td>
<td>94</td>
<td>18,032</td>
</tr>
<tr>
<td>R &amp; L</td>
<td>41.68</td>
<td>44.32</td>
<td>42.96</td>
<td>0.02%</td>
<td>1,122</td>
<td>272</td>
<td>20,510</td>
</tr>
<tr>
<td><strong>Mother and Both Sisters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right Break (R)</td>
<td>52.48</td>
<td>67.27</td>
<td>58.96</td>
<td>0.02%</td>
<td>3,860</td>
<td>2,017</td>
<td>25,306</td>
</tr>
<tr>
<td>Left Break (L)</td>
<td>52.33</td>
<td>67.04</td>
<td>58.78</td>
<td>0.02%</td>
<td>3,695</td>
<td>1,925</td>
<td>24,950</td>
</tr>
<tr>
<td>R &amp; L</td>
<td>37.80</td>
<td>37.89</td>
<td>37.85</td>
<td>6.60%</td>
<td>5,000</td>
<td>2,630</td>
<td>27,472</td>
</tr>
</tbody>
</table>

Table 5.1: Parsing results for categories contextualized with information about implicit prosodic breaks and local syntactic information.

substantial drop in performance from the baseline; given the similarly underwhelm-
ing performance of grammars incorporating both prosodic breaks in Chapter 4, it is not surprising that introducing even more distinctions exacerbates problems of data sparsity. All of the grammars explored here demonstrate the expected explosion of categories, nonce categories, and rules, increasing the number of rules over threefold and the number of categories over 150-fold in the most extreme case.

5.3 Experiments 2 & 3: Collapsing irrelevant distinctions

I proceed by following the same experimental layout as Chapters 3 and 4 by applying clustering techniques to the larger but more informative grammars obtained in Section 5.2 to recover a smaller, more general set of syntactic categories. The clustering techniques are the same as those described previously, obtaining a dendrogram to explore the similarity of the raw categories, and then extracting subdendrograms first according to merge height and then according to divergence from the overall category.
The clustering approaches will be applied to the grammar obtained by splitting according to Mother context and left boundary strength, as this obtained the greatest F-score in Table 5.1, and also to the grammar with Mother-Both-Sisters context and both left and right boundary strength, under the hypothesis that this grammar has the most latent information and data sparsity to be overcome by clustering.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Prec.</th>
<th>Recall</th>
<th>F</th>
<th>Failed</th>
<th>Categ.</th>
<th>Nonce</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline –</td>
<td>56.78</td>
<td>55.12</td>
<td>55.94</td>
<td>0%</td>
<td>32</td>
<td>3</td>
<td>8,907</td>
</tr>
<tr>
<td>Mother (from Chapter 3, for comparison)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclustered –</td>
<td>54.55</td>
<td>63.08</td>
<td>58.51</td>
<td>0%</td>
<td>187</td>
<td>33</td>
<td>11,292</td>
</tr>
<tr>
<td>KLD-Clust. 1.5</td>
<td>54.62</td>
<td>63.03</td>
<td>58.52</td>
<td>0%</td>
<td>133</td>
<td>15</td>
<td>11,144</td>
</tr>
<tr>
<td>Left break (from Chapter 4, for comparison)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclustered –</td>
<td>56.92</td>
<td>62.11</td>
<td>59.40</td>
<td>0%</td>
<td>270</td>
<td>156</td>
<td>14,028</td>
</tr>
<tr>
<td>Height-Clust. 0.1</td>
<td>57.03</td>
<td>62.15</td>
<td>59.48</td>
<td>0%</td>
<td>97</td>
<td>6</td>
<td>14,006</td>
</tr>
<tr>
<td>Mother Context with Left Break Strength</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclustered –</td>
<td>54.28</td>
<td>67.06</td>
<td>60.00</td>
<td>0.02%</td>
<td>518</td>
<td>94</td>
<td>18,032</td>
</tr>
<tr>
<td>Height-Clust. 0.1</td>
<td>55.11</td>
<td>65.64</td>
<td>59.92</td>
<td>0.02%</td>
<td>170</td>
<td>12</td>
<td>16,063</td>
</tr>
<tr>
<td>KLD-Clust. 1.5</td>
<td>54.84</td>
<td>65.99</td>
<td>59.90</td>
<td>0%</td>
<td>158</td>
<td>4</td>
<td>16,274</td>
</tr>
<tr>
<td>Mother and Both Sisters with Both Break Strengths</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclustered –</td>
<td>37.80</td>
<td>37.89</td>
<td>37.85</td>
<td>6.60%</td>
<td>5,000</td>
<td>2,630</td>
<td>27,472</td>
</tr>
<tr>
<td>Height-Clust. 1.5</td>
<td>49.48</td>
<td>45.86</td>
<td>47.60</td>
<td>0%</td>
<td>1,294</td>
<td>312</td>
<td>19,156</td>
</tr>
<tr>
<td>KLD-Clust. 0.5</td>
<td>46.26</td>
<td>42.66</td>
<td>44.39</td>
<td>0%</td>
<td>130</td>
<td>6</td>
<td>16,423</td>
</tr>
</tbody>
</table>

Table 5.2: Parsing results for grammars integrating prosodic and syntactic information after clustering, with other grammars for comparison.

Table 5.2 presents performance figures for these grammars. The performance of the Mother-both-sisters grammar with both break strengths improves substantially under height-clustering, but remains well below baseline. This suggests that the corpus is too small to support effective clustering of a grammar containing so many distinctions. The performance of the grammars derived by clustering the Mother-right-break grammar is more encouraging, however. Although a very small drop in F-score is observed
under both clustering approaches, the clustered grammars are substantially smaller than the unclustered grammar on all counts. Most interestingly, the grammars incorporating mother context and leftmost break strength outperform the best grammar explored so far (left break strength) in overall F-score, exhibiting an advantage in labeled constituent recall but a small disadvantage in labeled constituent precision.
CHAPTER 6

Conclusion

Extending ideas from Johnson (1998), I explored enriching the syntactic category distinctions in two very different Treebanks with the contextual information available within the local tree. PCFG parsing experiments using a grammar extracted from the enriched corpus confirm that significant information about the expansion properties of syntactic categories is immediately available in the local context. On one of the treebanks, this technique was extended to incorporate prosodic break strength (annotated during silent reading) into syntactic categories, and then to incorporate information about both local structural context and implicit prosodic break strength. In all cases where only one of leftmost break strength or rightmost break strength were used, the increased number of distinctions led to increased grammar performance. However, blindly introducing all contextually possible category distinctions results in well-known data sparsity issues, with many of the possible categories only rarely or
never occurring in the training data. I therefore explored introducing only those contextualized categories which are distributionally distinct, as measured by the probability distribution over the local expansions. On the Penn Treebank dataset, I showed that clustering can identify theoretically appealing clusters, automatically exploiting linguistic generalizations. Clustering based on the distance between contextualized categories was by and large equally effective as clustering based on the divergence of a contextualized category from the original category, even though the latter resulted in fewer clusters.

The relevance of the local context for defining or enriching the category set can be seen as an interesting reflection of the role of such context frames in the acquisition of categories during language acquisition Mintz (2003). The utility of these prosodic break strengths may also be relevant for language acquisition. Soderstrom et al. (2003) found, in an artificial language task, that their six-month-old and nine-month-old infants learn phrasal distinctions only in the presence of a contrast in break strength (realized on a number of dimensions, such as final vowel length, f0 peak height and alignment, and so on). The annotation used here was obtained during silent reading and so does not bear an obvious relationship to the acoustic realization of prosodic events that children hear. However, as discussed in Fodor, implicit prosody does have a psychological reality which children must at some point learn. Once learned, these experiments show that fairly simple grammar enrichment techniques are sufficient to exploit them.
Bibliography


