Alignment-Guided Chunking

Yanjun Ma, Nicolas Stroppa, Andy Way
National Centre for Language Technology
School of Computing
Dublin City University
Glasnevin, Dublin 9, Ireland
{yma,nstroppa,away}@computing.dcu.ie

Abstract

We introduce an adaptable monolingual chunking approach–Alignment-Guided Chunking (AGC)–which makes use of knowledge of word alignments acquired from bilingual corpora. Our approach is motivated by the observation that a sentence should be chunked differently depending the foreseen end-tasks. For example, given the different requirements of translation into (say) French and German, it is inappropriate to chunk up an English string in exactly the same way as preparation for translation into one or other of these languages.

We test our chunking approach on two language pairs: French–English and German–English, where these two bilingual corpora share the same English sentences. Two chunkers trained on French–English (*FE-Chunker*) and German–English (*DE-Chunker*) respectively are used to perform chunking on the same English sentences. We construct two test sets, each suitable for French–English and German–English respectively. The performance of the two chunkers is evaluated on the appropriate test set and with one reference translation only, we report F-scores of 32.63% for the *FE-Chunker* and 40.41% for the *DE-Chunker*.

1 Introduction

Chunking plays an important role in parsing, information extraction and information retrieval. Chunking is often a useful preprocessing step for many bilingual tasks, such as machine translation, cross language information retrieval, etc.

We introduce an adaptable chunking approach guided by word alignments automatically acquired from a bilingual corpus. Our approach is motivated by the observation that a sentence should be chunked differently depending the end-task in mind. Our approach employs bilingual word alignment in training and is tested on the monolingual chunking task. Our goal is to build adaptable monolingual chunkers for different language pairs, with the aim of facilitating bilingual language processing tasks.

We investigate our chunking approach on two language pairs: French–English and German–English, where these two bilingual corpora share the same English sentences. Two chunkers trained on French–English (*FE-Chunker*) and German–English (*DE-Chunker*) respectively are used to perform chunking on the same English sentences. We construct two test sets, each suitable for French–English and German–English respectively. The performance of the two chunkers is evaluated on the appropriate test set and with one reference translation only, we report F-scores of 32.63% for the *FE-Chunker*

and 40.41% for the *DE-Chunker*. We also extend our chunking approach with *Multi-level Chunking*, which is more tolerant of any chunking errors obtained.

The remainder of this paper is organized as follows. In Section 2, we review the previous research on chunking including monolingual chunking and bilingual chunking. Section 3 describes our chunking method. In Section 4, the experimental setting is described. In Section 5, we evaluate our chunking method on a one-reference 'gold standard' testset. Section 6 concludes the paper and gives avenues for future work.

2 Previous Research

2.1 Monolingual Chunking

Most state-of-the-art monolingual chunking methods are linguistically motivated. The CoNLL-2000 shared task (Tjong Kim Sang and Buchholz, 2000) defined chunking as dividing text into syntactically related non-overlapping groups of words. Chunks are directly converted from the Penn Treebank (Marcus et al., 1993) and each chunk is labelled with a specific grammatical category, such as NP, VP, PP, ADJP etc. This chunking method is sensitive to the grammars of a specific language and performs chunking in a monolingual context.

Marker-based chunking is another syntax-aware chunking strategy. This chunking approach is based on the "Marker Hypothesis" (Green, 1979), a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific closed set of lexemes or morphemes which signify context. Using a set of closed-class (or "marker") words, such as determiners, conjunctions, prepositions, possessive and personal pronouns, aligned source–target sentences are segmented into chunks. A chunk is created at each new occurrence of a marker word, with the restriction that each chunk must contain at least one content (or non-marker) word. Although marker-based chunking has been used in bilingual tasks such as machine translation between European languages (Gough and Way, 2004; Groves and Way, 2005; Stroppa and Way, 2006), which are relatively similar with regard to marker words and word orders, it is less appropriate for language pairs as different as Chinese and English (Ma, 2006).

2.2 Bilingual Chunking

Bilingual chunkers are usually based on parsing technology. (Wu, 1997) proposed Inversion Transduction Grammar (ITG) as suitable for the task of bilingual parsing. The stochastic ITG brings bilingual constraints to many corpus analysis tasks such as segmentation, bracketing, and parsing, which are usually carried out in a monolingual context. However, it is difficult to write a broad bilingual ITG grammar capable of dealing with long sentences. (Wang et al., 2002) proposed an algorithm integrating chunking and alignment and obtained good precision. However, this method needs quite a lot of syntax information and prior knowledge. (Liu et al., 2004) proposed an integrated probabilistic model for bilingual chunking and alignment independent of syntax information and grammatical rules.

3 Alignment-Guided Chunking

3.1 Notation

While in this paper, we focus on both French–English and German–English, the method proposed is applicable to any language pair. The notation however assumes the French–English task in what follows.

Given a French sentence f_1^I consisting of I words $\{f_1, \ldots, f_I\}$ and an English sentence e_1^J consisting of J words $\{e_1, \ldots, e_J\}$, $A_{F \to E}$ (resp. $A_{E \to F}$) will denote a French-to-English (resp. an English-to-French) word alignment between f_1^I and e_1^J. As 1-to-n alignments are quite common, $A_{F \to E}$ can be represented as a set of pairs $a_i = \langle f_i, E_i \rangle$ denoting a link between one single French word f_i and a few English words E_i (and similarly for $A_{E \to F}$). The set E_i is empty if the word f_i is not aligned to any word in e_1^J.

Given a French–English sentence pair $\langle f_1^I, e_1^J \rangle$, suppose f_i is aligned to a set of En-

glish words $E_i = \{e_j, \ldots, e_{j+m}\}$, and $E_{i+1}^I = E_{i+1} \bigcup \cdots \bigcup E_I = \{e_k, \ldots, e_{k+n}\}$ denotes a union of English words that are aligned to the set of French words $\{f_{i+1} \ldots, f_I\}$. There should be a partition between f_i and f_{i+1}, iff. $k > j + m$. We can partition the English sentence using the same method.

Given a French–English sentence pair and the word alignment between them, we can partition both French and English sentences following the criteria described above. As this chunking is guided by the word alignment, we call it *Alignment-Guided Chunking*.

Assume the French–English sentence pair and their word alignment in (1):

(1) *French*: Cette ville est chargée de symboles puissants pour les trois religions monothéistes .

English: The city bears the weight of powerful symbols for all three monotheistic religions .

Word alignment: 0-0 1-1 2-2 3-4 4-5 5-7 6-6 7-8 8-9 9-10 10-12 11-11 12-13

The AGC chunks derivable via our method are displayed in Figure 1.

Cette ||| ville ||| est ||| chargée ||| de ||| symboles puissants ||| pour ||| les ||| trois ||| religions monothéistes ||| .

The ||| city ||| bears ||| the weight ||| of ||| powerful symbols ||| for ||| all ||| three ||| monotheistic religions ||| .

Figure 1: Example of AGC chunks

Note that the method is able to capture adjective–noun combinations in each language, as well as the determiner-noun pair in English.

3.2 Data Representation

(Ramshaw and Marcus, 1995) introduced a data representation for baseNP chunking by converting it into a tagging task: words inside a baseNP were marked I, words outside a baseNP receive an O tag, and a special tag B was used for the first word inside a baseNP immediately following another baseNP. (Tjong Kim Sang and Veenstra, 1999) examined seven different data representations for noun phrase chunking and showed that the choice of data representation has only a minor influence on chunking performance.

In our chunking approach, every word is classified into a chunk and no fragments are left in a sentence. Accordingly, we do not need the tag O to mark any word outside a chunk. We can employ three data representations similar to (Tjong Kim Sang and Veenstra, 1999) named IB, IE, IBE1, IBE2, where the I tag is used for words inside a chunk. They differ in their treatment of chunk-initial and chunk-final words as shown in Table 1.

In our experiments, we use IE to represent the data, so that the problem of chunking is transformed instead into a binary classification task. The IE tag representation for the English sentence in Figure 1 is shown in (2):

(2) The/E city/E bears/E the/I weight/E of/E powerful/I symbols/E for/E all/E three/E monotheistic/I religions/E ./

Again, note the dependence of determiners and adjectives on their following head noun.

3.3 Parameter Estimation

In this section, we briefly introduce two well-known machine learning techniques we used for parameter estimation, namely Maximum Entropy (MaxEnt) and Memory-based learning (MBL). Both of them are widely used in Natural Language Processing (NLP).

Maximum Entropy was first introduced in NLP by (Berger et al., 1996). It is also used for chunking (Koeling, 2000). Memory-based learning (e.g. (Daelemans and Van den Bosch, 2005)) is based on the simple twin ideas that:

- learning is based on the storage of exemplars, and
- processing is based on the retrieval of exemplars, or for similarity-based reasoning, on the basis of exemplars.

IB	all chunk-initial words receive a B tag
IE	all chunk-final words receive a E tag
IBE1	all chunk-initial words receive a B tag, all chunk-final words receive a E tag; if there is only one word in the chunk, it receives a B tag
IBE2	all chunk-initial words receive a B tag, all chunk-final words receive a E tag; if there is only one word in the chunk, it receives a E tag

Table 1: Data Representation for Chunking

MBL can be used simply and effectively to perform a range of classification tasks.

3.4 Feature Selection

Feature selection is important for the performance for both machine learning techniques. In practice, the features we used are shown in Table 2. The information we used was contained in a 7-word window, i.e. the leftmost three words and their Part-of-Speech (POS) tags, the current word and its POS tag, and the rightmost three words and their POS tags.

3.5 Multi-level Chunking

3.5.1 Notation

Given a sentence s_1^I containing I words $\{w_1, \ldots, w_I\}$, chunking can be considered as the process of inserting a chunk boundary marker c_i between two consecutive words w_i, w_{i+1}. The probability of inserting a chunk boundary marker c_i between two consecutive words w_i, w_{i+1} (i.e. the partition probability) can be defined as:

$$\mathbb{p}(c_i|s_1^I) = p_{\lambda_1^M}(c_i|s_1^I)$$
$$= \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(c_i, s_1^I)]}{\sum_{c_i'} \exp[\sum_{m=1}^{M} \lambda_m h_m(c_i', s_1^I)]}$$

For sentence s_1^I, we can derive a set of partition probabilities with $I-1$ elements:

$$PP = \{\mathbb{p}(c_1|s_1^I), \ldots, \mathbb{p}(c_{I-1}|s_1^I)\}$$

By setting different thresholds for our partition probabilities, we can obtain different chunking results for the same sentence. This threshold can be adjusted depending on the task at hand with the result that different chunking patterns for the same sentence are obtained. We call this chunking model *Multi-level Chunking*.

If we relate this model to our IE data representation (cf. (2) above), it is equivalent to determining the probability of a word being labelled E. While most chunking approaches are essentially classification-based, our model attempts to transform the classification-based approach into a ranking problem and decide the partition point of a sentence by examining competitive scores at each point. We call this chunking approach *Ranking-based Chunking*.

The set of parameters in this model include (i) the set of partition probabilities, and (ii) estimates of thresholds for partition probabilities bearing in mind the specific task to be performed.

Figure 2 gives an example of the distribution of the partition probability.

The ||| city ||| bears ||| the ||| weight ||| of ||| powerful |||
0.7069 0.5307 0.5467 0.4527 0.3777 0.4098 0.4162
symbols ||| for ||| all ||| three ||| monotheistic ||| religions ||| .
0.4318 0.4253 0.3807 0.5655 0.5078 0.9796

Figure 2: Example of Multi-level chunking

If we take 2 words as our average chunk length, we can chunk sentence (2) as shown in Figure 3.

The ||| city ||| bears ||| the ||| weight of powerful symbols for all three ||| monotheistic religions ||| .

Figure 3: Example of chunking result using Multi-level chunking

Note that several words *weight ... three* have been combined into one chunk in Figure 3 based on the partition probabilities shown in Figure 2.

Word	w_{i-3}	w_{i-2}	w_{i-1}	w_i	w_{i+1}	w_{i+2}	w_{i+3}
POS	t_{i-3}	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}	t_{i+3}

Table 2: Features for chunking

3.5.2 Threshold Estimation

The average length of chunks can be estimated from training data acquired following the criteria described in Section 3.1. With an estimation of average chunk length, we can set a chunking threshold to chunk a sentence.

4 Experimental Setting

4.1 Evaluation

Using the Alignment-Guided Chunking approach described in Section 3, we can train two different chunkers on French–English (FE-Chunker) and German–English (DE-Chunker) bilingual corpora respectively. We use the two chunkers to perform chunking on the same English sentences. Two test sets are constructed, each suitable for the FE-Chunker and the DE-Chunker respectively. The performance of the two chunkers is evaluated on the appropriate test set.

4.2 Gold Standard Test Set

For each sentence E in the test set, there could be N translation references r_1^N. For each sentence pair $< E, r_i >$, a unique word alignment A_i can be acquired. Following the criteria described in Section 3.1, we can derive N chunking results C_1^N using $< E, A_i >$ $(i \in [0, N])$. All these chunking results should be considered to be correct. Chunking results for E using our approach are evaluated on C_1^N using just one 'gold standard' reference.

We firstly construct the test set automatically using the criteria described in Section 3.1. After that we check all the sentences manually to correct all the chunking errors due to word alignment errors.

4.3 Data

The experiments were conducted on French–English and German–English sections of the Europarl corpus (Koehn, 2005) Release V1.[1] This corpus covers April 1996 to December 2001, and we use the Q4/2000 portion of the data (2000-10 to 2000-12) for testing, with the other parts used for training. The English sentences in the French–English and German–English corpora are not exactly the same due to differences in the sentence-alignment process. We obtain the intersection of the English sentences and their correspondences to construct a new French–English corpus and German–English corpus, where these two corpus now share exactly the same English sentences.

In order to test the scalability of our chunking approach, we first use 150k of the sentence pairs for training, which we call the *Small Data* set. Then we use all the sentence pairs (around 300k sentence pairs) for training. We call this the *Large Data* set.

We tag all the English sentences in the training and test sets using a maximum entropy-based Part-of-Speech tagger-MXPOST (Ratnaparkhi, 1996), which was trained on the Penn Treebank (Marcus et al., 1993). We use the GIZA++ implementation of IBM word alignment model 4 (Brown et al., 1993; Och and Ney, 2003)[2] and refinement heuristics described in (Koehn et al., 2003) to derive the final word alignment.

We used the Maximum Entropy toolkit 'maxent',[3] and the Memory-based learning toolkit TiMBL[4] for parameter estimation.

4.4 Statistics on Training Data

To demonstrate the feasibility of adapting our chunking approach to different languages, we obtained some statistics on the chunks of two training sets derived from French–English (F-E, 300k-sentence pairs) and German–English

[1] http://people.csail.mit.edu/koehn/publications/europarl/

[2] More specifically, we performed 5 iterations of Model 1, 5 iterations of HMM, 5 iterations of Model 3, and 5 iterations of Model 4.

[3] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

[4] http://ilk.uvt.nl/timbl/

(D-E, 300k-sentence pairs) corpora respectively. There are 3,316,887 chunks identified in the F-E corpus and 2,915,325 chunks in the D-E corpus. A number of these chunks overlap: 42.08% in the F-E corpus and 47.87% in the D-E corpus (cf. Table 3). The number of overlapping chunks (OL chunks) between these two corpora is 1,395,627.

	F-E	D-E
No. of Chunks	3,316,887	2,915,325
OL Chunks[%]	42.08%	47.87%

Table 3: chunk statistics

We can also estimate the average chunk length on training data. Using the F-E corpus, the average chunk length for English is 1.84 words and 2.10 words using the D-E corpus. This demonstrates definitively that our approach does carve up sentences differently depending on the target language in question.

5 Experimental Results

5.1 Results

Two machine learning techniques—Maximum Entropy (MaxEnt) and Memory-based learning (MBL)—are used for chunking. In order to test the scalability of our chunking model, we carried out experiments on both the Small data and Large data sets described in Section 4.3.

The detailed results are shown in Table 4. Here we can see that the F-score is quite low because we have just one reference in the test set (see Section 4.2). Furthermore, we see no significant improvement with the maximum entropy method when more data is used.

F-scores for German chunks are on the whole between 25 and 33% higher than for French. For German, when using MaxEnt Precision scores are significantly higher than Recall, but the opposite is seen when MBL chunks are used. For French, Recall scores are higher in general than those for Precision.

Figure 4 gives an example of chunking results using MaxEnt. Note the differences between this output and that in Figure 3: the determiner *the* has now been properly grouped with the following N-bar *weight of powerful symbols ...*, and similarly *all* belongs more closely to *three monotheistic religions* than it did before.

The ||| city ||| bears ||| the weight of powerful symbols for ||| all ||| three ||| monotheistic ||| religions ||| .

Figure 4: Example of chunking result

5.2 Multi-level Chunking

As an extension to our classification-based chunking method, multi-level chunking can be regarded as an application of ranking. We obtain the global chunk length from the training data to derive the optimal partition threshold. We use the average chunk length from the training data described in Section 4.4, i.e. for the French–English task, the average English chunk length is 1.84 words, whereas it is 2.10 words for German–English. The results of applying the multi-level chunking method (Multi) are shown in Table 5.

By using the multi-level chunker, we can see a slight increase in recall together with a sharp decrease in precision. This demonstrates that deriving chunks using just a global average chunk length is likely to be sub-optimal for any given sentence.

6 Conclusions and Future Work

In this paper, we have introduced a novel chunking approach guided by the word alignment acquired from bilingual corpora. We investigate our chunking approach on two language pairs: French–English and German–English, where these two bilingual corpora share the same English sentences. Two machine learning techniques—Maximum Entropy and Memory-based learning—were employed to perform chunking. We demonstrate the impact of chunking results on the English side due to the differences between French–English word alignment and German English word alignment, demonstrating the merit of such a chunking approach in a bilingual context. We evaluate the performance of our chunking approach on a one-reference gold standard test set and report an F-score

	Accuracy		Precision		Recall		F-score	
	FR	DE	FR	DE	FR	DE	FR	DE
MaxEnt-Large	55.37	68.41	30.89	47.57	34.57	35.12	32.63	40.41
MBL-Large	52.70	65.75	24.08	38.00	30.43	41.61	26.88	39.72
MaxEnt-Small	55.08	68.37	30.83	47.37	35.26	34.93	32.90	40.21
MBL-Small	52.53	65.56	23.96	37.62	30.41	40.83	26.80	39.16

Table 4: Results of Classification-based Chunking[%]

	French			German		
	Precision	Recall	F-score	Precision	Recall	F-score
MaxEnt	30.89	34.57	32.63	47.57	35.12	40.41
MBL	24.08	30.43	26.88	38.00	41.61	39.72
MaxEnt-Multi	28.41	34.69	31.24	38.14	38.11	38.12
MBL-Multi	22.69	28.18	25.14	34.36	38.46	36.29

Table 5: Classification-based Chunking vs. Ranking-based Chunking[%]

of 32.63% for the *FE-Chunker* and 40.41% for the *DE-Chunker*. We also extend our chunking approach with *Multi-level Chunking*, which is more tolerant of the chunking errors, but lower Precision scores are seen across the board.

As for future work, we want to experiment with other methods of word alignment (e.g. (Tiedemann, 2004; Liang et al., 2006; Ma et al., 2007)) in order to establish which one is most appropriate for our task. We also want to apply this method to other corpora and language pairs, especially using IWSLT data where for 4 language pairs we have 16 reference translations. We anticipate that our chunking approach is likely to be of particular benefit, at least in theory, in a statistical machine translation task given the complexities of the decoding process. Nonetheless, the principal remaining concern is whether the better motivated yet considerably smaller number of bilingual chunks derived via our method will lose out in a real task-oriented evaluation compared to a baseline system seeded with phrase pairs produced in the usual manner.

Acknowledgments

This work is supported by Science Foundation Ireland (grant number OS/IN/1732). We would also like to thank the anonymous reviewers whose insightful comments helped improve this paper.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19**(2):263–311.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computatioanl Linguistics* **22**(1):39–71.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing.* Cambridge University Press, Cambridge, UK.

Nano Gough and Andy Way. 2004. Robust large-scale EBMT with marker-based segmentation. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, Baltimore, MD., pp.95–104.

T. Green. 1979. The necessity of syntax markers: two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18**:481–496.

Declan Groves and Andy Way. 2005. Hybrid example-based SMT: the best of both worlds? In *Proceedings of the workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, ACL 2005*, Ann Arbor, MI., pp.183–190.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pp.79–86, Phuket, Thailand.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pp.48–54, Edmonton, Canada.

Rob Koeling. 2000. Chunking with Maximum Entropy Models. In *In Proceedings of CoNLL-2000*, Lisbon, Portugal, pp.139–141.

Percy Liang, Ben Taskar and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, New York City, NY., pp.104–111.

Feifan Liu, Qianli Jin, Jun Zhao and Bo Xu. 2004. Bilingual chunk alignment based on interactional matching and probabilistic latent semantic indexing. In *Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, Sanya, Hainan Island, China, pp.416–425.

Yanjun Ma. 2006. *Automatic Identification and Alignment of Chinese-English Phrases based on Multi-strategies*. MA thesis, Tsinghua University, Beijing, China.

Yanjun Ma, Nicolas Stroppa and Andy Way. 2007. Bootstrapping Word Alignment Via Word Packing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Czech Republic, pp.304–311.

Michell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* **19**(2):313–330.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* **29**(1):19–51.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the third ACL Workshop on Very Large Corpora*, Somerset, NJ., pp.82–94.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 1996)*, Philadelphia, PA., pp.133–142.

Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the CoNLL-2000 shared task: chunking. 2000. In *Proceedings of CoNLL-2000 and LLL-2000*, Lisbon, Portugal, pp.127–132.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999)*, Bergen, Norway, pp.173–179.

Nicolas Stroppa and Andy Way. 2006. Matrex: DCU Machine Translation System for IWSLT 2006. In *Proceedings of IWSLT 2006 Workshop*, Kyoto, Japan, pp.31–36.

Jörg Tiedemann. 2004. Word to Word Alignment Strategies. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, Geneva, Switzerland, pp.212–218.

Wei Wang, Ming Zhou, Jinxia Huang, and Changning Huang. 2002. Structure alignment using bilingual chunking. In *Proceedings of COLING 2002*, Taipei, Taiwan.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpus. *Computational Linguistics* **23**(3):377–403.