Joint Word Segmentation and Phonetic Category Induction

Micha Elsner
Dept. of Linguistics
The Ohio State University
melsner0@gmail

Stephanie Antetomaso
Dept. of Linguistics
The Ohio State University
antetomaso.2@osu.edu

Naomi H. Feldman
Dept. of Linguistics and UMIACS
University of Maryland
nhf@umd.edu

Abstract

We describe a model which jointly performs word segmentation and induces vowel categories from formant values. Vowel induction performance improves slightly over a baseline model which does not segment; segmentation performance decreases slightly from a baseline using entirely symbolic input. Our high joint performance in this idealized setting implies that problems in unsupervised speech recognition reflect the phonetic variability of real speech sounds in context.

1 Introduction

In learning to speak their native language, a developing infant must acquire two related pieces of information: a set of lexical items (along with the contexts in which they are likely to occur), and a set of phonetic categories. For instance, an English-learning infant must learn that [*i*] and [*ɪ*] are different segments, differentiating between words like *beat* and *bit*, while for a Spanish-learning infant, [*i*] and [*ɪ*]-like tokens represent realizations of the same category. It is clear that these two tasks are intimately related, and that models of language acquisition must solve both together— but how?

This problem has inspired much recent work in low-resource speech recognition (Lee et al., 2015; Lee and Glass, 2012; Jansen and Church, 2011; Varadarajan et al., 2008), with impressive results. Nonetheless, many of these researchers conclude that their systems learn too many phonetic categories, a problem they attribute to the presence of contextual variants (allophones) of the different sounds. For instance, the [*a*] in *dog* is likely longer than the [*a*] in *dock* (Ladefoged and Johnson, 2010), but this difference is not phonologically meaningful in English— it cannot differentiate any pair of words on its own. Many unsupervised systems are claimed to erroneously learn these kinds of differences as categorical ones.

Here, we attempt to model the problem in a more controlled setting by extending work in cognitive modeling of language acquisition. We present a system which jointly acquires vowel categories and lexical items from a mixed symbolic/acoustic representation of the input. As is traditional in cognitive models of vowel acquisition, it uses a single-point formant representation of the vowel acoustics, and is tested on a simulated corpus in which vowel acoustics are unaffected by context. We find that, under these circumstances, vowel categories and lexical items can be learned jointly with relatively little decrease in accuracy from learning either alone. Thus, our results support the hypothesis that the more realistic problem is hard because of contextual variability. As a secondary point, we show that the results reflect problems with local minima in the popular framework of hierarchical Bayesian modeling.

2 Related work

This work aims to induce both a set of phonetic vowel categories and a lexical representation from unlabeled data. It extends the closely related model of Feldman et al. (2013a), which performs the same task, but with known word boundaries; this requirement is a significant limitation on the model's cognitive plausibility. Our model infers a latent word segmentation. Another extension, Frank et al. (2014), uses semantic information to disambiguate words, but still with known word boundaries.

A few models learn a lexicon while categorizing all sounds, instead of just vowels. Lee et al. (2015) and Lee and Glass (2012) use hierarchical Bayesian models to induce word and subword units. These models are mathematically very similar to our own, differing primarily using more complex acoustic representations and inducing categories for all sounds instead of just vowels. Jansen and Church (2011) learns whole-word Markov models, then clusters their states into phone-like units using a spectral algorithm. Their system still learns multiple allophonic categories for most sounds.

In the segmentation literature, several previous systems learn lexical items from variable input (Elsner et al., 2013; Daland and Pierrehumbert, 2011; Rytting et al., 2010; Neubig et al., 2010; Fleck, 2008). However, these models use pre-processed representations of the acoustics (phonetic transcription or posterior probabilities from a phone recognizer) rather than inducing an acoustic category structure directly. Elsner et al. (2013) and Neubig et al. (2010) use Bayesian models and sampling schemes similar to those presented here.

Acquisition models like Elsner et al. (2013),Rytting et al. (2010) and Fleck (2008) are designed to handle *phonological* variability. In particular, they are designed to cope with words which have multiple transcribed pronunciations (*[wan]* and *[want]* for "want"); this kind of alternation can insert or delete whole segments, or change a vowel sound from one perceptual category to another. Such variability is common in spoken English (Pitt et al., 2005) and presents a challenge for speech recognition (McAllaster et al., 1998).

In contrast, the system presented here models *phonetic* variability within a single category. It uses an untranscribed, continuous-valued representation for vowel sounds, so that different tokens within a single category may differ from one another. But it does so within an idealized dataset which lacks phonological variants. Moreover, although the phonetic input to the system is variable, the variation is not predictable; tokens within the category differ at random, independently from their environment.

Several other models also learn phonetic categories from continuous input, either from real or idealized datasets, without learning a lexicon. Varadarajan et al. (2008) learn subword units by incrementally splitting an HMM model of the data to maximize likelihood. Badino et al. (2014) perform k-means clustering on the acoustic representation learned by an autoencoder. Cognitive models using formant values as input are common, many using mixture of Gaussians (Vallabha et al., 2007; de Boer and Kuhl, 2003). Because they lack a lexicon, these models have particular difficulty distinguishing meaningful from allophonic variability.

3 Dataset and model

Our dataset replicates the previous idealized setting for vowel category induction in cognitive modeling, but in a corpus of unsegmented utterances rather than a wordlist. We adapt a standard word segmentation corpus of child-directed speech (Brent, 1999), which consists of 8000 utterances from Bernstein-Ratner (1987), orthographically transcribed and then phonetically transcribed using a pronunciation dictionary.

We add simulated acoustics (without contextual variation) to each vowel in the Brent corpus. Following previous cognitive models of category induction (Feldman et al., 2013b), we use the vowel dataset given by Hillenbrand et al. (1995), which gives formants for English vowels read in the context *h_d*. We estimate a multivariate Gaussian distribution for each vowel, and, whenever a monophthongal vowel occurs in the Brent corpus, we replace it with a pair of formants (f_1, f_2) drawn from the appropriate Gaussian. The ARPABET diphthongs "oy, aw, ay, em, en", and all the consonants, retain their discrete values. The first three words of the dataset, orthographically "you want to", are rendered: *y[380.53 1251.69] w[811.88 1431.96]n t[532.91 1094.14]*.

3.1 Model

Our model merges the Feldman et al. (2013a) vowel category learner with the Elsner et al. (2013) noisy-channel framework for word segmentation, which is in turn based on the segmentation model of Goldwater et al. (2009). In generative terms, it defines a sequential process for sampling a dataset. The observations will be surface strings S, which are divided into (latent) words $X_{i=1:n}$. We denote the j-th character of word i as S_{ij}. When S_{ij} is a vowel, the observed value is a real-valued formant pair (f_1, f_2); when it is a consonant, it is observed directly.

1. Draw a distribution over vowel categories, $\pi_v \sim DP(\alpha_v)$
2. Sample parameters for each category, $\mu_v, \Sigma_v \sim NIW(\mu_0, \Lambda, \nu)$
3. Draw a distribution over word strings, $G_0 \sim DP(\alpha_0, CV(\pi_v, p_c, p_{stop})$
4. Draw bigram transition distributions, $G_x \sim DP(\alpha_1, G_0)$

5. Sample word sequences, $X_i \sim G_{X_{i \ 1}}$
6. Realize each vowel token in the surface string, $S_{ij} \sim Normal(\mu_{X_{ij}}, \Sigma_{X_{ij}})$

The initial prior over word forms, $CV(\pi_v, p_c, p_{stop})$ is the following: sample a word length ≥ 1 from $Geom(p_{stop})$; for each character in the word, choose to sample a consonant with probability p_c or a vowel otherwise; sample all consonants uniformally, and all vowels according to the (possibly-infinite) probability vector π_v.[1] In practice, we integrate out π_v, yielding a Chinese restaurant process in which the distribution over vowels in a new word depend on those used in already-seen words. Vowels which occur in many word types are more likely to recur (Goldwater et al., 2006; Teh et al., 2006).

The hyperparameters for the model are α_0 and α_1 (which control the size of the unigram and bigram vocabularies), α_v (which weakly affects the number of vowel categories), μ_0, n, Λ and ν (which affect the average location and dispersion of vowel categories in formant space), and p_c and p_{stop} (which weakly affect the length and composition of words). We set α_0 and α_1 to their optimal values for word segmentation (3000 and 100 (Goldwater et al., 2009)) and α_v to .001. In practice, no value of α_v we tried would produce a useful number of vowels and so we fix the maximum number of vowels (non-probabilistically) to n_v; we explore a variety of values of this parameter below. The mean vector for the vowel category parameters is set to $[500, 1500]$ and the inverse precision matrix to $500I$, biasing vowel categories to be near the center of the vowel space and have variances on the order of hundreds of hertz. We set the prior degrees of freedom ν to 2.001. Since ν can be interpreted as a pseudocount determining the prior strength, this means the prior influence is relatively weak for reasonably-sized vowel categories. We set $p_c = .5$ and $p_{stop} = .5$; based on Goldwater et al. (2009), we do not expect these parameters to be influential.

These hyperparameter values were mostly taken from previous work. The vowel inverse precision and degrees of freedom differ from those in Feldman et al. (2013a), since our approach requires us to sample from the prior, but the uninformative prior used there was too poor a fit for the data. We chose a variance with units on the order of the overall data variance, but did not tune it.

3.2 Inference

We conduct inference by Gibbs sampling, including three sampling moves: block sampling of the analyses of a single utterance, table label relabeling of a lexical item (Johnson and Goldwater, 2009) and resampling of the vowel category parameters μ_v and Σ_v. We run 1000 iterations of utterance resampling, with table relabeling every 10 iterations.[2] Following previous work, we integrate out the mixing weight distributions G_0, G_1 and π_v, resulting in Chinese restaurant process distributions for unigrams, bigrams and vowel categories in the lexicon (Teh et al., 2006). Unlike Feldman et al. (2013a) and many other variants of the Infinite Mixture of Gaussians (Rasmussen, 1999), we do not integrate out μ_v and Σ_v, since this would create long-distance dependencies between different tokens of the same vowel category within an utterance and thus complicate the implementation of a whole-utterance block sampling scheme.

To block sample the analyses of a single utterance, we use beam sampling (Van Gael et al., 2008; Huggins and Wood, 2014), an auxiliary-variable sampling scheme in which we encode the model as an (infeasibly large) finite-state transducer, then sample cutoff variables which restrict our algorithm to a finite subset of the transducer and sample a trajectory within it. We then use a Metropolis-Hastings acceptance test to correct for the discrepancy between our finite-state encoding and the actual model probability caused by repetitions of a lexical item within the same utterance.

Specifically, for each vowel s_{ij}, we sample a cutoff $c_{ij} \sim U[0, P(s_{ij}|X_{ij})]$. This cutoff indicates the least probable category assignment we will permit for the surface symbol s_{ij}. This cutoff constrains us to consider only a finite number of vowels at each point; if there are not enough, we can instantiate unseen vowels by sampling their μ and Σ from the prior. We then construct the lattice of possible word segmentations in which s_{ij} is allowed to correspond to any vowel in any lexical entry, as long as all the consonants match up and the vowel assignment density $P(s_{ij}|x_{ij})$ is greater than the cutoff. We then propose a new trajectory by sampling from this lattice. See Mochihashi et al.

[1] Feldman et al. (2013a) assumes a more complex distribution over consonants, while Goldwater et al. (2009) assumes uniformity over all sounds.

[2] Annealing is applied linearly, with inverse temperature scaling from .1 to 1 for 800 iterations, then linearly from 1.0 to 2.0 to encourage a MAP solution. The Gaussian densities for acoustic token emissions are annealed to inverse temperature .3, to keep them comparable to the LM probabilities (Bahl et al., 1980).

(2009) for details of the finite-state construction.

As in Feldman et al. (2013a), we use a table relabeling move (Johnson and Goldwater, 2009) which changes the word type for a single table in the unigram Chinese restaurant process by changing one of the vowels. This recategorizes a large number of tokens which share the same type (though not necessarily all, since there may be multiple unigram tables for the same word type). The implementation is tricky because of the bigram dependencies between adjacent words, some of which may be tokens of the same lexical item. Nonetheless, this move is necessary because token-level sampling has insufficient mobility to change the representation of a whole word type: if the sampler has incorrectly assigned many tokens to the non-word *hav*, moving any single token to the correct *hæv* will raise the transducer probability but also catastrophically lower the lexical probability by creating a singleton lexical item.

Finally, because μ_v and Σ_v are explicitly represented rather than integrated out, their values must be resampled given the set of formant values associated with each vowel cluster. The use of a conjugate (Normal-Inverse Wishart) prior makes this simple, applying equations 250-254 in Murphy (2007).

4 Results

Despite using multiple block moves, mobility is a severe issue for the sampler; the inference procedure fails to merge together redundant vowel categories even when doing so would raise the posterior probability significantly. We demonstrate this by running the sampler with various numbers of vowel categories n_v. Posterior probabilities peak around the true value of 12, but models with extra categories always use the entire set.

With n_v set to 11 or 12 categories, quantitative performance is relatively good, although segmentation is not as good as the Goldwater et al. (2009) segmenter without any acoustics. In fact, the system slightly outperforms the Feldman et al. (2013a) lexical-distributional model with gold-standard segmentation. Results are shown in Table 1.

Word tokens are correctly segmented (both boundaries correct) with an F-score of 67%[3] (versus 74% in (Goldwater et al., 2009). Individual boundaries are detected with an F-score of 82% versus 87%. We also evaluate the lexical items, checking whether words are correctly grouped as well as segmented (for example, whether tokens of "is" and "as" are separated). Feldman et al. (2013a) evaluates the lexicon by computing a pairwise F-score on tokens (positive class: clustered together). Under this metric, their highest lexicon score for English words is 93%. We compute this metric on the subset of words for which the segmentation system performs correctly (it is not clear how to count "misses" and "false alarms" for tokens which were mis-segmented). On this subset, this metric scores our system with $n_v = 12$ at 91%, which indicates that we correctly identify most of the correctly segmented items.

System	Seg P	R	F	Vow P	R	F
Goldwater	76	72	74	-	-	-
Feldman	-	-	-	-	-	76
joint, n_v=12	64	69	67	87	80	83
joint, n_v=11	65	70	67	85	84	85

Table 1: Segmentation and vowel clustering scores.

We evaluate our phonetic clustering by computing the same pairwise F-score on pairs of vowel tokens. Our score is 83%; the Feldman et al. (2013a) model scores 76%. We conjecture that the improvement results from the use of bigram context information to disambiguate between homophones. Confusion between vowels (attached as supplemental material) is mostly reasonable. We find cross-clusters for *ah,ao*, *ey,ih*, and *uh,uw*. The model's successful learning of the vowel categories demonstrates that the high performance of cognitive models in this domain is not due solely to their access to gold-standard word boundaries (see also Martin et al. (2013)). We believe that the idealized acoustic values (sampled from stationary Gaussians reflecting laboratory production) are critical in allowing these models to outperform those which use natural speech.

Though solving the two tasks together is harder than tackling either alone, these results nonetheless demonstrate comparable performance to other models which have to cope with variability while segmenting. Fleck (2008) reports only 44% segmentation scores on transcribed English text including phonological variability; the noisy channel model of Elsner et al. (2013) yields a segmentation token score of 67%.[4]

Besides generic task difficulty, we attribute the

[3] The joint model scores are averaged over two sampler runs.

[4] Word segmentation scores from Lee et al. (2015), learning directly on acoustics, range between 16 and 20.

low scores to the model's inability to mix, which prevents it from merging similar vowel classes. Because table relabeling does not merge tables in the CRP hierarchy, even if it replaces an uncommon word with a more common one, the configurational probability does not change. Thus the model's sparsity preference cannot encourage such moves. The prior on vowel categories, $DP(p_v)$, does encourage changes which reduce the number of lexical types using a rare vowel, but relabeling a table can rearrange at most a single sample from this prior distribution and is easily outweighed by the likelihood.

A hand analysis of one sampler run in which */ɪ/* was split into two categories showed clear mixing problems. Many common words, such as "it" and "this", appeared as duplicate lexical entries (e.g. *[ɪ$_1$t]* and *[ɪ$_2$t]*). These presumably captured some chance variation within the category, but not an actual linguistic feature.

We suspect that this mobility problem is also a likely issue with models like Lee and Glass (2012) which use deep Bayesian hierarchies and relatively local inference moves. Since the problem occurs even in this idealized setting, we expect it to exacerbate the problems caused by contextual variability in more realistic experiments.

Some errors did result from the joint nature of the task itself. We looked for reanalyses involving both a mis-segmentation and a vowel category mistake. For instance, the model is capable of misanalyzing the word "milk" as "me" followed by the phonotactically implausible sequence "lk". Mistakes like these, in which the misanalysis creates a word, are relatively rare as a proportion of the total. The most common words created are "say", "and", "shoe", "it" and "a". More commonly, misanalyses of this type segment out single vowels or nonwords like *[luk]*, *[eŋ]*, and *[mɔ]*. Some such errors could be corrected by incorporating phonotactics into the model (Johnson and Goldwater, 2009). In general, the error patterns are neither particularly interpretable nor cognitively very plausible. This stands in contrast to the effects on word boundary detection found in a model of phonological variation (Elsner et al., 2013).

5 Conclusion

The main result of our work is that joint word segmentation and vowel clustering is possible, with relatively high effectiveness, by merging models known to be successful in each setting independently. The finding that success of this kind is possible in an idealized setting reinforces an argument made in previous work: that much of the difficulty in category acquisition is due to contextual variation.

Both phonological and phonetic variability probably contribute to the difficulty of the real task. Phonological processes such as reduction create variant versions of words, splitting real lexical items and creating misleading minimal pairs. Phonetic processes like coarticulation and compensatory lengthening create predictible variation within a category, encouraging the model to split the category into allophones. In future work, we hope to quantify the contributions of these sources of error and work to address them explicitly within the same model.

Acknowledgements

This research was funded by NSF grants 1422987 and 1421695. We are grateful for the advice of three anonymous reviewers, and to Sharon Goldwater for distributing the baseline DPSeg system.

References

Leonardo Badino, Claudia Canevari, Luciano Fadiga, and Giorgio Metta. 2014. An auto-encoder based approach to unsupervised learning of subword units. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7634–7638. IEEE.

Lalit Bahl, Raimo Bakis, Frederick Jelinek, and Robert Mercer. 1980. Language-model/acoustic-channel-model balance mechanism. Technical disclosure bulletin Vol. 23, No. 7b, IBM, December.

Nan Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6. Erlbaum, Hillsdale, NJ.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, February.

Robert Daland and Janet B. Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Bart de Boer and Patricia Kuhl. 2003. Investigating the role of infant-directed speech with a computer model. *Acoustic Research Letters On-Line*, 4:129–134.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 42–54, Seattle, Washington, USA, October. Association for Computational Linguistics.

Naomi H. Feldman, Thomas L. Griffiths, Sharon Goldwater, and James L. Morgan. 2013a. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 4:751–778.

Naomi H. Feldman, Emily B. Myers, Katherine S. White, Thomas L. Griffiths, and James L. Morgan. 2013b. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio, June. Association for Computational Linguistics.

Stella Frank, Naomi Feldman, and Sharon Goldwater. 2014. Weak semantic context helps phonetic learning in a model of infant language acquisition. In *ACL (1)*, pages 1073–1083.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney, Australia, July. Association for Computational Linguistics.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, 97:3099.

Jonathan Huggins and Frank Wood. 2014. Infinite structured hidden semi-Markov models. *arXiv preprint arXiv:1407.0044*, June.

Aren Jansen and Kenneth Church. 2011. Towards unsupervised training of speaker independent acoustic models. In *INTERSPEECH*, pages 1693–1692.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado.

Peter Ladefoged and Keith Johnson. 2010. *A course in phonetics*. Wadsworth Publishing.

Chia-ying Lee and James Glass. 2012. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40–49, Jeju Island, Korea, July. Association for Computational Linguistics.

Chia-ying Lee, Timothy J O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning phonemes with a protolexicon. *Cognitive Science*, 37:103–124.

Don McAllaster, Lawrence Gillick, Francesco Scattone, and Michael Newman. 1998. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *ICSLP*.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore, August. Association for Computational Linguistics.

Kevin Murphy. 2007. Conjugate Bayesian analysis of the gaussian distribution. Technical report, University of British Columbia.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *11th Annual Conference of the International Speech Communication Association (InterSpeech 2010)*, pages 1053–1056, Makuhari, Japan, 9.

Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Carl Edward Rasmussen. 1999. The infinite Gaussian mixture model. In *NIPS*, volume 12, pages 554–560.

Anton Rytting, Chris Brew, and Eric Fosler-Lussier. 2010. Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, 37(3):513–543.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Gautam K. Vallabha, James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33):13273–13278.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. Beam sampling for the infinite Hidden Markov model. In *Proceedings of the 25th International Conference on Machine learning*, ICML '08, pages 1088–1095, New York, NY, USA. ACM.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of the Association for Computational Linguistics: Short Papers*, pages 165–168.