Word Segmentation as Unsupervised Constituency Parsing

Raquel G. Alhama
Tilburg University,
Warandelaan 2, 5037AB Tilburg,
the Netherlands
rgalhama@uvt.nl

Abstract

Word identification from continuous input is typically viewed as a segmentation task. Experiments with human adults suggest that familiarity with syntactic structures in their native language also influences word identification in artificial languages; however, the relation between syntactic processing and word identification is yet unclear. This work takes one step forward by exploring a radically different approach of word identification, in which segmentation of a continuous input is viewed as a process isomorphic to unsupervised constituency parsing. Besides formalizing the approach, this study reports simulations of human experiments with DIORA (Drozdov et al., 2019), a neural unsupervised constituency parser. Results show that this model can reproduce human behavior in word identification experiments, suggesting that this is a viable approach to study word identification and its relation to syntactic processing.

1 Introduction

When exposed to speech in an unknown language, humans are faced with the task of finding out what are the basic combinatorial units of the language, such as phonemes, syllables, words and phrases. Since speech is continuous, humans need to rely on implicit cues –such as statistical information– to find out the building blocks of the language. One approach that studies which statistical cues can be used by humans in this task is Artificial Grammar Learning (AGL). Experiments in AGL are characterized by the use of artificial languages with carefully controlled statistical properties. To investigate word identification with this paradigm, participants in a typical AGL experiment are first exposed to a speech-like sample of the artificial language (usually recorded with synthetic voice). Then, they participate in a test that has been designed to show whether participants identified the words in the artificial language.

To formalize theories of how humans identify words in AGL tasks, a range of computational models have been proposed over the last two decades. These models have explained a wide arrange of phenomena, using a variety of algorithms such as Bayesian inference (Frank et al., 2010), normative statistics (Swingley, 2005), cognitively inspired processes implementing recognition or memorization (Alhama and Zuidema, 2017; Perruchet and Vinter, 1998), and neural networks (French et al., 2011; Endress and Johnson, 2021).

There is, however, one phenomenon that has not been addressed in the computational literature in AGL: the fact that participant's knowledge of their native language influences performance in this type of AGL experiments. In particular, results seem to be influenced by co-occurrence statistics of sublexical units (Onnis et al., 2005; Siegelman et al., 2018; Elazar et al., 2022), and interestingly, also by the presence of left- or right-branching syntactic structures in the native language, which predict the statistics that subjects use to identify words (Onnis and Thiessen, 2013).

One likely reason why this has not been the focus of prior models of word identification in AGL is that we are in need a computational framework that can represent this information on the first place. While sensitivity to co-occurrences of sublexical patterns could potentially be accounted for with at least some of the existing models (in particular the neural network approaches, which should show similar output to input with similar representations), the influence of prior syntactic knowledge cannot be readily explained with the existing approaches, as none of these models incorporate syntactic processing.

Thus, a preliminary step before modelling the influence of prior knowledge is to develop a modelling framework that can relate word identification in AGL to syntactic processing in the first place[1].

[1] In the field of word identification from naturalistic input,

This work aims to fill this gap by presenting a radically different account of word identification that is isomorphic to syntactic processing: namely, word segmentation as *unsupervised constituency parsing*.

This paper is structured as follows. Section 2 reviews the experimental record that this work focuses on. The approach of modelling word segmentation as unsupervised constituency parsing is formalized in section 3. Next, section 4 reports an empirical study using DIORA (Drozdov et al., 2019), an unsupervised neural inside-outside constituency parser. The results, reported in section 5, show that this approach can be effectively used to model human word identification in AGL experiments with human adults. Finally, implications of this new perspective on word identification are discussed in section 6, and directions for future studies are proposed in section 7.

2 Experimental Record

A long tradition of AGL experiments have used artificial languages to discover how humans identify words from a continuous speech-like stream. Studies show that humans can segment words based on statistics over syllables, such as frequency of co-occurence (Aslin et al., 1998), transitional probabilities (Saffran et al., 1996a,b; Perruchet and Desaulty, 2008) predictive dependencies between non-adjacent syllables (Peña et al., 2002; Endress and Bonatti, 2007; Frost and Monaghan, 2016), or phonotactic patterns (Onnis et al., 2005).

Here, the focus is on the two experiments reported in Perruchet and Desaulty (2008) (P&D onwards). These experiments showed that humans have the ability to keep track of both forward and backward transitional probabilities (as explained next) and use them for identifying words. It is precisely this ability that is susceptible of being influenced by prior syntactic knowledge (Onnis and Thiessen, 2013), motivating the choice to focus on these experiments as a starting point.

In Experiment 1, the authors used an artificial language consisting of 9 bi-syllabic 'words', formed with combinations of 12 different syllables. There were two conditions in the experiment: *forward* and *backward*. In the forward condition, the first syllable of each word uniquely predicted the second syllable (e.g. if A and B were syllables and AB was word, then A was only followed by B). In other words, the forward TP (TP_{fw}) *within* words was consistently 1, while it was much lower *between* words:

$$TP_{fw}(AB) = p(B|A) = \begin{cases} 1 \; if AB \in \{words\} \\ 0.11 \; otherwise \end{cases}$$

The *backward* condition follows exactly the same design, except that it is the second syllable in the word which uniquely predicts the first:

$$TP_{bw}(AB) = p(A|B) = \begin{cases} 1 \; if AB \in \{words\} \\ 0.11 \; otherwise \end{cases}$$

The participants were familiarized with a sample of synthesized speech of this language, consisting of a random concatenation of 115 repetitions of each word. With this design, the co-occurrence frequency of syllables within a word was 3 times larger than for syllables spanning word boundaries. The total duration of the recorded stream was 8 minutes, and there were no pauses or any other acoustic indication that separated the words. Thus, the only two cues that participants could use to identify words were the TPs between syllables and the co-occurrence frequency of syllables (as it was 3 times higher for syllables within words than for syllables spanning word boundaries).

Condition	**Words**
Forward	AX, BX, CX, DY, EY, FY, GZ, HZ, IZ
Backward	XA, XB, XC, YD, YE, YF, ZG, ZH, ZI

Table 1: Words in the artificial languages used in Experiments 1 and 2 in Perruchet and Desaulty (2008). The symbols A, ..., Z were arbitrarily mapped to syllables.

After listening to this stream of artificial words, the participants were presented with a 2-Alternative Forced Choice (2AFC) test. Each trial in the test consisted of a choice between a word of the language, and a 'partword', i.e. a sequence of two syllables that spanned across word boundaries. For instance, in the forward condition, a test trial could involve the word CX and the partword XD (see

models that allow for some level of hierarchical representations have been proposed (De Marcken, 1995; Johnson and Goldwater, 2009; Lignos, 2012) but have not been evaluated for unsupervised parsing at the syntactic level.

Table 1). Participants were instructed to choose the item that seemed more like a word of the artificial language. In both conditions, participants chose words more frequently than partwords (with a slight advantage for the *backward* condition). This finding suggests that words can be identified based on statistical properties such as syllable co-occurrence frequency and TPs, in either directions.

To disentangle the contribution of each cue, in a second experiment, the authors designed an artificial language in which the frequency of words and partwords in the familiarization stream was controlled. Thus, the only way to identify words was to keep track of TPs. Results of Experiment 2 showed that participants were statistically above chance in both conditions, with a slight advantage for the forward condition (although the difference between directions did not reach significance). The authors concluded that human adults can track TPs in both directions, and use them to identify words in a continuous stream.

3 Formalization of the Approach

The approach presented this paper is to model the task of word identification from a continuous input using the same process for discovering syntactic constituents. A number of adaptations and considerations are required, as described next.

3.1 Word Segmentation as Unsupervised Constituency Parsing

Constituency parsing is the task of identifying which word spans form constituents, and how are those constituents are hierarchically combined into larger constituents to form the correct syntactic tree. The nodes that occupy the lowest positions in the tree (considering that the root is the highest node) correspond to the 'tightest' constituents, i.e. those that span over words that form *cohesive phrases* that can be further combined (Onnis and Thiessen, 2013). As an example, given the sentence *the singer yelled*, a constituency parser needs to decide whether a grouping like *((the, singer), yelled)* is more likely than *(the, (singer, yelled))*. A successful parser would conclude that *(the, singer)* forms a cohesive constituent (concretely, a noun phrase), while *(singer, yelled)* does not.

More generally, given a sentence $S = ABC$ where A, B and C are basic units (in this case, words), the parser needs to decide whether to group together AB or BC to form a higher-order unit (a constituent). Likewise, a segmentation algorithm presented with a stream $S = ABC$, where A,B and C are basic units (e.g. syllables or phonemes), also needs to decide whether the most cohesive higher-order unit (in this case, a word) is AB or BC[2]. Thus, with this simile, word segmentation can be cast in terms of a process that is isomorphic to (unsupervised) constituency parsing.

3.2 Input

Participants in the experiments by P&D were exposed to a speech stream formed with a randomized concatenation of the bisyllabic words in the artificial language. Similarly, to train a parsing model, a stream of 'syllables' (which is coded simply using the same symbols as P&D, i.e. A-D, X-Z) is generated with the same procedure described in the original paper. Thus, these symbols are the basic units (or vocabulary) for the parser.

As in most AGL experiments, the stimuli in P&D consisted of one single stream, which was not separated into different sentences. However, the training data used for parsing typically consists of a large number of sentences, likely much shorter than the stimuli in AGL experiments. Moreover, the adults participating in the experiment are presumably not deriving one single parse during the 8 minute exposure to the artificial language, as this input greatly exceeds the average sentence length of natural language. More likely, humans separately processed subsequences of the stimuli, as would be expected given limited attention span and short term memory. This intuition is captured in some models of segmentation in AGL, which operate over subsequences of random length (Perruchet and Vinter, 1998), or an all possible subsequences up to a predefined maximum length (Alhama and Zuidema, 2017).

Similarly, the approach proposed here is to divide the stream into subsequences ('sentences'), the length of which is determined with a stochastic procedure. Unlike previous models, this approach samples the length of the subsequences from a Poisson distribution, with parameters derived from spoken natural language: the mean and standard deviation of the distribution were computed from the monolingual French corpus in OpenSubtitles

[2] In practice, ABC could also form a word. For simplicity and consistency with the stimuli in P&D, this work focuses exclusively on bisyllabic words. However, the approach can be extended to words of any length.

(Lison and Tiedemann, 2016)[3]. The corpus consisted of over 100 million sentences, and the mean sentence length was 5.93 (with standard deviation of 4.55). A constraint is set such that the minimum sentence length is 4, and the maximum is 10. This prevents too much fragmentation of the input and keeps the distribution centered around the peak. Figure 1 shows the distribution of the subsequences derived from the stimuli.

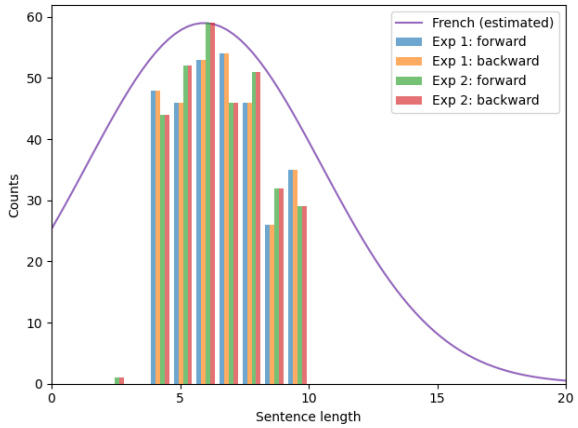

Figure 1: Distribution of subsequence ('sentence') length after partitioning the input stream, for each experimental condition. The continuous line corresponds to the estimated Poisson distribution from the mean and standard deviation in OpenSubtitles (for French).

It must be noted that, by breaking the stream into subsequences, boundaries are introduced in an otherwise continuous stream, and it is therefore imperative that these are not consistently aligned with word boundaries, as otherwise this would provide additional information to the model (which was not available to participants in the experiments). By using a stochastic procedure, the boundaries are not consistently set either within or between words, and thus no artificial cue is introduced.

3.3 Evaluation

In the experiments reported in P&D, participants responded to a 2AFC test that paired words with partwords, i.e. sequences of syllables that spanned word boundaries. A preference for words at group level was taken as indication of having successfully identified the words of the artificial language.

From a modelling perspective, what is required to implement the 2AFC choices is some 'score' that conditions the choice for for words vs. partwords. Previous models of segmentation derived scores based on internal counts of the model, i.e. the amount of times that a sequence was encountered (Frank et al., 2010) or memorized (Perruchet and Vinter, 1998; Alhama and Zuidema, 2017); or alternatively, based on the reconstruction error of these items in an autoencoder (French et al., 2011).

In this work, a different approach is required, since scores need to be derived from the predicted parse trees. The proposal presented here is to assign a score to each test item (word or partword) based on to what extent the parser identifies this syllable sequence as a cohesive constituent. Given that all the tested items are bisyllabic, the most straightforward approach is to quantify cohesiveness as the number of times a word or partword has been placed at the lowest level of the trees predicted from the familiarization stimuli (or, in other words, the amount of times that the syllables in a word or partword are siblings; see table 2 for an example). This computation can easily be extended to longer items by considering additional higher nodes in the tree.

	Score	
Tree	**Words**	**Partwords**
$[[A[XE]][YB]]$	AX: 0 EY: 0	XE: 1 YB: 1
$[[[AX][EY]]B]$	AX: 1 EY: 1	XE: 0 YB: 0
$[[AX][E[YB]]]$	AX: 1 EY: 0	XE: 0 YB: 1

Table 2: Example of different parse trees for the sentence AXEYB, and the scores for test items (words: AX, EY; partwords: XE, YB).

Then, for each item pair in the test, the item that has the largest score is chosen (or randomly determined in the unlikely case of a tie). Finally, as in the original experiments, the accuracy is the mean number of choices for words over the total number of test items.

4 Simulations

This section presents simulations with Deep Inside-Outside Recursive Autoencoder (DIORA, Drozdov et al., 2019), an unsupervised neural constituency parser. DIORA is an autoencoder network, trained with a fill-in-the-blank objective: it encodes all the words in a sentence except one in a single vector, and then decodes from this vector, predicting all the words (including the removed one). The encoder uses a chart to build a constituency tree, with each cell consisting of a weighted average all the possible subtrees covering the represented span.

[3]French was the native language of the participants in the experiments of P&D

These subtrees are encoded as independent vectors with their corresponding score, both of which are computed recursively using a composition function. In a recent empirical comparison, DIORA exhibited some of the best results in unsupervised constituency parsing for English, and outperformed all the competing models in most of the experiments in Japanese (Li et al., 2020).

To reproduce the original experiments in P&D, I trained DIORA with the input data generated according to the procedure described in 3.2[4]. DIORA can be used with different composition functions: a multilayer feed-forward network (MLP), a version of the MLP that shares the inside and outside parameters (MP$_{shared}$), and a TreeLSTM (Tai et al., 2015). The model can be optimized with either Max-Margin or Cross-Entropy loss (Softmax). Simulations are reported with all these variants, with the rest of hyperparameters fixed to the default values, except: batch size=20, hidden layer size= 16, maximum epochs=50 [5]).

I trained 30 individual models for each configuration and experimental condition. This is roughly the larger number of participants in the experimental conditions in P&D (n=31), and the models only differed in their initial state. At the end of training, the models were presented with the stimuli one more time, to produce the final parse trees that would be used for evaluation. The evaluation metric described in 3.3 was computed for each model, and —as in the original paper— the mean performance of the 30 models is submitted to a one-sided Student's t-test to find whether the performance is significantly above chance level[6].

5 Results

5.1 Experiment 1

The first experiment reported in Perruchet and Desaulty (2008) used an artificial language in which words could be identified based on the TPs between syllables (either in the forward or the backward direction, depending on the condition). Table 3 reports the mean performance (i.e. mean number of correct choices in the 2AFC test), and the statistical significance when comparing against chance level.

Comp.	Loss	Cnd.	Acc. (SE)
TreeLSTM	margin	fw	0.77(0.02)***
		bw	0.76(0.03)***
	softmax	fw	0.78(0.03)***
		bw	0.78(0.03)***
MLP	margin	fw	0.77(0.03)***
		bw	0.75(0.03)***
	softmax	fw	0.74(0.03)***
		bw	0.65(0.03)***
MLP$_{shared}$	margin	fw	0.75(0.03)***
		bw	0.74(0.03)***
	softmax	fw	0.77(0.02)***
		bw	0.76(0.02)***
Humans		fw	0.60(0.51)
		bw	0.67(0.56)**

Table 3: Performance of the model variants on Experiment 1, for the forward (fw) and backward (bw) conditions. The accuracy (standard error) is averaged over 30 models that differ in initialization. Asterisks indicate statistical significance over chance performance (***:p<0.001;**:p<0.01;*:p<0.05).

As can be seen, all the model variants are successful in distinguishing words from partwords. The mean accuracies of all the models are statistically above chance, and do not differ greatly in terms of model choices (with TreeLSTM-softmax having the best performance). Thus, word identification in this condition can be achieved with DIORA, slightly outperforming humans.

5.2 Experiment 2

The second experiment used an artificial language with controlled frequency, such that words and partwords would not differ on this regard. The results of simulations with this stimuli are reported in table 4. The pattern of results is notably different from Experiment 1: only the model with Tree-LSTM combined with Max-Margin reconstruction loss is successful in this task (with the exception of MLP-shared for the backward condition). Thus, this variant of DIORA, which was also successful in identifying words in Experiment 1, successfully reproduces the observed behavior of human adults, and is capable of identifying words in continuous input based solely on the transitional probabilities between syllables, regardless of whether these are more reliable in the forward or the backward direction.

[4] I used a fork of the original model, with a small adjustment to the code that prevented the model from loading pre-trained embeddings (https://github.com/rgalhama/diora)

[5] Performance with the default hyperparameters DIORA was low for the reported experiments, possibly due to the very reduced amount of data of the current experiment.

[6] The code used for these simulations is available at https://github.com/rgalhama/segmentation_as_unsup_parsing

Comp.	Loss	Cnd.	Acc. (SE)
TreeLSTM	margin	fw	0.60(0.04)*
		bw	0.58(0.03)*
	softmax	fw	0.55(0.04)
		bw	0.53(0.03)
MLP	margin	fw	0.50(0.04)
		bw	0.52(0.04)
	softmax	fw	0.52(0.04)
		bw	0.55(0.04)
MLP_{shared}	margin	fw	0.55(0.04)
		bw	0.58(0.05)
	softmax	fw	0.51(0.04)
		bw	0.58(0.04)*
Humans		fw	0.66(0.43)***
		bw	0.61(0.51)*

Table 4: Performance of the model variants on Experiment 2, for the forward (fw) and backward (bw) conditions. The accuracy (standard error) is averaged over 30 models that differ in initialization. Asterisks indicate statistical significance over chance performance (***:p<0.001;**:p<0.01;*:p<0.05).

However, the fact that the accuracy dropped for the other model variants is intriguing. Since the evaluated performance is the mean over 30 simulations, there could be at least two reasons behind the tendency to perform at chance. One would be that most of these simulations do perform individually at chance, and are simply not well suited for distinguishing between words and partwords based on TPs. Alternatively, the mean may be around chance due to a similar number of well-performing and failing models, as would be the case if the initial state was highly influential on the final performance of the individual models. The greater variance found in this experiment (compared to experiment 1) suggests that this may be the case. To find out more, the distribution of scores is graphically reported in Figure 2.

As can be seen, the distributions are much tighter for Experiment 1, and the spread of the scores in Experiment 2 cover almost the entire range of scores, suggesting that, as suspected, the initial state is highly influential on performance.

5.3 Subjective Frequencies

The experimental design in P&D involves the use of a 2AFC test to discover whether the words in the speech sample have been discovered. However, the extent to which 2AFC tests reflect the discovery of words has been put into question before (Alhama et al., 2015; Kidd et al., 2020). In particular, success in 2AFC can happen even when words are not that clearly distinguished from partwords. Thus, to gain further insight on the status of words, Fig. 3 shows the amount of times that the best-performing DIORA model (TreeLSTM-margin) –which was successful in the 2AFC test– recognized each test item as a constituent. This quantity is known as the 'subjective' frequencies of the model (Alhama and Zuidema, 2016).

As can be seen, the frequencies for words in Experiment 1 are much higher than those of partwords. A Student's t-test confirms that counts for words are statistically different from partwords (backward: $[t(30) = 14.54, p = 1.19e^{-40}]$, forward: $[t(30) = 13.40, p = 1.52e^{-35}]$). However, in Experiment 2, the difference between words and partwords is less obvious, and a few partwords are identified more often than some of the words. The slight superiority of words was enough for this model to be successful in the 2AFC test. A Student's t-test over counts of words vs. partwords does not yield evidence of significant differences (backward:$[t(30) = 1.62, p = 0.10]$, forward:$[t(30) = 1.96, p = 0.05]$). Together, these results suggest that the 2AFC test reveals only a slight superiority of words over partwords.

6 Discussion

From a computational perspective, word identification from continuous (artificial) input has always been portrayed as a *segmentation* task, concerned with breaking the continuous stream into combinatorial pieces. This work explores a completely different perspective, in which the identification of words is carried out with a syntactic constituency parser, which groups the syllables hierarchically into tree structures.

The results for experiments 1 and 2 show that a model like DIORA (with TreeLSTM and Max-Margin loss) can successfully reproduce human behavior in the experiments. From a mechanistic perspective, a tentative conclusion is that, when exposed to speech-like input in an unknown language, human adults group syllables that follow statistically coherent patterns, and this grouping is hierarchical –akin to the hierarchical structures attributed to syntax.

How, then, does the process of identifying words relate to finding the syntactic relations between the identified words? Given the hierarchical nature of

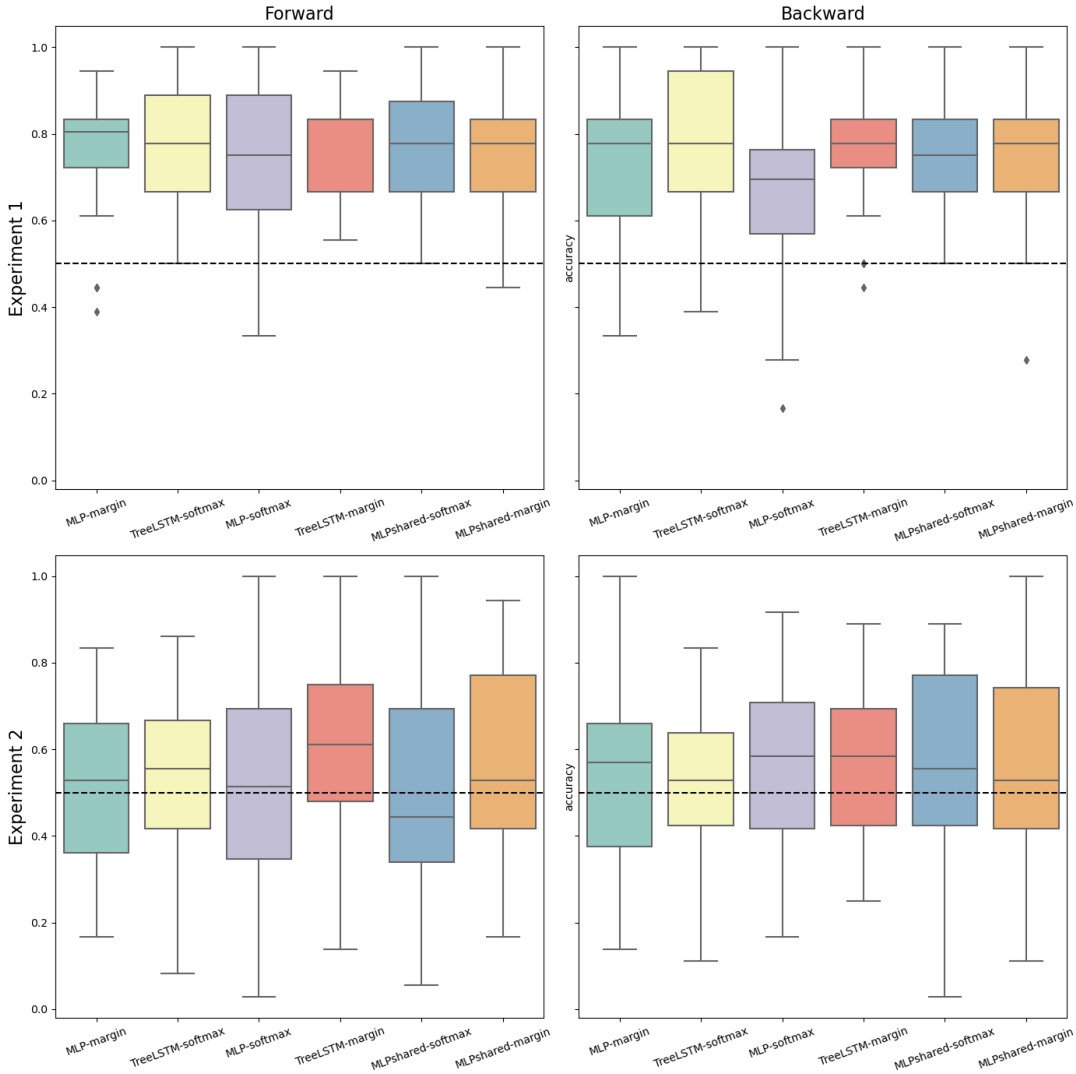

Figure 2: Distribution of accuracies for DIORA model variants on Experiments 1 and 2, for forward and backward TPs.

the process, a possibility is that one single process builds a bottom-up hierarchy of units, grouping subword sequences into words and combining those into syntactic constituents. This is consistent with some usage-based theories of language (Kay and Fillmore, 1999; Goldberg, 2006, p.5), which deem all levels of grammatical analyses as homologous. This interpretation would explain the results in Onnis and Thiessen (2013), which show that humans identify words consistent with TPs in the forward or backward direction, depending on grammatical patterns in the native language (in particular, the tendency for head-directionality).

Although DIORA reproduced, to a great extent, the pattern of results reported in P&D, there are some differences. To begin with, DIORA is better than humans in identifying words when those are more frequent than partwords. This is evidenced by the performance in Experiment 1, as well as by the distribution of frequency counts reported in section 5.3. On the other hand, only one of the variants of DIORA identified words in Experiment 2, when frequency information was removed. As shown above, there is large variance in the performance of the models, depending on their initial state. This is again consistent with the results observed in Onnis and Thiessen (2013): in the absence of frequency information, humans seem to rely on prior knowledge to guide the discovery of words. Nevertheless, to confirm whether the current results speak to the observed behavior in Onnis and Thiessen (2013), simulations using the same stimuli are required. Thus, a prediction from this work is that pre-training the parser with Korean or

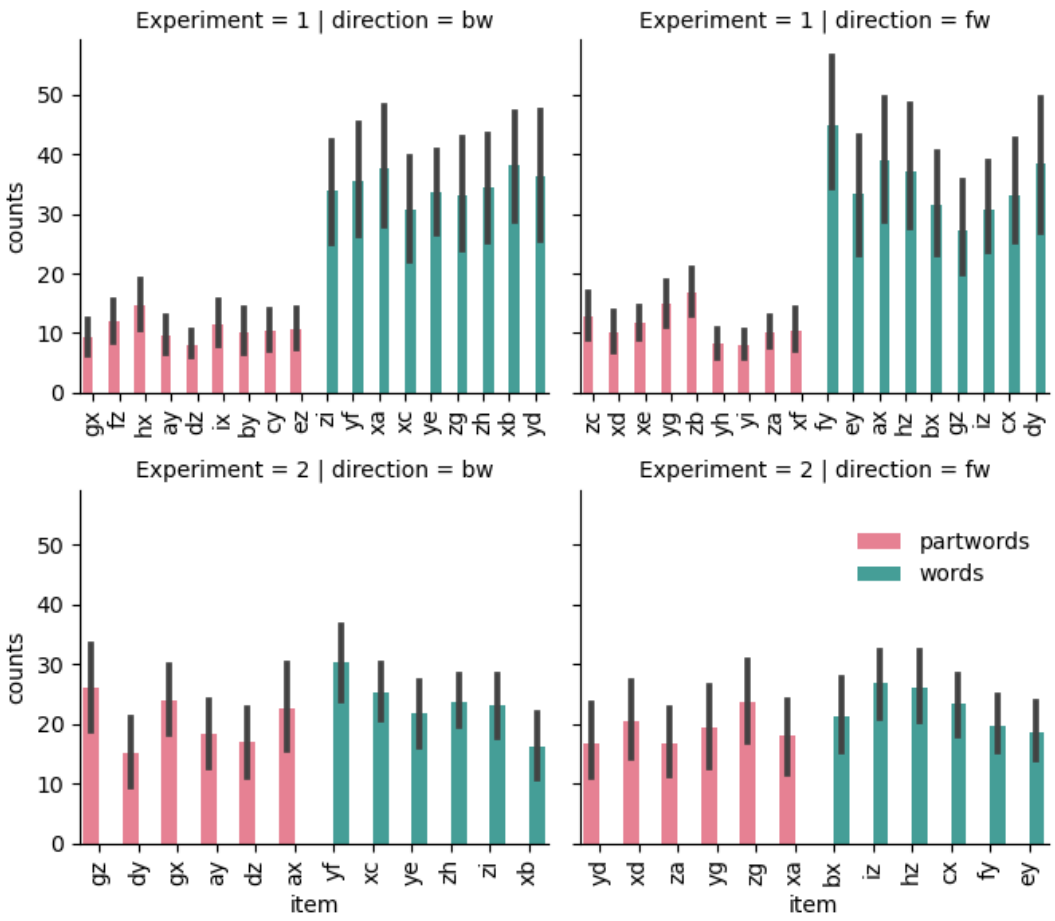

Figure 3: Subjective frequency counts of test items, as identified by the best configuration of DIORA (TreeLSTM+Margin), averaged over 30 individual models.

English could set a bias in the model to discover words based on either TP_{fw} or TP_{bw}. The fact that DIORA was successful in both English and Japanese (a language that, like Korean, has a tendency for left-branching syntactic structures) bodes well for such experiment (Li et al., 2020).

Finally, it must be noted that, to fully understand the role of TPs in word identification –specially in the absence of frequency cues– it would be useful to have experimental procedures with stricter tests, as the analyses of subjective frequencies revealed thatsuccess in the 2AFC can be achieved with only a slight difference between words and partwords.

7 Conclusion

This paper proposes a novel approach for word identification from continuous speech-like input: word segmentation as unsupervised parsing. Using this framework with DIORA revealed that word identification in AGL can be explained from the perspective of unsupervised constituency parsing, suggesting this framework can be effectively used to bridge the gap between models of word identification and syntactic syntactic processing. This work paves the way for addressing unanswered questions on the influence of syntactic knowledge in subsequent learning; in particular, an immediate next step for future work is to pre-train DIORA with head-first and head-last languages to find whether the model can be biased towards tracking forward or backward TPs.

The implications of this study are not limited to Cognitive Modelling: the use of techniques from Natural Language Processing to investigate human learning can also be fruitful for this field. In particular, one finding is that, unlike humans, DIORA discovers constituents best when those are identifiable by the frequency of co-occurrence of the related units –rather than by transitional probabilities–. Although this model was not designed to mimic human learning, incorporating the inductive biases of humans (i.e. a tendency for tracking forward or

backward dependencies depending on the degree of left- or right-branchness of the language) may be a fruitful avenue to pursue, as humans are, after all, the best-performing syntactic parsers.

Acknowledgements

I am grateful to Phong Le, Jelle (Willem) Zuidema and Afra Alishahi for their helpful comments on a previous version of this article. I also thank Phong for insightful discussions.

References

Raquel G. Alhama, R. Scha, and W. Zuidema. 2015. How should we evaluate models of segmentation in artificial language learning? In *Proceedings of 13th International Conference on Cognitive Modeling*, pages 172–173.

Raquel G. Alhama and Willem Zuidema. 2016. Generalization in artificial language learning: Modelling the propensity to generalize. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 64–72.

Raquel G. Alhama and Willem Zuidema. 2017. Segmentation as retention and recognition: the R&R model. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, page 1531–1536.

Richard N Aslin, Jenny R Saffran, and Elissa L Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4):321–324.

Carl De Marcken. 1995. The unsupervised acquisition of a lexicon from continuous speech. *arXiv preprint cmp-lg/9512002*.

Andrew Drozdov, Patrick Verga, Yi-Pei Chen, Mohit Iyyer, and Andrew McCallum. 2019. Unsupervised labeled parsing with deep inside-outside recursive autoencoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1507–1512, Hong Kong, China. Association for Computational Linguistics.

Amit Elazar, Raquel G. Alhama, Louisa Bogaerts, Noam Siegelman, Cristina Baus, and Ram Frost. 2022. When the "tabula" is anything but "rasa:" what determines performance in the auditory statistical learning task? *Cognitive Science*, 46(2):e13102.

Ansgar D. Endress and Luca L. Bonatti. 2007. Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, 105(2):247–299.

Ansgar D Endress and Scott P Johnson. 2021. When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, page 104621.

Michael C Frank, Sharon Goldwater, Thomas L Griffiths, and Joshua B Tenenbaum. 2010. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125.

Robert M French, Caspar Addyman, and Denis Mareschal. 2011. Tracx: a recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological review*, 118(4):614.

Rebecca Frost and Padraic Monaghan. 2016. Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, 147:70–74.

Adele E Goldberg. 2006. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand.

Zellig S Harris. 1970. From phoneme to morpheme. In *Papers in structural and transformational linguistics*, pages 32–67. Springer.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric 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*, pages 317–325, Boulder, Colorado. Association for Computational Linguistics.

Paul Kay and c Fillmore. 1999. Grammatical constructions and linguistic generalizations. *Language*, 75(1):1–33.

Evan Kidd, Joanne Arciuli, Morten H Christiansen, Erin S Isbilen, Katherine Revius, and Michael Smithson. 2020. Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, 200:104964.

Jun Li, Yifan Cao, Jiong Cai, Yong Jiang, and Kewei Tu. 2020. An empirical comparison of unsupervised constituency parsing methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3278–3283, Online. Association for Computational Linguistics.

Constantine Lignos. 2012. Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 30, pages 13–15.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Luca Onnis, Padraic Monaghan, Korin Richmond, and Nick Chater. 2005. Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, 53(2):225–237.

Luca Onnis and Erik Thiessen. 2013. Language experience changes subsequent learning. *Cognition*, 126(2):268–284.

Marcela Peña, Luca L Bonatti, Marina Nespor, and Jacques Mehler. 2002. Signal-driven computations in speech processing. *Science*, 298(5593):604–607.

Pierre Perruchet and Stéphane Desaulty. 2008. A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, 36(7):1299–1305.

Pierre Perruchet and Annie Vinter. 1998. Parser: A model for word segmentation. *Journal of memory and language*, 39(2):246–263.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Jenny R Saffran, Elissa L Newport, and Richard N Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621.

Noam Siegelman, Louisa Bogaerts, Amit Elazar, Joanne Arciuli, and Ram Frost. 2018. Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, 177:198–213.

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, 50(1):86–132.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.