CROSS-SITUATIONAL MODELING OF INPUT COMPLEXITY

COGNITIVE MODELS OF LANGUAGE LEARNING

DANIEL AKKERMAN, DJOURDAN GOMES-JOHNSON, LAURA GOZZO

1 INTRODUCTION TO RESEARCH QUESTION

Language acquisition is a complex cognitive process that has been the subject of extensive research within the field of cognitive science. Although there are many plausible mechanisms at play, cross-situational learning may be one of the most essential ones. Cross-situational learning concerns the inference of word meanings by observing source and target co-occurrences across multiple situations. This project specifically focuses on cross-situational learning for word acquisition, and its relation to input complexity.

Child directed language is a phenomenon where adults communicate in a modified and simplified way to children, to adapt to their current level of linguistic development. One important target of modification of language is input complexity. This can be supported by the Input Hypothesis (Krashen, 1985), which states the language learner demonstrates better learning when the complexity of the next input is i+1 compared to the current input. In other words, if the input data complexity is not i+1 (slightly more complex) compared to the current level, then language learning is observed as sub-optimal.

Previous research has shown that there is a correlation between the parents' use of child directed speech and their child's language development: simplifying the language used to communicate with children helps them to learn more efficiently (Furrow, Nelson, & Benedict, 1979). It is also known that children exposed to shorter sentences learn the meaning of words faster compared to children exposed to longer sentences (Fazly, Alishahi, & Stevenson, 2010). Parents adapt their language to match the children's increasing linguistic abilities and use more rare words as the child grows up (Prichard & Tamminga, 2012). Therefore, training the model with input of increasing complexity reflects how children are exposed to language.

This project specifically focuses on cross-situational learning for word acquisition, relating it to input complexity. The central question guiding this research is:

> *How does the complexity of input data affect the rate and accuracy of word acquisition in a cross-situational word learning model, and what insights can this model provide into real-world language learning?*

2 DATASET

This project utilizes the English-Dutch sentences from the Tatoeba corpus, preprocessed to remove punctuation. Tatoeba is an open-source bilingual corpus of 420 example languages with 10,800,000 sentences. Specifically, the English-Dutch dataset that was used

contains 75298 sentence pairs varying from one-word phrases to longer and more complex sentences.

There are five principal motivations for choosing this corpus. Firstly, as word acquisition can be seen as a mapping problem of symbols (words) to referents (meanings), a bilingual corpus is suitable because the source language words can be seen as the symbols, and the target language words can be seen as the meanings. So the two tasks essentially represent the same mapping problem. Secondly, the dataset is extensive and does not provoke any limitations regarding the quantity of input data. Thirdly, the variation of sentence length allows for a precise focus on the influence of input complexity for word learning. The fourth advantage is that the bilingual data represents a fairly realistic input, containing elements that are to be expected in real-world scenarios, such as noise. As one sentence in English might not consist of the same amount of words as its equivalent Dutch translation, not every symbol cleanly maps to one referent. Thus, the difference in word count is considered noise. The fourth motivation for choosing the dataset is the facility with which it can be evaluated. As the model will produce mappings of words from the target sentence to words from the translated sentence, even a simple dictionary would suffice to enable an evaluation.

3 MODELING AND ANALYSIS

This study simulates a cross-situational learning environment using bilingual inputs, presenting sentences in one language alongside their translations in another. The bilingual characteristic of the input data is not explanatory to the study, it is solely used as a tool in the learning environment.

The model works based on basic statistical principles, by keeping track of the co-occurrence of words. When a new sentence pair is presented to the model, for each word in the source sentence it adds a single co-occurrence count for each word in the target sentence. As more sentences are presented to the model, it will collect more information theoretically resulting in better performance.

In terms of word translation, the model estimates the most likely meaning of an input word by assessing a score for each word that has co-occurred with the target word. This score is determined by taking the co-occurrence count of the source and target word, and dividing it by the total occurrence of the target word. Without this second step, words that are generally more common would be greatly overemphasized. Finally, the co-occurring word with the highest score is returned as the most likely translation.

In order to investigate the influence of random versus progressive input complexity, two versions of the dataset were used. One contained the input in ascending order with sentences with fewer words appearing first, representing the progressive input complexity. The other version was shuffled to represent the random input complexity. The simplicity of the model is an advantage as the phenomenon of cross-situational learning is isolated from interference with other mechanisms. Additionally, the model is agnostic to input order, so if different instances of the model are fed the same data but in different order, these models will end up at the same place.

The model is assessed at different phases of development with each step representing an additional 1 percent of words processed from the dataset. In the first overview of the models, measurements are taken for every 1 percent of unique words processed. For the second overview, measurements are taken every 1 percent of total words processed. As the

progressive model will encounter less unique words in the beginning, these two methods of displaying the results provide better information to judge their relative performance. This is important because speakers in a real language learning scenario will attune the language complexity to each other's respective abilities. A corpus with too many repetitions of the same word however, would result in poor learning performance as this is the equivalent of providing language input with low complexity. The measurement interval based on unique words should be able to compensate for this possibility.

The assessment consists of asking the respective model for its translation of all source words it has encountered so far. These translations are then compared to the top 5 translations acquired with the word2word python library. If there is a match, the word is considered learned, and if not it is considered inaccurate. Both accuracy and total words learned are used as metrics to enable a good comparison of model behavior.

4 PRESENTATION OF RESULTS

Figure 1 presents the result with all plots in the left column measured in intervals of 1 percent of unique words processed, and all plots in the right column measured in intervals of 1 percent total words processed.The top two rows show the unique words learned and the accuracy for both the random and progressive input, whereas the bottom two rows show the difference between progressive and random input complexity.

Progressive input complexity outperformed random input complexity, resulting in more words learned earlier in the learning process with a higher average accuracy. The ratio of learned words was especially higher during the early phases of the learning process. This could potentially be linked to the concept of vocabulary spurt. At the end of the model cycle, when all data has been processed, the models have equal performance in both metrics of unique words learned and accuracy, as the final resulting model is not dependent on the order of the input data.

The total words measurement can represent a lower bound, whereas the unique words measurement can represent an upper bound of the power of the learning mechanism in context of the model and corpus used. The differences are most pronounced in the initial higher accuracy and relative words learned for the unique words measurement (Figure 1, rows 3 and 4).

A real speaker, for example a parent, can attempt to keep the level of complexity around i+1 for their child. A corpus, however, cannot do this. So, the total word measurement is likely too generous for the random input complexity, as there is no incentive to increase the complexity when needed, and random complexity is exposed to more useful trials early in the process.

On the other hand, the unique word measurement may be too generous for the progressive input complexity, as it mimics a scenario where input is always at a sufficient complexity. This exposes the progressive model to more total words before having processed the same amount of unique words, as it will encounter the same words more often in the beginning of the learning process as opposed to the other model.

Altogether, although there are some noticeable differences between the two methods of presenting the results, both make a strong case in terms of the research question: our models seem to confirm that providing progressively complex input data increases the rate and accuracy of word learning.

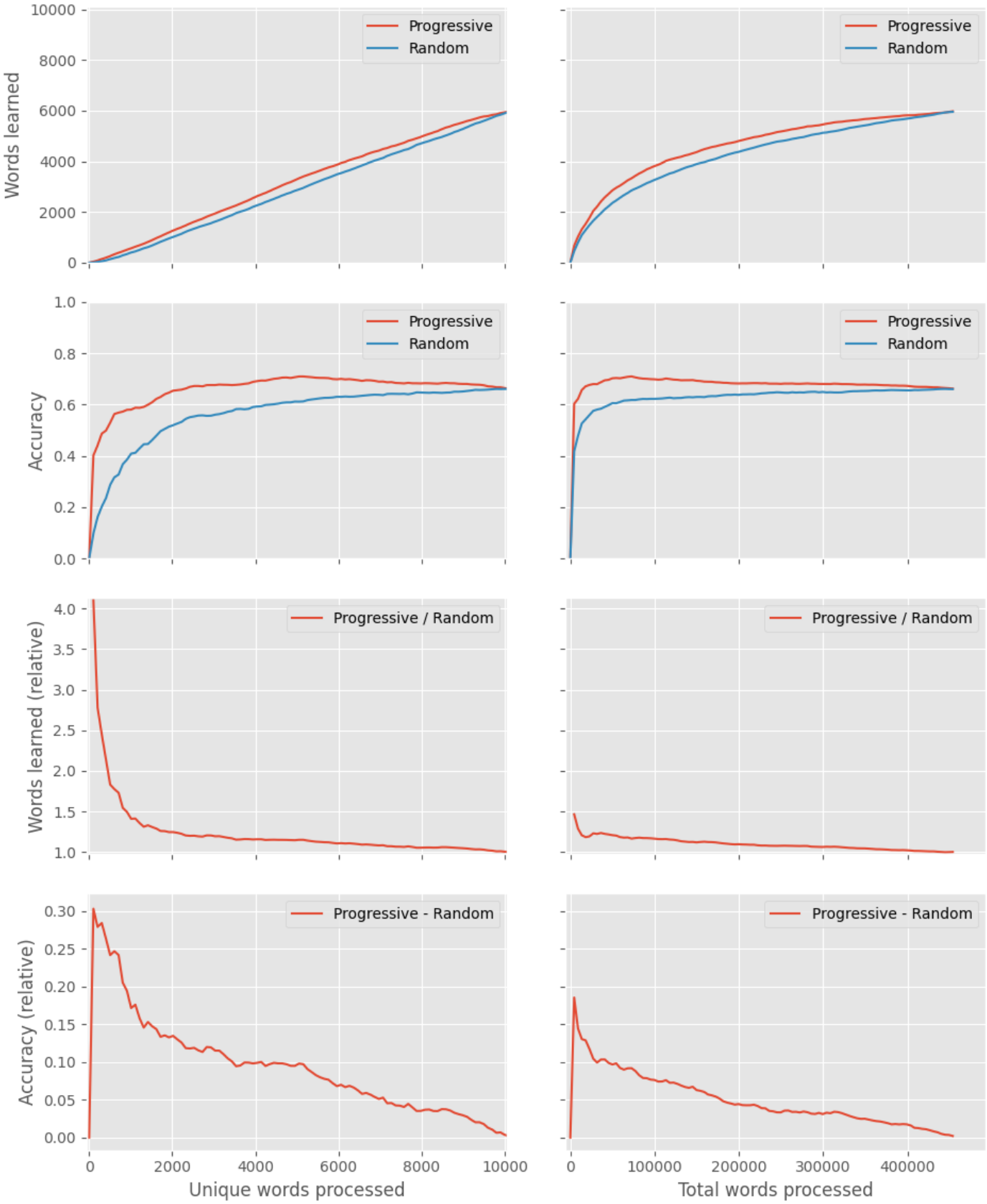

Figure 1: Progressive vs Random input complexity, measured in unique words processed (left column) and total words processed (right column)

5 CONCLUSION

While the nature of this study may be simple in design, the principal of input complexity that we address provides insight on word learning mechanisms. The data seems to support that there is an advantage focusing on simple phrases/words in the beginning of language learning, as it results in quicker increase of vocabulary size and improved accuracy of the used words. These results align with Krashen's Input Complexity Hypothesis (Krashen, 1985). The nature of progressive input complexity reflects the experience of child language learning, as child-directed speech from parents adapts input complexity to the language level of the child, which facilitates communication.

The modeling framework includes several strengths and limitations. A primary strength of the model is its simplicity of its mechanism due to reliance on the co-occurrence statistic, making the model easily explainable. Another strength lies in the possibility of future work to utilize the framework for studies on other language pairs. The languages of English and Dutch are both Germanic, thus there are several commonalities at a sentence-level between the two. Could this be the case for Mandarin and English word pairs?

In terms of limitations, the modeling framework assumes word learning as an effect of co-occurrence counts, yet neglects to address any other possible word learning mechanisms. Children use prosody and other sources of information during language acquisition. Different mechanisms may work together, compensating for their respective strengths and weaknesses. As an illustration, the initial benefits of a faster growth of vocabulary and accuracy may propel a child to be an active participant in conversations more frequently, enabling it to gain more from those experiences.

The cross-situational learning mechanism could be considered cognitively plausible because it relies on information that is realistically available to real language-learners and some of the behaviors of the model reflect real human language learning data. The model's ability to adapt to input complexity supports the human capacity to learn under words in various contexts. The progressive input complexity reflects the child language development process, and reinforces the idea that language acquisition is a cumulative process of contextual understanding. On the other hand, there are various discrepancies between the model and human cognition, such as in regards to memory. A human language learner would not be capable of remembering every word co-occurrence, and would likely require more observations to reinforce the mappings of symbols and referents.

Conclusively, this study incorporated a cross-situational model with a co-occurrence statistic to measure the effect of input complexity. Background research provided an understanding on input complexity, sentence length, and general child-directed speech. An analysis on translation accuracy and word processing determined the success, and allowed for a comparison between the two models that had different input complexities. While the modeling framework is limited to its assumptions, the simplicity of the co-occurrence mechanism reinforced the statistical properties in linguistics. This study provides further evidence for Krashen's Input Complexity Hypothesis and motivates future research on the input complexities for other symbol-referent pairs.

REFERENCES

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross situational word learning. *Cognitive Science*, *34*(6), 1017–1063. Retrieved from https://doi.org/10.1111/j.1551-6709.2010.01104.x doi: 10.1111/j.1551-6709.2010.01104.x

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language*, *6*(3), 423–442.

Krashen, S. D. (1985). *The input hypothesis: Issues and implication*. New York: Longman.

Prichard, H., & Tamminga, M. (2012). The impact of higher education on local phonology. , *18*(2), 87–95.