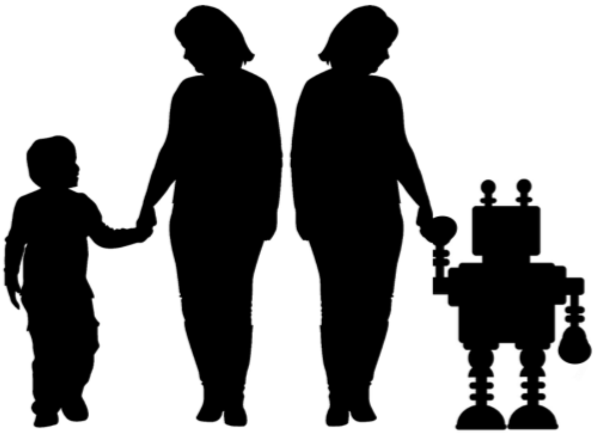
MILLing 2024

Proceedings of
the 2024 CLASP Conference on Multimodality and Interaction in Language Learning

Editors: Amy Qiu, Bill Noble, David Pagmar, Vladislav Maraev, and Nikolai Ilinykh



Gothenburg, Sweden and online
14–15 October 2024

CLASP centre for
linguistic theory
and studies in probability

CLASP Papers in Computational Linguistics, Volume 6
University of Gothenburg

CLASP Conference Proceedings, Volume 3

Front-cover art: David Pagmar.

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISSN 2002-9764
ISBN 979-8-89176-163-6

Message from the organisers

We are happy to welcome you to the CLASP Conference on Multimodality and Interaction in Language Learning (MILLing 2024)! This volume consists of the archival papers presented at the MILLing conference held at the Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg on October 14–15, 2024. The purpose of this conference was to bring together researchers in linguistics and computational linguistics to discuss learning through linguistic interaction, from the perspectives of both human language acquisition and machine learning. The conference covers areas such as theoretical linguistics, experimental linguistics, pragmatics, computational linguistics, artificial intelligence, and cognitive science.

Recent transformer-based chat systems have impressed users with their ability to sustain coherent conversations on a wide range of topics. These models are created by fine-tuning large language models (LLMs) trained on massive text corpora. In spite of their impressive performance, these models are still fundamentally different from human speakers both in their linguistic ability and in the process by which they *learn* language. This leads to a number of questions about how the LLM learning procedure compares to language acquisition for human children: Can we design more data-efficient AI models? Audio, visual, and haptic information is available to humans both prior to and throughout language acquisition. Can such multi-modal information benefit AI language learning as well? Human language acquisition follows distinct stages, and different environmental inputs are available at different stages. Can curriculum learning benefit AI models? Pre-training + RLHF is fundamentally a batch learning procedure. Can AI models be designed to incorporate learning signals on the fly, as in online reinforcement learning? Explicit instruction is, colloquially, an important way that humans learn. Chat bots can follow prompt instructions, but prompts are ephemeral and the prompting window is finite. Can AI models be designed to learn from explicit instruction?

MILLing invited papers on topics from these and closely related areas, including (but not limited to): language acquisition (especially formal, statistical, experimental, and machine learning-based work); language learning through dialogue (in both humans and machines); multi-modality and figurativeness in language learning and dialogue; linguistic variation, adaptation, and audience design; low-resource and ecologically plausible language modelling (e.g., BabyLM); cognitive architectures for language learning; information state update in humans and machines; cognitive approaches to second language acquisition; dialogue systems for language learning; online, reinforcement and curriculum learning in NLP; atypical development and language learning; and ethical considerations in AI-assisted language learning.

Accepted papers and invited talks included topics ranging from socially intelligent agents, the relationship between probabilistic spaces and language learning, insights from neurodivergent language learners and people with language disorders, pre-linguistic communication, prosody, lexical meaning acquisition and adaptation, joint attention, language-as-action, spatial perspective coordination and more. The conference, and by extension these proceedings, is a discussion about these related topics and that examines various approaches and how they can mutually inform each other. The event included presentations of 10 accepted peer-reviewed papers, including 7 archival short papers and 2 archival long papers, 4 invited keynote talks, a panel discussion, and a poster session with 10 posters. We would like to thank all our contributors, programme committee members and volunteers, with special thanks to CLASP for organising the hybrid conference and the Swedish Research Council for funding CLASP.

Amy Qiu, Bill Noble, David Pagmar, Vladislav Maraev, and Nikolai Ilinykh

Gothenburg, Sweden

October 2024

Organising Committee

Program Chairs

Amy Qiu, University of Gothenburg, Sweden
Bill Noble, University of Gothenburg, Sweden
David Pagmar, University of Gothenburg, Sweden
Vladislav Maraev, University of Gothenburg, Sweden

Proceedings Chair

Nikolai Ilinykh, University of Gothenburg, Sweden

Local Arrangements Organisers

Susanna Myyry, University of Gothenburg, Sweden
David Pagmar, University of Gothenburg, Sweden
Amy Qiu, University of Gothenburg, Sweden

Programme Committee

Maxime Amblard	University of Lorraine
Fahima Ayub Khan	University of Gothenburg
Alexander Berman	University of Gothenburg
Ellen Breitholtz	University of Gothenburg
Stergios Chatzikyriakidis	University of Crete
Robin Cooper	University of Gothenburg
Simon Dobnik	University of Gothenburg
Eleni Gregoromichelaki	University of Gothenburg
Xudong Hong	Saarland University
Julian Hough	Swansea University
Christine Howes	University of Gothenburg
Nikolai Ilinykh	University of Gothenburg
Elisabetta Jezek	University of Pavia
Richard Johansson	Chalmers Technical University & University of Gothenburg
Ruth Kempson	King's College London
Nikhil Krishnaswamy	Colorado State University
Shalom Lappin	University of Gothenburg & King's College London
Staffan Larsson	University of Gothenburg
Sharid Loáiciga	University of Gothenburg
Andy Lücking	Goethe University Frankfurt
Chiara Mazzocconi	Aix Marseille University
Louise McNally	Universitat Pompeu Fabra
Gregory Mills	Kingston University
Massimo Poesio	Queen Mary University of London
Matthew Purver	Queen Mary University of London
James Pustejovsky	Brandeis University
Mehrnoosh Sadrzadeh	University College London
Asad Sayeed	University of Gothenburg
David Schlangen	University of Potsdam
Sabine Schulte im Walde	University of Stuttgart
Vidya Somashekarappa	University of Gothenburg
Carl Vogel	Trinity College Dublin
Sina Zarrieß	University of Bielefeld

Invited Speakers

Napoleon Katsos, University of Cambridge
Catherine Pelachaud, ISIR, UPMC
Charles Yang, University of Pennsylvania
Robin Cooper, University of Gothenburg

Invited talk: Napoleon Katsos

"Deficit", "difficulty", "difference": perspectives into autistic people's pragmatic skills and their implications for research methodology

It is widely reported that autistic people face pervasive challenges with producing and understanding pragmatics, i.e. context-dependent aspects of language. These are often attributed to challenges with mentalising, i.e. the ability to attribute the correct beliefs and intentions to other people. In this talk I will select influential papers from the past three decades of research in autism and language, each of which reveal a radically different perspective on the architecture of the linguistic system and on what it means to face challenges with linguistic competence (in our case, pragmatics). I will conclude that the recent perspective of neurodiversity implies a radical re-think of how we define pragmatics and how we assess the acquisition and processing of pragmatic competence.

Invited talk: Catherine Pelachaud

Multi-forms Adaptation for Socially Interactive Agents

Interacting with others enhances learning. Getting feedback on results, being encouraged and motivated . . . all help the learning process. During interaction, participants adapt to each other to show affiliation, group belongings, or to support social bonding. Adaptation can take place at different levels, through verbal alignment, imitation, and conversational strategies. Social resonance can also serve as a marker of adaptation. Socially Interactive Agents SIAs are virtual agents with a human-like appearance, capable of communicating verbally and nonverbally with their human interlocutors. In this talk, I will present our latest works aimed at endowing an SIA with various adaptive capabilities when interacting with its partners. The adaptation mechanisms are learned from human-human interaction data and evaluated by experimental studies involving human-agent interaction.

Invited talk: Charles Yang

Why language learning is not probabilistic

It seems harmless, and certainly mathematically convenient, to treat language learning as acquiring a probabilistic distribution over a space of linguistic patterns. The goal is to find or approximate an optimal hypothesis with respect to the data. Such is the mainstream machine learning approach, and the so-called Evaluation Procedure in generative grammar can be viewed as a particular instantiation. Despite having pursued it vigorously in my earlier work, I now believe this approach is wrong (and wrong-headed). On the one hand, language is not a zero sum game: even overwhelming presence of one linguistic form does not necessarily inhibit or penalise alternative forms. On the other, the grammar can be a partial function: there are inputs for which no output form is acceptable even though some will always be most highly valued in a probabilistic framework. The alternative is a theory of learning that does not even try to optimise but only sastifice. The coverage of the data only needs to be good enough up to a point; failure to do so may just result in the memorisation of the input – nothing in the cognitive system mandates generalisation under all circumstances. I will review the psychological and computational studies of the Tolerance Principle, a parameter-free learning theory that also appears operative beyond the domain of language.

Invited talk: Robin Cooper

Types in a Theory of Interactive Learning

In this talk I will present some CLASP research on using types in a theory of interactive learning. In the first part of the talk I will introduce the notion of type we have been using and how it relates to a general theory of action, including linguistic acts. In the second part of the talk I will present some work I have been doing with Staffan Larsson and Jonathan Ginzburg on how such a theory relates to communicative acts by prelinguistic children and how such communicative acts serve as a basis for the development of linguistic acts by children. In the third part of the talk, I will present some preliminary work with Staffan, Jonathan and Andy Lücking on how the kind of types we are using might relate to the approach to neural modelling that Chris Eliasmith and colleagues have been developing. While this work is still in the very early stages, the hope is that ultimately we can propose an explanatory account of interactive learning which is grounded in biologically plausible neural activity.

Table of Contents

Critical Size Hypothesis: How Model Hyperparameters Correlate with Its Linguistic Abilities

Ekaterina Voloshina
University of Gothenburg
Chalmers University of Technology
ekaterina.voloshina@chalmers.se

Oleg Serikov
KAUST
oleg.serikov@kaust.edu.sa

Abstract

In recent years, the models were tested on different probing tasks to examine their language knowledge. However, few researchers explored the very process of models' language acquisition. Nevertheless, the analysis of language acquisition during training could shed light on the model parameters that help to acquire the language faster. In this work, we show how the model architecture seems not to influence the language acquisition process. We experiment with model hyperparameters and reveal that the hidden size is the most essential factor for model language acquisition.

1 Introduction

Modern deep learning models have achieved significant results in the field of language modeling and text generation (Krause et al., 2019; Niu et al., 2020). Therefore, language models (LMs) are often used in linguistic research to find systematic similarities in the language data. Performance of the state-of-the-art models, such as Transformer-based ones (Vaswani et al., 2017), on linguistic tasks show that they have learned measurable language structures during the training process (Warstadt and Bowman, 2022).

Consequently, it is interesting to explore how the LMs acquire the language during their training process and what part of their architecture helps to acquire a language better. In this work, we study the correlation between the acquisition process in the BERT model and different model sizes. Linguistic tasks are meant to represent three levels of language grammar structure: morphology, syntax, and discourse. In other words, we pose the following questions: which parameters of models influence the language acquisition process?

2 Related work

The first work on probing of neural networks across time was carried by Saphra and Lopez (2018). The authors showed that first, a LSTM model (Hochreiter and Schmidhuber, 1997) acquires syntactic and semantic features and later information structure. Chiang et al. (2020) looked at the training process of ALBERT (Lan et al., 2019) and concluded that semantic and syntactic information is acquired during the early steps while accuracy on world knowledge fluctuates during the training. Liu et al. (2021) showed similar results on RoBERTa (Liu et al., 2019): the model shows good results on linguistic probing tasks starting from early stages, and later it learns factual and common sense knowledge. (Blevins et al., 2022) studied training dynamics of multilingual models, they reveal that while linguistic information is acquired early, transfer learning abilities are evolving during the entire training process. Choshen et al. (2022) examined the trajectories of models' language acquisition, and they find no impact of architecture or a model size on training trajectories. Warstadt and Bowman (2022) provides survey and theoretical discussions on how neural networks can help us learn more about language acquisition. Following one of the ideas we conduct an ablation study of model's hyperparameters.

3 Methods

3.1 Models

We train small models to see how language acquisition trajectories vary depending on the model hyperparameters. Since previous research shows that the acquisition of most of the linguistic features stops after 500,000 steps, we look at the first training steps. We regard number of layers, embedding size and number of attention heads to be crucial. Therefore we train four models:

1. A base model: the hidden size of 128, 2 layers, and 2 attention heads;
2. A model with increased number of attention heads: the hidden size of 128, 2 layers, and 4 attention heads;

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 1–7
October 14–15, 2024. ©2024 Association for Computational Linguistics

3. A model with increased hidden size: the hidden size of 256, 2 layers, and 2 attention heads;
4. A model with increased number of layers: the hidden size of 128, 4 layers, and 2 attention heads.

Our hypothesis states that if any of these models show a significantly different result on any group of tasks, this parameter causes a better acquisition process. If all the models show similar results, different sizes of models do not correlate with the acquisition process, therefore, it depends on language features rather than on model parameters.

We train models with the same computational resources and data corpus, which included Wikipedia articles limited to 10,000,000 tokens. We choose this threshold as an optimal one, as according to Zhang et al. (2020), models can acquire basic linguistic information from this amount of data. We compare our model to **MultiBERT** (Sellam et al., 2021), the model with 12 layers and embedding size 768. Unlike the original BERT (Devlin et al., 2018), it was trained with 25 different seeds. We use the model with seed 0 and we use the same seed to train small models to make our results more comparable.

To explore the combination of different hyperparameters, we train several other models. We are interested in what size the model should have to behave as the model of standard size (768 embedding size, 12 layers and 12 attention heads). To calculate that, we first train a model of the same size as the multiBERT we compared to in the experiments before. Then we use it as a standard of comparison and train several models of different sizes on the same data and with the same setup as the standard BERT. We limit the training process to 100,000 iterations to find minimal parameters that help the model to achieve the accuracy of the standard model. Table 1 summarise models we trained to find the proper combination of parameters.

Model	Size	Layers	Att. heads
1	256	4	4
2	256	8	4
3	512	4	4
4	512	8	8
5	512	12	8
6	768	8	8
7	768	12	8

Table 1: Summarisation of trained models: for each model we state the hidden size of embeddings, number of layers, and number of attention heads.

3.2 Probing tasks

We use probing tasks from several probing datasets, such as SentEval (Conneau et al., 2018), Morph Call (Mikhailov et al., 2021), DisSent (Nie et al., 2019), DiscoEval (Chen et al., 2019), and BLiMP (Warstadt et al., 2020) (see examples in Tables 3 and 4):

- **Transitive verbs** includes minimal pairs of sentences with different verbs, where only one verb is transitive.
- **Passive verbs** consists of pairs that have different verbs, where only one verb can be used in a passive form.
- **Island effects** tests a model's sensibility to syntactic order. An island is a structure from which a word cannot be moved (Ross, 1967).
- **Principle A** shows the use of reflexives. According to Chomsky (1981), a reflexive should have a local antecedent, and if it does not, the sentence is ungrammatical.
- **Subject number** is a binary classification task with labels NNS and NN (plural and singular number, respectively).
- **Person** is a binary classification with labels 0 and 1, which signifies if a subject has a person marker or not.
- **Tree depth** contains six classes, each of which stand for a depth of the syntactic tree of a given sentence.
- **Top constituents** requires to identify the number of constituents located right below the sentence (S) node.
- **Connectors** includes pairs of sentences originally connected with one of 5 prepositions, and the task is to choose the omitted preposition.
- **Sentence position** contains sequences of 5 sentences, and the first sentence is placed in the wrong place. Therefore, the aim is to detect the original position of these sentences.
- **Penn Discourse Treebank** is based on Penn Discourse Treebank annotation (Marcus et al., 1994). The aim is to choose the right discourse relation between two discourse items from Penn Treebank.
- **Discourse coherence** is a binary classification with classes 1 and 0. Class 1 means that

the given paragraph is coherent, and class 0 should be assigned to paragraphs with shuffled sentences.

For tasks from BLiMP, we mask each word in a sentence, and sum probabilities of all words. The probability of an acceptable sentence should be higher than the probability of an unacceptable sentence.

For most tasks we take a sentence embedding via mean pooling. A logistic regression as a classifier model is used to classify embedded sentences. For the Sentence Position task, we calculate the difference between the first embedding and the other pairwise. The first embedding and its differences with others are concatenated and put as an input to a classifier. For other discourse tasks, we concatenated sentence embeddings, which were calculated as the mean of token embeddings.

4 Results and Discussion

First, we conducted the same experiments on four small models described in section 3.1.

As Figure 1 shows, compared to MultiBERT, models show worse accuracy. However, among small models, the one with the increased hidden size shows the best results in all cases, except for Penn Discourse Treebank and Tree depth, where the model with the increased number of layers shows the best results. This model shows the second best results on other tasks.

The behaviour of the model with the increased number of attention heads is inconsistent compared to the *base* model (hidden size of 128, 2 attention heads, and 2 layers). On some tasks, such as Penn Discourse Treebank and Discourse Coherence, it shows worse accuracy than the base model. On other tasks, it shows better quality than the base model but worse than other models.

Nevertheless, these observations are not applicable to the tasks from BLiMP. As charts show, on tasks, such as Passive and Principle A, the base model shows better quality than any other models, including the MultiBERT model. At the same time we see that small models encounter difficulties with the acceptability of sentences with transitive verbs and with islands.

The described above leads to the conclusion that bigger models are more successful in language acquisition. Different parameters of model size give different level of improvement. Thus, the most important parameter for language acquisition is hidden size, since it leads to better results for most features. The second best parameter is the number of layers.

Level	Task	Model size
morphology	subject number	768/8
morphology	person	128/2
morphology	passive	512/4
morphology	transitive	512/8
syntax	top constituents	768/8
syntax	tree depth	768/8
syntax	adjunct island	768/8
syntax	principle A	512/4
discourse	discourse coherence	128/2
discourse	Connectors	768/8
discourse	Sentence Position	512/4
discourse	Penn Treebank	128/4

Table 2: The comparison of tasks' acquisition

The results of our experiments with model sizes show that the increase of hidden size has the biggest impact on the quality of models. The number of layers was the second important parameter and improved quality better than the number of attention heads. Our results are similar to the results reported in Wang et al. (2019): they show that larger hidden size tend to improve quality.

The hidden size might be important for smaller models because different layers code different information. For example, Rogers et al. (2020) summarise that the first layers are task-invariant and contain general linguistic information while the latest layers are usually task-specific.

On the contrary, attention heads are usually more detailed, for example, they are known to remember specific syntactic patterns (Htut et al., 2019). Kovaleva et al. (2019) reveal that attention heads learn the same patterns. Therefore, when the resources to encode information are limited, attention heads do not add much new information.

Regarding the hidden size, our results are different from the results in (Wang et al., 2019). While they postulate that number of layers is the most essential parameter, our results show that hidden size is better for performance improvement.

Since on some tasks small models did not reach the level of a base model, we train more models but following the results we achieved in our experiments with model parameters, we limit our experiments to hidden size and number of layers

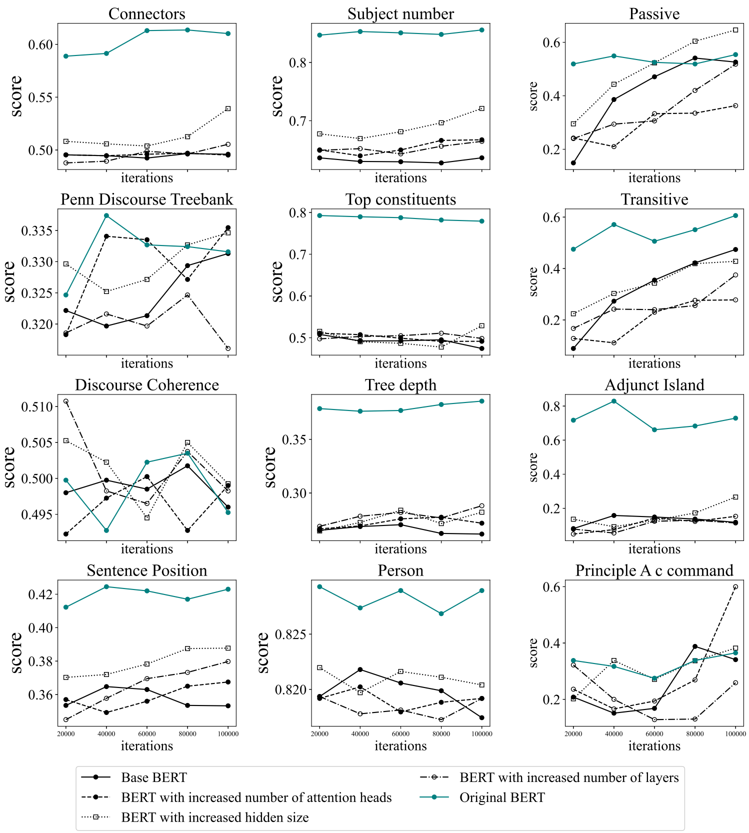


Figure 1: Small models' results on different tasks.

leaving behind the number of attention heads as an insignificant factor.

These experiments summarised in table 2 show that most of the 'morphosyntactic' tasks are acquired by models with hidden size of 768. At the same time, on discourse-based tasks, models with much smaller size show results comparable to the base model.

For most BLiMP tasks the level of the base models is achieved by models of hidden size of 518 with 4 or 8 layers, which is a smaller size than for other 'morphosyntactic' probing tasks.

The results of the experiments prove that increasing hidden size shows better results than increasing number of layers.

Moreover, models with the hidden size of 768 and 8 layers show results close to the model with the same hidden size and 12 layers. Therefore, we conclude that hidden size is the crucial parameter for language acquisition.

5 Conclusion

This works addresses the problem of language acquisition in state-of-the-art models and answers which factors influence the language acquisition process.

To display correlation between language acquisition and different model parameters, we trained four models: one with the minimal hidden size and minimal number of layers and attention heads and three models with one parameter increased and others frozen. These experiments reveal that hidden size appears to be the most essential parameter for language acquisition, whereas attention heads do not significantly increase a model's performance.

Finally, we compared all tasks with the size of a model that shows the quality comparable with the base model used before. The idea behind this comparison is to find any correlation between different language levels and probing measures. As a result, models distinguish discourse from morphology and syntax but there is almost no difference between 'morphological' and 'syntactic' tasks.

References

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proc. of EMNLP*.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*.

Noam Chomsky. 1981. Lectures on government and binding (dordrecht: Foris). *Studies in generative grammar*, 9.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Ben Krause, Emmanuel Kahembwe, Iain Murray, and Steve Renals. 2019. Dynamic evaluation of transformer language models. *arXiv preprint arXiv:1904.08378*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021. Morph call: Probing morphosyntactic content of multilingual transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 97–121, Online. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Tong Niu, Semih Yavuz, Yingbo Zhou, Huan Wang, Nitish Shirish Keskar, and Caiming Xiong. 2020. Unsupervised paraphrase generation via dynamic blocking. *arXiv preprint arXiv:2010.12885*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

John Robert Ross. 1967. Constraints on variables in syntax.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2019. Cross-lingual ability of multilingual bert: An empirical study. *arXiv preprint arXiv:1912.07840*.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2020. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

A Examples of Tasks

Task	Sentence examples	Labels
Subject number	*Her employer had escaped with his wife for several afternoons this summer.*	NN
	Your Mackenzie in-laws have sordid reputations few decent families wish to be connected with.	NNS
Person	*So I still can recomend them but prepare pay twice as much as they tell you initially.*	has a person marker
	The service was friendly and fast, but this just does nt make up for the lack - luster product.	does not have a person marker
Tree depth	*We have done everything we can for her .*	11
	Alvin Yeung of Civic Party	3
Top constituents	*Did it belong to the owner of the house ?*	VBD_NP_VP_.
	How long before you leave us again ?	WHNP_SQ_.
Connectors	*He 'd almost forgotten about that man . Sarah had somehow brought him back , just as she had his nightmares .*	but
	I let out a slow , careful breath . Felt tears sting my eyes .	and
Sentence position	*Quneitra Governorate (/ ALA-LC : “ Muhāfaẓat Al-Qunaytrah “) is one of the fourteen governorates (provinces) of Syria . The governorate had a population of 87,000 at the 2010 estimate . Its area varies , according to different sources , from 685 km ^{2} to 1,861 km ^{2} . It is situated in southern Syria , notable for the location of the Golan Heights . The governorate borders Lebanon , Jordan and Israel .*	1
	The bossom and the part of the xhubleta covered by the apron are made out of crocheted black wool . The bell shape is accentuated in the back part . The xhubleta is an undulating , bell-shaped folk skirt , worn by Albanian women . It usually is hung on the shoulders using two straps . Part of the Albanian traditional clothing it has 13 to 17 strips and 5 pieces of felt .	4
Penn Discourse Treebank	*Solo woodwind players have to be creative,they want to work a lot*	Pragmatic Cause
	The U.S. , along with Britain and Singapore , left the agencyl, its anti-Western ideology , financial corruption and top leadership got out of hand	List
Discourse Coherence	*Within the fan inlet case , there are anti-icing air bosses and probes to sense the inlet pressure and temperature .', 'High speed center of pressure shifts along with fin aeroelasticity were major factors . At the 13th (i.e .', 'the final) compressor stage , air is bled out and used for anti-icing . The amount is controlled by the Pressure Ratio Bleed Control sense signal (PRBC) . The “ diffuser case “ at the aft end of the compressor houses the 13th stage .*	a text is not coherent
	This experience of digital circuitry and assembly language programming formed the basis of his book “ Code : The Hidden Language of Computer Hardware and Software ” . Petzold purchased a two-diskette IBM PC in 1984 for $ 5,000 . This debt encouraged him to use the PC to earn some revenue so he wrote an article about ANSI.SYS and the PROMPT command . This was submitted to PC Magazine for which they paid $ 800 . This was the beginning of Petzold 's career as a paid writer . In 1984 , PC Magazine decided to do a review of printers .	a text is coherent

Table 3: Examples of tasks

Task	Acceptable sentence	Unacceptable sentence
Transitive	*The pedestrians question some people.*	*The pedestrians wave some people.*
Passive	*Tracy isn't fired by Jodi's daughter.*	*Tracy isn't muttered by Jodi's daughter.*
Principle A c command	*This lady who is healing Charles wasn't hiding herself.*	*This lady who is healing Charles wasn't hiding himself.*
Adjunct Island	*Who does John leave while alarming Beverly?*	*Who does John leave Beverly while alarming?*

Table 4: BLiMP Minimal pairs examples

INIKOL - Collocational Database for Learning Croatian as a Foreign Language

Goranka Blagus Bartolec[1], Gorana Duplančić Rogošić[2], Antonia Ordulj[3]

[1] Institute for the Croatian Language Zagreb
[2] Faculty of Economics, Business, and Tourism Split
[3] Faculty of Croatian Studies Zagreb

gblagus@ihjj.hr, gduplanc@efst.hr, antoniasvetic@gmail.com

Abstract

This paper describes the ongoing work on the INIKOL project - the development of a collocation database for learning Croatian as a foreign language. The main goal of the project is to contribute to easier mastery of collocations as fixed phrases in Croatian as a foreign language.

1 Introduction

Collocations, which are multi-word expressions (MWEs) with a fixed structure and meaning, are a challenge for non-native speakers of Croatian who have difficulty understanding and using them in terms of: 1. recognizing the elements of a collocation and understanding their meanings (see Ordulj and Naumoska-Giel, 2022; Goh, 2000; Graham, 2006), 2. recognizing the part of speech when selecting collocates and choosing the correct morphological form of the inflected word (Ordulj, 2018a; 2018b), 3. using the correct preposition with the appropriate case form of the noun, 4. linking individual words to other words within a phrase or sentence.

The number of students learning Croatian has been steadily increasing over the last two decades, so the need for high-quality manuals and an applied description of the Croatian language is also growing. The groups of students learning Croatian as a foreign language are extremely heterogeneous in terms of mother tongue, age, gender and previous knowledge; their motives for learning Croatian are also different. Foreigners who learn Croatian as non-native speakers come from different language areas: many come from Slavic countries as students of Croatian studies and the Croatian language, so the structure and vocabulary of Croatian are close to them, but they also come from other language areas where the structure of Croatian is unfamiliar and more difficult to learn. Many participants come from South America and are descendants of Croats who emigrated from Croatia at the beginning or middle of the 20th century. There are also many learners from other European countries who are in Croatia for professional or private reasons and want to learn Croatian. In recent years, the number of immigrants, i.e. workers from Asian countries who need to learn Croatian at a basic level, has also increased.

Textbooks and exercise books for learning and teaching Croatian as a foreign language as well as the more recent *Basic Croatian Grammar for Croatian Language Learners* (Matovac, 2022) created in Croaticum, the largest institution for teaching, research and description of Croatian as a second and foreign language, offer non-native speakers a good insight into the structure and lexical potential of the Croatian language from beginner to intermediate level.

Given the frequency of collocations in everyday use and all the above-mentioned challenges that non-native speakers face when learning Croatian, the INIKOL project was developed to build a collocation database as an additional, publicly accessible resource that non-native speakers can use as an online tool for

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 8–12
October 14–15, 2024. ©2024 Association for Computational Linguistics

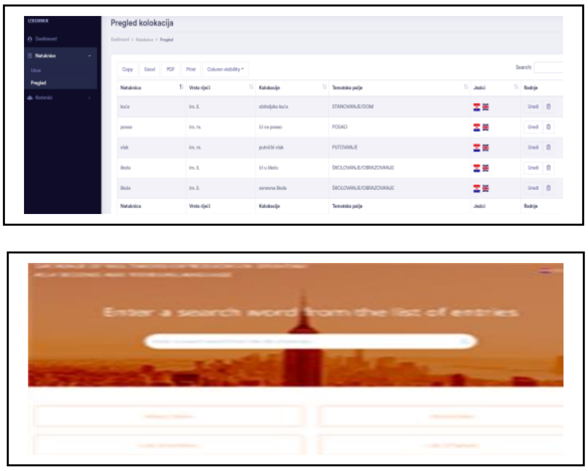
searching, understanding and applying these expressions when learning Croatian and when communicating in Croatian in various contexts of use.

2 About the INIKOL database

The INIKOL database is part of two larger projects: MWE-Cro: Multiword Expressions in Croatian - Lexicological, Computational Linguistic and Glottodidactic Approach, and VIBA: Database of Croatian MWEs, which aim to build a complete, publicly accessible online platform for multi-word expressions in Croatian, which will include: a) several monolingual databases, namely of idioms, proverbs and multi-word expressions in general use and in languages for specific purposes, b) a multilingual database of verb collocations and c) a multilingual database of collocations in Croatian as a foreign language (INIKOL). Acquiring and understanding fixed word combinations in Croatian, as in other languages, is a challenging and demanding task for non-native speakers, especially for non-Slavic ones. The morphology of inflected nouns and verbs and the selection of the appropriate collocation are often a significant obstacle that makes it difficult for non-native speakers to learn and reproduce MWEs in Croatian. Therefore, the basic aim of building INIKOL is to contribute to the easier acquisition and use of collocations in Croatian as a foreign language. During the four-year project period[1], a total of 800 to 1,000 collocations from the basic vocabulary of Croatian as a foreign language will be entered into the INIKOL database. The collocations entered into the database follow the *Croatian A2: Descriptive Framework of Reference Level A2* (Grgić and Gulešić Machata, eds., 2017), which is based on the Common European Framework of Reference for Languages (CEFR) and the content of textbooks for teaching Croatian as a foreign language at level A2. All collocations are grouped into thematic areas (e.g. *MAN, EDUCATION, LIVING, TRAFFIC AND TRAVEL, FOOD AND DRINK, SHOPPING, SERVICES*). The main entries in the database are nouns (e.g. *family, school, house, train, city, language, sea, park, glass...*) and verbs (e.g. *to be, to go, to eat, to drink, to write, to study, ...*). Verbal and noun entries are key words under which collocations are listed in INIKOL. For example, the collocation *biti gladan*, eng. 'to be hungry,' is listed under the verb *biti*, eng. 'to be' as a verb entry, and the collocation *obiteljska kuća*, eng. 'family house' is listed under the noun *house* as a noun entry. Other parts of speech, such as prepositions, adverbs, and adjectives, will also be included as entries in the INIKOL in the following phase.

2.1 INIKOL structure: user interface outline

Collocations are entered in the online working interface (backend) and all entered data is visible in the user interface (frontend) after saving. The interfaces were created by an external IT developer according to the ideas of the project member (see Figure 1). The project members access the user interface (backend) via a user name and password, and while working on the project, the user interface (frontend) is only visible and accessible to the project members, but not publicly visible and searchable.



Natuknica	Vrsta riječi	Kolokacije	Tematska polja	Jezici	Radnje
kuća	im. ž.	obiteljska kuća	STANOVANJE/DOM	🇭🇷 🇬🇧	Uredi
posao	im. m.	ići na posao	POSAO	🇭🇷 🇬🇧	Uredi
vlak	im. m.	putnički vlak	PUTOVANJE	🇭🇷 🇬🇧	Uredi
škola	im. ž.	ići u školu	ŠKOLOVANJE/OBRAZOVANJE	🇭🇷 🇬🇧	Uredi
škola	im. ž.	osnovna škola	ŠKOLOVANJE/OBRAZOVANJE	🇭🇷 🇬🇧	Uredi
Natuknica	Vrsta riječi	Kolokacije	Tematska polja	Jezici	Radnje

Figure 1: INIKOL – backend and frontend

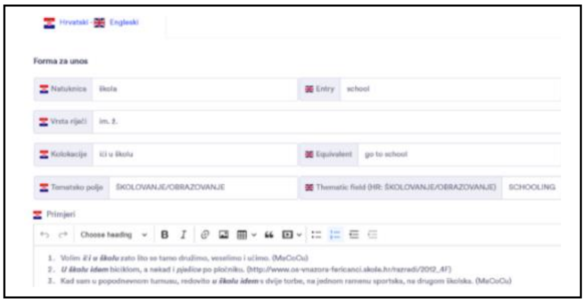
The frontend of the INIKOL database, i.e. the final result of the project, will be publicly accessible and searchable under the Croatian domain jezik.hr after the project is completed.

When the interface opens to the public, users will be able to search: a) by individual entries (one-word lexemes), which are elements of multi-word expressions in Croatian and English; b) by multi-word expressions in Croatian (by selecting from a

[1] INIKOL is an ongoing project that will run from December 2023 to December 2027 at the Institute for the Croatian Language in Zagreb.

drop-down menu, which will be sorted alphabetically by the initial word of the multi-word expression); and c) by the thematic area to which the multi-word expression belongs. At the moment, it is only possible to search for Croatian and English entries in the test version of the frontend.

The user interface for INIKOL consists of seven fields (see layout of the backend interface in Figure 2). For Croatian, four fields are filled in: 'Entry' (one-word lexeme as collocation element under which collocations are entered in INIKOL), 'Part of speech' (for the entry); 'Collocation/MWE', 'Thematic field' to which the collocation belongs, 'Examples from the corpus' and three fields for English equivalents: 'Entry', 'Collocation/MWE' and 'Thematic field'.



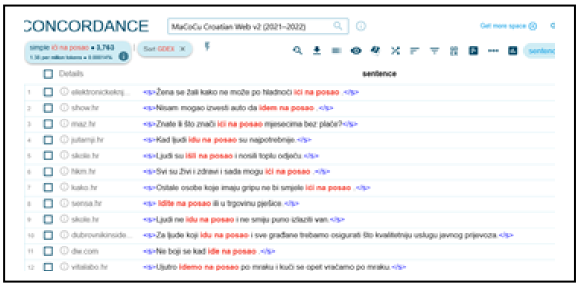
Natuknica	škola	Entry	school
Vrsta riječi	im. ž.		
Kolokacije	ići u školu	Equivalent	go to school
Tematsko polje	ŠKOLOVANJE/OBRAZOVANJE	Thematic field (HR: ŠKOLOVANJE/OBRAZOVANJE)	SCHOOLING

1. Volim ***ići u školu*** zato što se tamo družimo, veselimo i učimo. (MaCoCu)
2. ***U školu idem*** biciklom, a nekad i pješice po pločniku. (http://www.os-vnazora-fericanci.skole.hr/razredi/2012_4F)
3. Kad sam u popodnevnom turnusu, redovito ***u školu idem*** s dvije torbe, na jednom ramenu sportska, na drugom školska. (MaCoCu)

Figure 2: Layout of backend interface in Croatian and English for collocation *ići u školu*, eng. 'go to school')

2.2 Corpus-based examples in INIKOL

An essential part of the development of the INIKOL database is the inclusion of example sentences from the corpus. Two or more example sentences are provided for each entry in order to confirm the use of the collocations in practice. The source corpus is the Croatian corpus *MaCoCu Croatian Web v2 2021–2022* (Bañón et al., 2023), which is available in Sketch Engine as the Croatian version of the multilingual corpus platform MaCoCu corpora, which were built by indexing the Internet's top-level domains in 2021 and 2022. In the Croatian version of MaCoCu, users can search for examples using the simple query option, with the possibility of searching for lemmatic forms of the elements of collocations. In this way, it is possible to find: 1. all paradigmatic forms in which collocations are recorded in the corpus, which is important because Croatian is an inflectional language, 2. the valency patterns between words at the syntagmatic and syntactic levels. After obtaining the overall results through a simple query, the GDEX option is selected for the automatic selection of sentences that are easily understandable for non-native speakers and suitable for teaching. Figure 3 shows the results of a corpus search for the collocation *ići u školu*, eng. 'go to work' in MaCoCu using the GDEX option.


CONCORDANCE
MaCoCu Croatian Web v2 (2021–2022)
Get more space
simple ići na posao • 3,763
1.36 per million tokens • 0.00014%
Sort GDEX
sentence

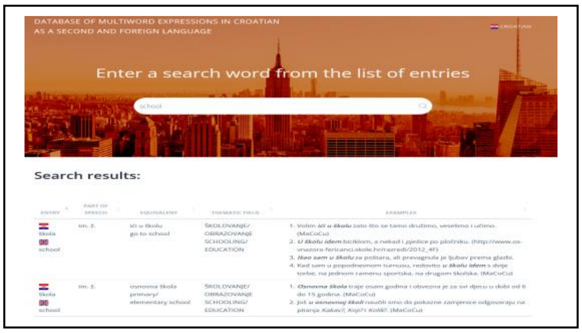
	Details	sentence
1	elektronickekoj...	<s>Žena se žali kako ne može po hladnoći ići na posao .</s>
2	show.hr	<s>Nisam mogao izvesti auto da idem na posao .</s>
3	maz.hr	<s>Znate li što znači ići na posao mjesecima bez plaće?</s>
4	jutarnji.hr	<s>Kad ljudi idu na posao su najpotrebnije.</s>
5	skole.hr	<s>Ljudi su išli na posao i nosili toplu odjeću.</s>
6	hkm.hr	<s>Svi su živi i zdravi i sada mogu ići na posao .</s>
7	kako.hr	<s>Ostale osobe koje imaju gripu ne bi smjele ići na posao .</s>
8	sensa.hr	<s> Idite na posao ili u trgovinu pješice.</s>
9	skole.hr	<s>Ljudi ne idu na posao i ne smiju puno izlaziti van.</s>
10	dubrovnikinside...	<s>Za ljude koji idu na posao i sve građane trebamo osigurati što kvalitetniju uslugu javnog prijevoza.</s>
11	dw.com	<s>Ne boji se kad ide na posao .</s>
12	vitalabo.hr	<s>Ujutro idemo na posao po mraku i kući se opet vraćamo po mraku.</s>

Figure 3: The first 12 examples of the MWE *ići na posao*, eng. 'to go to work', retrieved using GDEX

Based on the retrieved examples, the entry is entered and edited in the backend by selecting sentences from the corpus that a non-native speaker could understand and that include different morphological forms of the individual elements of the collocation. The use of Croatian prepositions in certain cases is rather challenging for non-native speakers as the noun used determines the choice of preposition, e.g. the prepositions *u* and *na* (literally *in* and *on* respectively in English) are in Croatian '*ići na posao*', eng. 'to go to work' and *ići u školu*, eng. 'to go to school'. The inclusion of such collocations in INIKOL will make it easier for non-native speakers to use such collocations correctly in Croatian.

2.3 INIKOL database frontend

The frontend of the INIKOL database follows the structure of the backend interface and enables the display of entries in Croatian and English interfaces. Figure 4 shows the search for the entry *school* in the English interface. The user will be able to use both the Croatian and the English interface. The English equivalents will be retrievable through the English interface and the Croatian equivalents through the Croatian search engine.



DATABASE OF MULTIWORD EXPRESSIONS IN CROATIAN
AS A SECOND AND FOREIGN LANGUAGE

Enter a search word from the list of entries

school

Search results:

ENTRY	PART OF SPEECH	EQUIVALENT	THEMATIC FIELD	EXAMPLES
škola school	im. ž.	ići u školu go to school	ŠKOLOVANJE/ OBRAZOVANJE SCHOOLING/ EDUCATION	1. Volim *ići u školu* zato što se tamo družimo, veselimo i učimo. (MaCoCu) 2. *U školu idem* biciklom, a nekad i *pješice* po pločniku. (http://www.os-vnazora-fericanci.skole.hr/razredi/2012_4F) 3. *Išao sam u školu* za poštara, ali prevagnula je ljubav prema glazbi. 4. Kad sam u popodnevnom turnusu, redovito *u školu idem* s dvije torbe, na jednom ramenu sportska, na drugom školska. (MaCoCu)
škola school	im. ž.	osnovna škola primary/ elementary school	ŠKOLOVANJE/ OBRAZOVANJE SCHOOLING/ EDUCATION	1. *Osnovna škola* traje osam godina i obvezna je za svi djecu u dobi od 6 do 15 godina. (MaCoCu) 2. Još *u osnovnoj školi* naučili smo da pokazne zamjenice odgovaraju na pitanja *Kakav?, Koji?i Koliki?*. (MaCoCu)

Figure 4: Layout of the INIKOL frontend with the entry *škola,* eng. 'school'

A list of the entries entered in INIKOL is available in Croatian and English on the home page of the frontend (see Figure 5) and can be searched using the Croatian or English search engine.

Enter a search word from the list of entries

Most searched

Last added

Figure 5: INIKOL English home page with a list of searchable and added entries

3 Further work on INIKOL

The INIKOL database is planned as a dynamic project that offers the possibility of updating existing data and entering new collocations according to user needs and changes in extra-linguistic reality. The database will be expanded to include additional fields and a sound recording of the pronunciation of each collocation.

Collocations at advanced levels will be added to the INIKOL database, which in the first phase includes the basic vocabulary, as the ultimate goal is to include collocations from all CEFR levels, with each collocation labelled with the level at which it is acquired. In addition to the English equivalents, the database will be expanded to include other languages from the countries from which non-native speakers who learn Croatian come. In a further step, in addition to the MaCoCu corpus, examples from the CroLTeC - CROatian Learner TExt Corpus (Mikelić Preradović et al., 2015) will be entered. CroLTeC contains essays collected from learners of Croatian as a second and foreign language (from A1 to C1 level).

The INIKOL project will provide online tools for Croatian as a foreign language and contribute to its visibility, prominence and use by non-native speakers. It will also help place Croatian in line with other European languages that already have such tools in wider use such as English (e.g. Oxford Collocations Dictionary; Collins Dictionary), Spanish (Dictionary of Collocations - DICE) or French (Le Robert).

Acknowledgments

This work was supported by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-7697, and under the project VIBA - Database of Croatian MWEs funded by the EU - NextGenerationEU. Author(s) opinions do not necessarily reflect those of the European Union or the European Commission. Neither the EU nor the EC can be held responsible for them.

References

Ana Grgić and Milvia Gulešić Machata (Eds.). 2017. *Hrvatski A2: Opisni okvir referentne razine A2 [Croatian A2: Descriptive framework of reference level A2]*. FF press, Zagreb.

Antonia Ordulj. 2018a. *Kolokacije u hrvatskom kao inom jeziku: Uvid u receptivno i produktivno znanje imenskih kolokacija*. HSN, Zagreb.

Antonia Ordulj. 2018b. Analiza odgovora imenskih kolokacija kod neizvornih govornika hrvatskoga jezika, *Journal for Foreign Languages*, 10 (1): 133-153. https://doi.org/10.4312/vestnik.10.133-153.

Antonia Ordulj and Karina Naumoska-Giel. 2022. “Razumijete li me?” ili o slušanju u nastavi hrvatskoga kao inoga jezika. In *Od ucha do ucha.* 243-255. Uniwersytet im. Adama Mickiewicza w Poznaniu, Poznań.

Collins. 2024. *Collins Online Dictionary*. https://www.collinsdictionary.com.

Christine M. Goh 2000. A cognitive perspective on language learners' listening comprehension problems. *System*, 28(1): 55–75. https://doi.org/10.1016/S0346-251X(99)00060-3.

Darko Matovac. 2022. *Basic Croatian Grammar: For Croatian Language Learners*. HSN, Zagreb.

DiCE: Diccionario de Colocaciones del Español. Facultade de Filoloxía (Universidade da Coruña). https://www.dicesp.com/paginas.

Dictionnaire français gratuit – Dico en ligne Le Robert. Le Robert. https://dictionnaire.lerobert.com.

Marta Bañón et al. 2023. *Croatian web corpus MaCoCu-hr 2.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042,1.

Nives Mikelić Preradović, Monika Berać, Damir Boras. 2015. Learner Corpus of Croatian as a Second and Foreign Language. *Multidisciplinary Approaches to Multilingualism*. In K.Cergol Kovačević, S.L.Udier. Peter Lang, F am M.107-126.

Oxford collocations dictionary at Oxford learner's dictionaries. 2024. Oxford University Press.

Suzanne Graham. 2006. Listening comprehension: The learners' perspective. *System*, 34(2): 165–182. https://doi.org/10.1016/j.system.2005.11.001.

Textbooks and workbooks – Croaticum. https://croaticum.ffzg.unizg.hr/?page_id=1632.

How Does an Adjective Sound Like? Exploring Audio Phrase Composition with Textual Embeddings

Saba Nazir and **Mehrnoosh Sadrzadeh**
Department of Computer Science
University College London, United Kingdom
saba.nazir.19@ucl.ac.uk, m.sadrzadeh@ucl.ac.uk

Abstract

We learn matrix representations for the frequent sound-relevant adjectives of English and compose them with vector representations of their nouns. The matrices are learnt jointly from audio and textual data, via linear regression and tensor skipgram. They are assessed using an adjective similarity benchmark and also a novel adjective-noun phrase similarity dataset, applied to two tasks: semantic similarity and audio similarity. Joint learning via Tensor Skipgram (TSG) outperforms audio-only models, matrix composition outperforms addition and non compositional phrase vectors.

1 Introduction

Natural language data consists of words arranged into phrases and sentences. Words have statistical representations and phrases/sentences symbolic forms. The formers, mined from co-occurrence counts, fall within the remit of distributional lexical semantics. The latters, often formalised within logic frameworks, are obtained from rules of grammar. A model of natural language should ideally take both into account. Consider a simple adjective-noun phrase. On the lexical side, statistical vector embeddings are learnt for adjectives and nouns. On the symbolic side, e.g. in Combinatory Categorial Grammar (CCG) (Steedman, 2002), an adjective is a function applied to a noun. The lexical and the symbolic sides are brought together by providing a statistical representation for the CCG rules. For the adjective-noun phrase rule, this is achieved by representing adjectives as matrices, nouns as vectors, and function application by matrix-vector multiplication (Baroni and Zamparelli, 2010). This unified model has been applied to multimodal image-text data (Lewis et al., 2022), but never to other combinations such as audio-text. For example, in an audio-text context, adjectives like "loud" or "soft" can modify nouns like "music," where the meaning is enriched by integrating corresponding audio features with their textual representations. Our aim in this paper is to fill this gap. We represent the sounds of adjectives by matrices, the sounds of nouns by vectors, and test whether their matrix-vector multiplication is a good representative of the sound of adjective-noun phrase. To this end, we work with two tasks: a semantic similarity task and an audio similarity one. We develop a new dataset of audio relevant adjective-noun phrases and collect human annotations for them. The matrix representations are from the audio data gathered from FreeSound[1], a collaborative repository of sounds. The correlation between the model's predictions and human annotations is tabulated. These show that matrix-vector adjective-noun composition works better than simple vector addition and non-compositional vectors of adjective-noun phrases. The quality of the audio adjectives significantly improved after auditory and textual data were combined and textual data used as a signal in audio adjective learning. These results show that matrix composition leads to better representations for audio phrases, with potential applications to audio classification (Xie and Virtanen, 2021) and captioning tasks (Mahfuz et al., 2023).

2 Related Work

Using vector addition for composing adjectives with nouns was proposed in (Mitchell and Lapata, 2008). Later, in a series of papers (Grefenstette and Sadrzadeh, 2011; Baroni and Zamparelli, 2010; Maillard and Clark, 2015), it was argued that vector addition is not appropriate for composition as it is commutative. Furthermore, an adjective needs to *modify* the meaning of a noun, thus its representation should be a map, rather than a vector. In finite dimensions, maps are approximated by matrices and adjective-noun phrase composi-

[1] https://freesound.org

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 13–18
October 14–15, 2024. ©2024 Association for Computational Linguistics

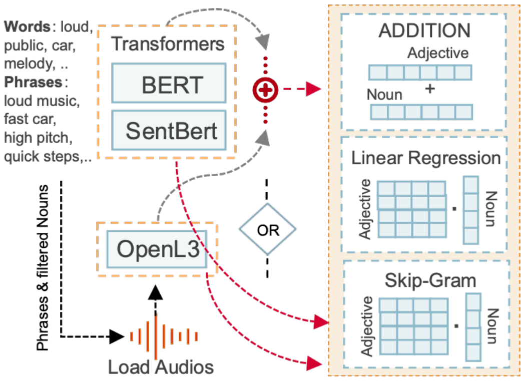


Figure 1: For audio vectors, we used the pre-trained OpenL3 (Cramer et al., 2019) library, trained on environmental and musical data from AudioSet (Gemmeke et al., 2017). OpenL3 uses a convolutional architecture initialised on a Mel-spectrogram time-frequency representation with 256 bands; its vectors are 512 dimensional. For textual vectors, we used 768 dimensional pre-trained BERT embeddings (Devlin et al., 2018) for words and SBERT (Reimers and Gurevych, 2019) for phrases.

tion becomes matrix-vector multiplication, a non-commutative operation. Different methodos were put forwards for learning the adjective matrices; (Baroni and Zamparelli, 2010) used linear regression and (Maillard and Clark, 2015; Wijnholds and Sadrzadeh, 2019) developed a tensorial extension of the word2vec skipgram model (Mikolov et al., 2013). Learning multimodal image-text embeddings for words was proposed in (Bruni et al., 2014; Lazaridou et al., 2015); extended to sound-text in (Kiela and Clark, 2015). Matrix composition of images and text was explored in (Lewis et al., 2022).

3 Single and Multi Modal Learning

An overview of multimodal phrase composition is presented in Figure 1. To learn the matrices, we used linear regression (LR) and the tensorial extension of skipgram (TSG). For LR, we trained adjective matrices $\boldsymbol{A}$ given observed adjective-noun vectors $\boldsymbol{p}$ and noun vectors $\boldsymbol{v}$, using the formula $\boldsymbol{p} = \boldsymbol{A}\boldsymbol{v}$.

The original word2vec skipgram model had the following objective function, where $\boldsymbol{n}$ is a vector, and $\mathcal{C}$ and $\overline{\mathcal{C}}$ sets of positive and negative contexts.

$$\sum_{\boldsymbol{c}' \in \mathcal{C}} \log \sigma\left(\boldsymbol{w} \cdot \boldsymbol{c}'\right) + \sum_{\overline{\boldsymbol{c}}' \in \overline{\mathcal{C}}} \log \sigma\left(-\boldsymbol{w} \cdot \overline{\boldsymbol{c}}'\right)$$

This model learns a vector for a word w regardless of its grammatical type. Its tensorial extension, dubbed as **tensor skipgram** has an objective function that depends on the grammatical role of the words. For adjective-noun phrases, this is as follows, where $\mathbf{A}$ is the adjective matrix, $\boldsymbol{n}$ the vector of the noun it modifies, and the rest is as before.

$$\sum_{\boldsymbol{c}' \in \mathcal{C}} \log \sigma\left(\mathbf{A}\boldsymbol{n} \cdot \boldsymbol{c}'\right) + \sum_{\overline{\boldsymbol{c}}' \in \overline{\mathcal{C}}} \log \sigma\left(-\mathbf{A}\boldsymbol{n} \cdot \overline{\boldsymbol{c}}'\right)$$

The above function is only for adjective-noun phrases. It generalises to any phrase in (Wijnholds and Sadrzadeh, 2019). TSG significantly outperforms LR on text (Maillard and Clark, 2015; Wijnholds and Sadrzadeh, 2019).

The audio and textual representations were combined with two different methods. In the first method, we concatenated their vectors (**AT-Concat**) and used the result as an input to training. In the second method, we trained a joint audio-text matrix (**AT-Joint**), where one representation was used as a signal to improve the other.

AT-Concat Regression uses the following adaptation of the above single modality regression:

$$\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle = \mathbf{A}\langle \boldsymbol{v}^a, \boldsymbol{v}^t \rangle$$

where $\boldsymbol{v}^a$ is the audio representation of a noun, $\boldsymbol{v}^t$ its textual counterpart, and $\langle \boldsymbol{v}^a, \boldsymbol{v}^t \rangle$ their concatenation. Similarly, $\boldsymbol{p}^a$ is the audio representation of an adjective-noun phrase, $\boldsymbol{p}^t$ its textual counterpart, and $\langle \boldsymbol{p}^a, \boldsymbol{p}^t \rangle$ their concatenation.

AT-Joint Regression uses the following variant of the original regression formula $\boldsymbol{p}^a = \mathbf{A}\boldsymbol{v}^t$ for training, where the audio adjective-noun phrase vector $\boldsymbol{p}^a$ uses the textual representation of its noun $\boldsymbol{v}^t$ as a signal to learn an adjective matrix $\mathbf{A}$, which has a combined audio-text meaning.

AT-Concat Tensor Skipgram is based on the modified training objective of the single modality TSG and has the following objective function (to save space we only provide the positive sampling part):

$$\sum_{(\boldsymbol{c}'^{\boldsymbol{a}}, \boldsymbol{c}'^{\boldsymbol{t}}) \,\in\, \mathcal{C}^a \times \mathcal{C}^t} \log \sigma\left(\mathbf{A}\langle \boldsymbol{n}^{\boldsymbol{a}}, \boldsymbol{n}^{\boldsymbol{t}} \rangle \cdot \langle \boldsymbol{c}'^{\boldsymbol{a}}, \boldsymbol{c}'^{\boldsymbol{t}} \rangle\right)$$

Here, $\langle \boldsymbol{n}^a, \boldsymbol{n}^t \rangle$ is the concatenation of the fixed pre-trained audio and textual embeddings of a noun, and $\mathcal{C}^a, \mathcal{C}^t$ are sets of positive and negative contexts of the adjective-noun phrase. For positive contexts, we use the fixed pretrained embeddings of the actual audio and text representations of the adjective-noun phrases. For negative contexts, we fix the adjective and randomly chose a subset of nouns different from n. For example, to learn a matrix $\mathbf{A}$ for the adjective *happy*, $\boldsymbol{n}^t$ is the textual embedding of *cat* and $\boldsymbol{n}^a$ the average of all its audio vectors; $\boldsymbol{c}'^{\boldsymbol{a}}$ indexes over all the audio embeddings we have for *happy cat* and $\boldsymbol{c}'^{\boldsymbol{t}}$ is its textual embedding. For negative contexts, $\overline{\boldsymbol{c}}'^{\boldsymbol{a}}$ indexes over all the audio embeddings we have for *happy noun*, where *noun* is a random noun different from *cat*, e.g. *baby* and *car*.

AT-Joint Tensor Skipgram changes the objective function to the following, for the same $\boldsymbol{n}^t$ and $\mathcal{C}^a$ as above.

$$\sum_{\boldsymbol{c}'^{\boldsymbol{a}} \in \mathcal{C}^a} \log \sigma\left(\mathbf{A}\boldsymbol{n}^t \cdot \boldsymbol{c}'^{\boldsymbol{a}}\right) + \sum_{\overline{\boldsymbol{c}}'^a \in \overline{\mathcal{C}}} \log \sigma\left(-\mathbf{A}\boldsymbol{n}^t \cdot \overline{\boldsymbol{c}}'^a\right)$$

Here, the audio adjective is learnt from an audio-only context, but in such a way that when multiplied with the textual vector of a noun, it is forced to be closer to the audio context.

4 Implementation

We implemented an audio-text TSG, by extending the image-text TSG model of (Lewis et al., 2022) to audio data. The positive context is the audio files representing a target phrase. For instance, for *loud melody* we had 100 audio files and for *loud cat* 82. The negative context is determined by random selection of nouns during the training process with each adjective. We treat these nouns as a hyper parameter and choose them by tuning on the validation segment of the dataset.

For skipgram models, we learn 50 dimensional phrase vectors with a learning rate of 10^{-6} and a batch size of 512, trained for 200 epochs. The models were trained on NVIDIA T4 and V100 depending on their availability on Google Colab. The training was done in batches over a period of 3 months, totalling 80 hrs. We used Binary Cross-Entropy loss and the Adam optimiser in the training process to refine the performance.

5 Evaluation Tasks and Results

Our main hypothesis is that combining text and audio improves over audio-only learning. To test this, we trained audio-only variants of LR and TSG models. In these, the adjective matrices were learnt using only the audio vectors of their nouns and contexts. A second hypothesis is that non-commutative matrix multiplication models (LR and TSG) outperform simple commutative models. To test this, we implemented an additive model where an adjective's representation is added to its nouns. Finally, we hypothesise that compositional models outperform non-compositional ones. For this, we compared the results to the holistic OpenL3 audio vector of adjective-noun phrases.

5.1 Adjective Similarity

Following (Maillard and Clark, 2015), we first evaluate our methods on an adjective similarity task. Starting from the word similarity dataset SimLex-999 (Hill et al., 2015), We identified 13 sound-relevant (adj, adj) pairs with audio files in FreeSound. These pairs represent 11 out of 30 adjectives from our dataset. We call it *Simlex-Audio*. Examples are (*happy, cheerful*) and (*fast, rapid*).

5.2 Adjective-Noun Similarity

Existing adjective-noun phrase similarity benchmarks, such as (Mitchell and Lapata, 2010; Vecchi et al., 2017) were unsuitable due to limited sound relevance. This led us to develop a new audio phrase dataset.We selected frequent *audio adjectives* from the UKWaC corpus (top 1000 adjectives with at least 200 occurrences) and those with strong auditory relevance in FreeSound (800+ mentions)[2], resulting in 30 suitable adjectives, each paired with a noun. Nouns were refined grammatically and filtered to those with 100+ mentions on Freesound.

This procedure resulted in a dataset of 30 adjectives, 721 unique nouns, and 1,944 adjective-noun phrases. The number of nouns modified by each adjective varied; for example, *low* modified 46 nouns, while *quick* modified 114, with an average of 65 nouns per adjective. For audios, we selected 100 audio files per noun and on average 50 files per adjective-noun phrase, each 10-20 seconds long. The number of audio files per adjective-noun varied, e.g., *human cough* had 97 audios and *angry girl* had 45. The dataset contained 271,766 files (about 760 hours), split into 80% training, 10% testing, and 10% validation for experimentation.

5.3 Semantic and Audio Similarity Tasks

The new audio phrase dataset includes both semantic and audio similarity judgments, scored from 1 (least similar) to 5 (most similar). Annotators scored pairs based on semantic relatedness and perceived sound similarity. A pilot study with 100 randomly chosen phrase pairs and 10 annotators yielded an inter-annotator agreement of 0.45. To improve this, pairs with identical adjectives were categorized as *environmental* (e.g., *happy cat*, *loud wind*) or *musical* (e.g., *loud piano*). The data was

[2] We refer to these adjectives as audio-relevant due to their strong association with sounds.

Model	Adjective Similarities		Phrase Similarities			
	Simlex-Audio		SemPhrase		AudPhrase	
	LR	TSG	LR	TSG	LR	TSG
AT-Concat	0.731	0.755	0.762	0.856	0.779	0.876
AT-Joint	0.635	0.79	0.668	0.882	0.581	0.894
Audio-Only	0.683	0.743	0.716	0.783	0.753	0.825
ADD-Audio	0.455		0.689		0.743	
ADD-AT	0.499		0.647		0.669	
Non-Comp Audio	–		0.511		0.578	

Table 1: Similarities computed for Simlex-Audio, SemPhrase, and AudPhrase datasets. Non-Comp, ADD, LR, and TSG denote Non-Compositional, Addition, Linear Regression, and Tensor Skipgram; **AT** is Audio-Text, and **Concat** is concatenation.

arranged into forms of 10 pairs; each with only either musical or environmental phrases. 4 forms were grouped together to create 1 questionnaire.

Human Judgements: We used Amazon Mechanical Turk to collect annotations, selecting annotators with a HIT approval rate above 95% and over 1000 approved HITs. They were paid £10.42/hr. Tasks were batched with gold standards to filter automated responses, excluding unexpectedly fast annotations. To manage costs, we limited the nouns per adjective to 15-20, with 100 sound files each. This resulted in 3,144 adjective-noun pairs across 77 questionnaires, each annotated by 15 different annotators, totalling 113. Inter-annotator agreement was 0.69 for semantic similarity and 0.67 for audio similarity. We call these datasets SemPhrase and AudPhrase and they will be available on github[3].

5.4 Results

We measured the Spearman correlation ρ_s between the human annotations and cosine similarities, see Table 1 for the results. For semantic similarity and in both SimLex and our new dataset, the best performing model was the audio-text joint learning (**AT-Joint**) via TSG. The second best performing model was audio-text concatenation (**AT-Concat**) via TSG. They both improved on their LR counterparts, and outperformed the audio-only, additive, and non compositional models. In LR, only **AT-Concat** outperformed all the baselines; but itself fell short of TSG. A very similar trend was observed for the audio similarity task, where TSG applied to **AT-Joint** was the best performing model again, outperforming all baselines. The second best model was TSG with **AT-Concat**. For LR, again only **AT-Concat** outperformed the baselines.

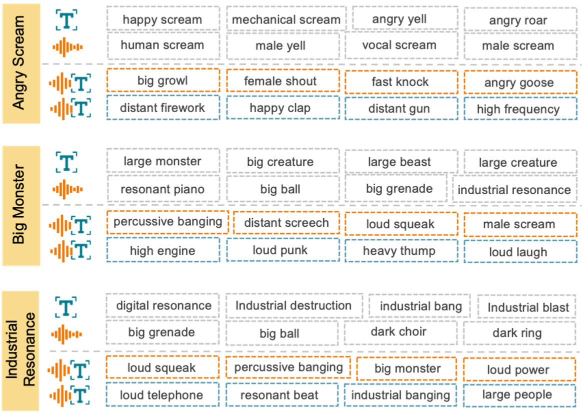


Figure 2: Query and its top 4 closely related phrases(left to right). Grey rows indicate non-comp audio and text-based similarities, while orange and blue signify similar phrases for compositional audio and semantic similarities, using AT-Joint.

6 Discussion and Conclusion

We conducted a case study to understand the better performance of compositional multimodal audio-text embeddings using k-means clustering with optimal k values determined via Silhouette method (Rousseeuw, 1987). Cosine similarities were computed within each cluster, to find closet neighbours to the random queries from the evaluation split. Some examples are provided in Figure 2. We found out that holistic singular text and audio only models predicted either semantic or audio relevance, often getting close to opposite concepts or literal sounds. On the other hand, multimodal composition managed to predict a more accurate phrase meaning. When non-compositional models struggle to predict, e.g. in the second example, the audio-only model predicted *resonant piano* and *big ball* as synonyms of *big monster*, multimodal composition predicted *loud squeak* and *heavy thump* and bridged the gap. Another example is the prediction of *distant firework*, *distant gun*, and *high frequency* for *angry scream* by multimodal composition, where a text-only model guessed the opposite, i.e. *happy scream*.

Similar is the case for *industrial resonance*, predicted to be close to *percussive banging* and *loud telephone* by the compositional model, improving over the audio-only model which predicted *big monster* and the text-only model which again predicted opposite, i.e. *industrial blast*.

These findings show that reflecting the textual grammatical structure in adjective-noun composition and considering both audio and text modalities improves the quality of audio data. Extending the setting to verb phrases is work in progress.

[3] https://github.com/audio-comp

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2461–2470.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Martha Lewis, Qinan Yu, Jack Merullo, and Ellie Pavlick. 2022. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*.

Rehana Mahfuz, Yinyi Guo, and Erik Visser. 2023. Improving audio captioning using semantic similarity metrics. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jean Maillard and Stephen Clark. 2015. Learning adjective meanings with a tensor-based skip-gram model. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 327–331.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Mark Steedman. 2002. Mark steedman, the syntactic process (language, speech, and communication). cambridge, ma: Mit press, 2000. pp. xiv 330. *Journal of Linguistics*, 38(3):645–708.

Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. 2020. Pre-training audio representations with self-supervision. *IEEE Signal Processing Letters*, 27:600–604.

Eva M Vecchi, Marco Marelli, Roberto Zamparelli, and Marco Baroni. 2017. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive science*, 41(1):102–136.

Gijs Wijnholds and Mehrnoosh Sadrzadeh. 2019. Evaluating composition models for verb phrase elliptical sentence embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 261–271, Minneapolis, Minnesota. Association for Computational Linguistics.

Huang Xie and Tuomas Virtanen. 2021. Zero-shot audio classification via semantic embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1233–1242.

Learning through gesture: embodied repetitions in tandem interactions

Loulou Kosmala
Paris-Est Créteil University (IMAGER)
loulou.kosmala@u-pec.fr

Abstract

Grounded in an interactional framework, this corpus-based study presents an analysis of multimodal tandem interactions held in English between tandem partners (L1 and L2 speakers) to study other-repetitions across different levels and modalities. In particular, I investigate cases of embodied repetitions in contexts of co-construction and repair whereby tandem partners negotiate meaning. Based on careful micro-analyses of data fragments, analyses reveal different types of temporal coordination between the repetition of the target item and/or of the gesture, addressing specific issues at different linguistic levels. While repetitions typically occur in linguistic-oriented contexts, emerging gestures may further contribute to mutual understanding and alignment.

1 Introduction

Research in Second Language Acquisition (SLA) has increasingly gained an interest in the study of gesture in L2 learning. Gestures are said to provide a window onto cognition (Goldin-Meadow, 1999), and a series of perception experiments have highlighted the facilitative role of gesture for vocabulary (Huang et al., 2019), phoneme acquisition (Hoetjes and Van Maastritch, 2020), or L2 pronunciation more generally (Gluhareva and Prieto, 2017). In particular, the present study is grounded in an interactional approach to language learning, drawing on conversation-analytic (CA) methods, and thus considers learning processes at the *heart* of face-to-face interactions. The study of *CA-for-SLA* (Pekarek Doehler 2006; Pekarek Doehler and Pochon-Berger 2011; Mondada and Pekarek Doehler 2004), for instance, has highlighted the socially situated dimension of L2 learning captured in actual language use in its natural ecology. In this respect, tandem settings (Calverts and Brammerts, 2003) are a particularly relevant context to study situated language learning, since they rely on a friendly and low-hierarchy relationship between tandem partners, as opposed to more institutional teacher-student relations, during which the interactants also engage in authentic conversations, rather than artificial perception or production tasks in experimental tasks. The present study is conducted on a selection of the SITAF Corpus (Horgues and Scheuer, 2015) which comprises English interactions between tandem partners at university (English L1 speakers and French L2 speakers) during a narrative task. Previous work on the same data has highlighted the multimodal dimension of tandem interactions, with a focus on the role of gesture in corrective feedback, communication breakdowns, fluency mechanisms, and chains of reference (Debras and Beaupoil-Hourdel, 2019) and the aim of the present study is to explore the role of gesture in L2 learning and understanding at different linguistic levels, through embodied repetitions.

2 Repetitions in L2 interaction: from speech to gesture

Speech repetition is a key aspect of L2 acquisition and has been used successfully in L2 teaching and learning, including repetitions and imitations of words and sentences (Ghazi-Saidi and Ansaldo 2017). Repetition has been regarded as a way of “providing learners greater access to language forms […] as a means of enabling learners to develop automaticity in the target language” (Duff 2000, 109). While the present work does not dwell on classroom interactions, it is relevant to note that some studies have also highlighted the collaborative and intersubjective nature of utterance repetitions during students’ joint writing assignments (e.g. DiCamilla and Anton, 1997). In

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 19–25
October 14–15, 2024. ©2024 Association for Computational Linguistics

addition, repetitions in L2 interactions are not solely associated with learning and acquisition, but also point to affective participation (Skehan 1998). In addition, several studies conducted in corrective feedback and miscommunication point to the sequential organization of repetitions (e.g. Debras et al., 2020; Horgues & Scheuer, 2017). While their main focus has not been on repetitions per se, these studies have shown a tendency for L1 speakers to repeat L2 speakers' utterances when they have trouble understanding, or to provide corrective feedback. When corrective feedback is provided, the L2 learners would also frequently provide *uptake,* i.e., a repetition of the prior target form with the correction.

When it comes to *embodied* repetitions, with the repetition of gestures in particular, previous studies also conducted in the classroom have shown that matching gestures can be used to highlight aspects of learning, as well as to display recipiency and co-participation. For instance, in a study conducted on Swedish, grounded in a conversation analytic framework, Majlesi (2015) has shown that gesture repetitions in the context of L2 learning may serve two functions: (1) to address prior actions and maintain intersubjectivity, and (2) to provide learning opportunities during correction and instructional sequences. In another study on Mexican Spanish learners of English, Eskildsen and Wagner (2015) have illustrated the joint construction of speech and gesture during collaborative picture-describing activities, with instances of repeated gestures produced with variations in pace (e.g. gestures repeated more slowly by the instructor). The authors also point to two embedded functions of gesture in these learning situations, namely displaying shared understanding, and integrating these processes of understanding.

Studies conducted outside the classroom have also described the roles of gesture as communication strategies to solve different types of linguistic problems at the lexical, syntactic, or pragmatic level (Gullberg 2011). In adult-child conversations more specifically, Graziano et al., (2011) have described examples of "parallel gesturing" (also known as "gesture mimicry" by Kimbara, (2006) or "gestural alignment" by Bergmann and Kopp, 2012, among other terms) during which the adult or the child repeats speech and gesture to display understanding or to provide corrective feedback, among other actions.

In sum, embodied repetitions, involving speech and gesture, do not only exemplify learning processes, but also point to the structural and intersubjective nature of interaction itself, whereby interactants demonstrate different forms of involvement and participation. The focus here is on tandem interactions held outside the classroom during specific learning contexts which may result from a trouble or repair sequence. Analyses will show how embodied repetitions may not only assist the learner in their target language at different linguistic levels (lexical, syntactic, phonological), but also address and resolve issues in understanding, as well as to mark recipiency and alignment.

3 Data and Method

The data under study is based on the SITAF Corpus (Horgues and Scheuer, 2015) which comprises 24 videotaped dyadic interactions between L1 and L2 speakers in English and French. The dyads were paired through a tandem program at university, and the participants regularly met outside the recording sessions to exchange in their respective L1 and L2. The selected sample comprises eight pairs (selected randomly) from the corpus, during which the participants performed a narrative task called "Liar Liar" in English. The aim of the task was for one of the participants to retell a story in which they had to insert three lies that their partner later had to identify. In this case, the stories were told in the French speaker's L2 (English). The selected sample, including the duration of the exchanges, are reported in Table 1.

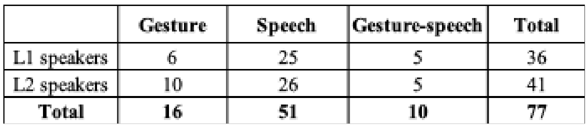
Pair 1	06:40
Pair 2	03:22
Pair 5	05:17
Pair 8	03:04
Pair 9	08:13
Pair 10	05:29
Pair 11	04:46
Pair 15	05:09

Table 1: Data sample (duration in mins)

All instances of other-repetitions (identical repetition of the interlocutor's previous word, utterance, or gesture) were categorized in the data, distinguishing between speech repetition, gesture repetition, and gesture-speech repetition. In addition, the different possible functions of these repetitions were identified, based on the SLA and gesture literature: (1) corrective feedback, (2) uptake, (3) misunderstanding, (4) confirmation, and (5) alignment.

4 Overview of the data

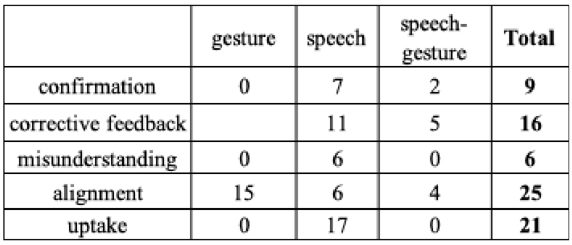
Results show a total of 77 repetitions across the 8 pairs, with a majority of speech repetitions (N=51/77) but also instances of embodied repetitions, (N=26/77, including gesture and speech-gesture repetitions).



	Gesture	Speech	Gesture-speech	Total
L1 speakers	6	25	5	36
L2 speakers	10	26	5	41
Total	**16**	**51**	**10**	**77**

Table 2: Number of repetitions across the 8 pairs

Analyses further show the different functions of these repetitions, based on the modality, as reported in Table 3:



	gesture	speech	speech-gesture	Total
confirmation	0	7	2	**9**
corrective feedback		11	5	**16**
misunderstanding	0	6	0	**6**
alignment	15	6	4	**25**
uptake	0	17	0	**21**

Table 3: Distribution of functions across modalities

While a majority of the repetitions occur in corrective feedback sequences (N=37, including 'uptake'), it is interesting to note that 25 instances are used to display alignment and understanding, with a majority being in the visual-gestural modality (through gesture and speech-gesture repetitions).

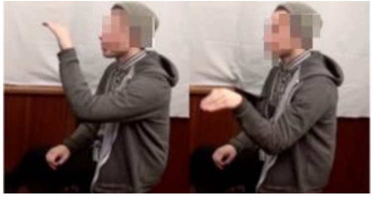
These quantitative results are further illustrated in the following data fragments, which are based on three pairs of the corpus (Pairs 8, 10, and 11). They focus on cases of embodied repetitions more specifically, as well as the linguistic levels involved (lexical, syntactic, phonological and morphological) looking more closely at the relationship between gesture and speech.

5 Illustrative cases: gesture and speech repetition

5.1 Lexical level

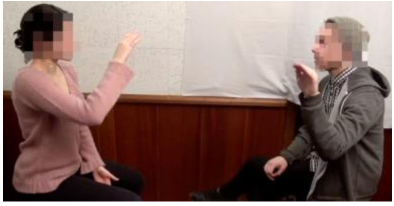
In this first example, taken from Pair 11, the L2 speaker is describing a place where she spent her winter in a castle surrounded by big hills, and explains how she and her family were stuck inside the castle for three days because of the snow.

L2: it was in December so it was really really cold.
L1: yeah yeah ((nods))
L2: and it's snowing [a lot a lot.
L1: [ok
L2: and we:e (.) we have to stay in the castle three days
L1: ok ((laughs))
L2: because there's ((both hands raised in the air to represent a pile of snow))
L2: because we're on the:e what's the word on meadow↗ ((left hand raised high up with palm down))
L1: yeah
L2: but a a big meadow so hhh. when [you're
L1: [meadow↗
L2: yeah a sort [of mead[ow↘
L1: [ok
L1: [ok ((nods))
L2: not a mountain but a little mmm between meadow and [mountains ((moves both hands up and down in alternating motions))
L1: [o::ok ok yeah yeah
L2: a:[and
L1: [like big hills↗↘ ((raises his left hand and waves it in the air to represent the hills, pic.1))



1.

L2: yeah big hills↘((produces a similar gesture in synchrony, pic. 2))



2.

L1: ok ((nods))

In this example, the L2 French speaker is experiencing lexical difficulties when describing the castle's surroundings. After explicitly displaying her ongoing search ("what's the word") she offers the word "meadow", but pronounced incorrectly *[*midoʊ]*, which seems to lead to a trouble in understanding on the L1's speaker part, who repeats the target word with a rising intonation, using the correct pronunciation. The L2 speaker then further elaborates her description of the surroundings, introducing the word 'mountain', to which the L1 speaker replies with several tokens of understanding ("ok" and "yeah"). In the subsequent turn, the L1 speaker suggests the noun phrase "big hills" introducing a novel lexical item, during which he raises his left hand in the air in a waving motion to represent the shape of the hills (pic. 1). The L2 speaker then repeats the target words and reproduces a similar gesture in synchrony (pic. 2). As previous studies have already suggested, these matching and synchronized gestures may serve two simultaneous functions: (1) to resolve the current misunderstanding and display alignment and recipiency, (2) to orient to the novel lexical item "hill" as a *learnable* (Maljesi, 2015).

5.2 Syntactic level

In this second exchange, held a few minutes after the first excerpt, the same L2 speaker is describing the insides of a car after it had been snowed in.

L2: there was interior in leather

L1: inter – oh the leather interior – so switch those words leather interior (("switch" U-shaped gesture, pic 3))

3.

L1: [not interior leather

L2 yeah yeah [because he has] a Ferrari he has leather in (.) in [the car

L1: yeah yeah [it's just what you said (.) you just switch the word ((switching hand gesture+ "two" handshape, pic 4.))

4.

L2: ok

L1: so it's leather interior not interior leather

((L2's both index fingers raised and held, followed by a similar switching hand gesture, 5.))

5.

L2: leather interior ((repeats the same switching hand gesture))

L1: yeah there you go!

In this case, the issue does not seem to be lexical, but syntactic, with the matter of word order. Unlike the previous example, the target words "leather interior" become a highlighted pedagogical focus, where the L1 speaker adopts a much more instructional posture as he takes some time to explain the switch in word order from "leather interior" to "interior leather". His "switching" gesture, produced with both hands using two fingers in alternating motions (pic. 3), is very similar to what have been labeled 'pedagogical gestures' in instructional conversations (e.g. Tellier and Yerian, 2022). It takes some time for the L2 speaker to understand this shift in tone and orient to this pedagogical sequence, and when she does, she repeats a similar switching gesture (pic.5), but produced with index fingers moving sideways in a cross. Once again, the two gestures are produced simultaneously, and once the L2 speaker provides *uptake*, i.e., repeats the correct target form with the right word order, the L1 speaker provides praise and uptake validation ("yeah there you go!").

5.3 Phonological and morphological level

In the following example (analyzed in detail in Kosmala et al., 2023) the L1 speaker adopts a similar instructional posture and explains the plural form of "geese" using his hands.

L1: you can say for (.) um there's one than more goose (.) they're geese ((right hand curved into a U shape moved to the side))

L2: geese ((stretched lips))

L1: geese yeah it changes to "ee" in the middle ((spells the vowel digraph in the air))

L2: ok yeah ((repeats a similar hand-spelling gesture))

L2: so geese ((stretched lips))

Once again, the target word 'geese' becomes a relevant pedagogical topic to which the two tandem partners jointly orient to. The L1 speaker provides both morphological and phonological explanations from the change of 'goose' to 'geese' with the shift to the plural form. He illustrates this shift with a specific U handshape (similar to the previous excerpt) and moves it from left to right. As the L2 speaker repeats the target word in a hyperarticulated way, the L1 speaker then spells the vowel digraph "ee" in the air to further illustrate a change in pronunciation. The L2 speaker then first repeats the target word without the gesture, and then reproduces the same hand-spelling gesture, perhaps to better help her visualize the word.

6 Illustrative cases: gesture or speech only repetition

These examples have shown cases of both gesture and speech repetitions in contexts of co-construction to highlight several linguistic aspects of the target language (lexical, syntactic, and phonological). The next examples illustrate cases in which the repeated elements are either speech or gesture only, following the repair initiated by the L1 speaker.

6.1 Speech-only repetition

In the next example, taken from Pair 10, the L2 speaker is talking about a dance class she had over the summer and retells a moment during which she playfully fought with her dance partner using ballet shoes.

L2: um (..) I (..) I uh ((moves her hands in space)) [!] my my ballet shoes is uh ((mimics the action of throwing something away with her right hand))

L2: I give up of my hand↗↘((frowns and looks towards her interlocutor))

L1: um ((frowns))

L2: uh we fight and ((repeats the gesture with both hands+ winces+ laughs, pic 6))

6.

L2: ((in French)) je l'ai lâché↗↘ ((repeats the same throwing-away gesture))

L1: um (.) ((looks sideways)) you let it - you let it go↗↘ ((repeats the same throwing-away gesture))

L2: yeah I let it go↘ ((places both open palms opposite her and towards her interlocutor))

In this example, the L2 speaker is demonstrably having difficulties at the lexical and syntactic levels, as she does not know how to verbally express the action of throwing one's shoes away. She first offers the structure "I give up of my hand" which the L1 speaker does not understand, and then resorts to her first language (French) to describe the action. As she does so, she repeatedly produces a sort of "throwing-away" gesture by which she mimics the action of throwing or letting go of an object in the air (pic. 6). After some delay (marked by filled and unfilled pauses), the L1 speaker provides the correct target structure ("you let it go") and repeats the same throwing-away gesture. However, when the L1 speaker repeats the target words (with a turn-initial "yeah" marking agreement), she does not repeat the same gesture, but places both her palm-up open hands opposite her and towards her interlocutor to convey her alignment and understanding. While she does use speech to repeat the linguistic target, she does not resort to gesture to do so. The repetition of the target item was thus only performed at the verbal level. The next example illustrates the opposite tendency, with the repetition of the gesture, but not of speech.

6.2 Gesture-only repetition

In this excerpt, taken from Pair 8, the L2 speaker is retelling an experience she had with a playboat (kayak) in the summer.

L2: when I was doing kayak (.) I::I uh – it was very quick in the water ((left hand reproduces the movement of the water with an open palm facing down, moving sideways))

L1: ((nods))

L2: so:o (..) so my (.) kayak uh (..) turned upside down↗ ((places both palm-up open hands then moves her right hand above her left hand facing down, pic 7)

7.

L1: flipped over↘ ((produces a similar flipping-over gesture in synchrony, pic 8))

8.

L2: yeah ((repeats the same gesture more quickly))

In this sequence, the L2 speaker is also experiencing difficulties with the description of a specific action, and resorts to a sort of “flipping-over” gesture (pic. 7) to describe her incident with the playboat. She offers the verbal phrase “turned upside down” which is immediately corrected by the L1 speaker who suggests “flipped over” instead, while repeating the gesture (pic. 8). Unlike the previous example, the L1 speaker repeats the same gesture once more in her subsequent turn, in a faster pace, but she does not repeat the target word. Instead, she produces a verbal agreement token (“yeah”). This is very similar to the cases of gestural alignment explored in previous studies (see Rasenberg, Özyürek, and Dingemanse 2020 for example) with matching gestures to display aligning responses.

The last two examples have illustrated how gestures may assist learners with the spatial description of actions in motion when they did not have sufficient knowledge of the L2 to provide the accurate verbal expressions or prepositions. While these gestures also helped the L1 speakers gain visual access to what the L2 speakers were describing, they also contributed to the overall flow of the interaction, matching the interactants’ mental representations of the event. In addition, these examples did not foreground a specific pedagogical sequence, which was not treated as relevant in these cases by both parties, but still contributed to mutual understanding.

7 Conclusion

The aim of this preliminary corpus-based study was to highlight the role of gesture in L2 interactions in contexts of repair and co-construction during other-repetitions. Even though repetitions tend to be mostly verbal and relate to linguistic content, several cases of embodied repetitions have shown that gestures may further contribute to mutual understanding and alignment. As the literature has suggested, matching gestures can be used to serve several functions, both interactional- and pedagogical-oriented to display alignment, understanding, and recipiency, or to gain access to a linguistic feature in the L2, using repetition as a way for the L2 speaker to perhaps better memorize the target words all the while being engaged in the interaction. Embodied repetitions were shown to emerge across three different types of linguistic issues, at the lexical, syntactic, morphological and phonological levels. In addition, the analyses illustrated different types of temporal coordination between the repeated elements with cases of speech- or gesture- only repetition, following the repair initiated by the L1 speaker, highlighting different types of orientations towards the learning sequence. In some cases, the gestures epitomized the pedagogical-oriented sequence initiated by the L1 speaker, leading to a joint instructional focus, while in other cases it was mostly used to display more interaction-oriented features, such as intersubjectivity, alignment, and mutual understanding. However, the number of occurrences in the data under study remains relatively limited, so more work should be done on the rest of the corpus to complement these preliminary findings.

References

Bergmann, Kirsten, and Stefan Kopp. 2012. “Gestural Alignment in Natural Dialogue.” In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34.

Calvert, Mike, and H. Brammerts. 2003. “Learning by Communication in Tandem.” In *Autonomous Language Learning in Tandem*, edited by T. Lewis and L. Walker. Sheffield: Academy Electronic Publication.

Debras, Camille, and Pauline Beaupoil-Hourdel. 2019. “Gestualité et Construction Des Chaînes de Référence Dans Un Corpus d’interactions Tandem.” *Cahiers de Praxématique*, no. 72.

Debras, Camille, Pauline Beaupoil-Hourdel, Aliyah Morgenstern, Céline Horgues, and Sylwia Scheuer. 2020. “Corrective Feedback Sequences in Tandem Interactions: Multimodal Cues and Speakers’ Positionings” in S; Raineri, M. Sekali and A. Leroux. *La correction en langue(s) – Linguistic Correction/Correctness*, 91-115

DiCamilla, Frederick J., and Marta Anton. 1997. “Repetition in the Collaborative Discourse of L2 Learners: A Vygotskian Perspective.” *The Canadian Modern Language Review* 53 (4): 609–33.

Duff, Patricia A. 2000. “Repetition in Foreign Language Classroom Interaction.” *Second and Foreign Language Learning through Classroom Interaction*, 109–38.

Eskildsen, Søren W., and Johannes Wagner. 2015. “Embodied L2 Construction Learning.” *Language Learning* 65 (2): 268–97.

Ghazi-Saidi, Ladan, and Ana Ines Ansaldo. 2017. “Second Language Word Learning through Repetition and Imitation: Functional Networks as a Function of Learning Phase and Language Distance.” *Frontiers in Human Neuroscience* 11:277215.

Gluhareva, Daria, and Pilar Prieto. 2017. “Training with Rhythmic Beat Gestures Benefits L2 Pronunciation in Discourse-Demanding

Situations." *Language Teaching Research* 21 (5): 609–31.

Goldin-Meadow, Susan. 1999. "The Role of Gesture in Communication and Thinking." *Trends in Cognitive Sciences* 3 (11): 419–29.

Graziano, Maria, Adam Kendon, and Carla Cristilli. 2011. "Parallel Gesturing in Adult-Child Conversations." *Integrating Gestures*, 89–102.

Gullberg, Marianne. 2011. "Multilingual Multimodality: Communicative Difficulties and Their Solutions in Second-Language Use." In *Embodied Interaction: Language and Body in the Material World*, edited by Jürgen Streeck, Charles Goodwin, and C. LeBaron, 137–51. Cambridge, UK: Cambridge University Press.

Hoetjes, Marieke, and Lieke Van Maastricht. 2020. "Using Gesture to Facilitate L2 Phoneme Acquisition: The Importance of Gesture and Phoneme Complexity." *Frontiers in Psychology* 11.

Horgues, Céline, and Sylwia Scheuer. 2015. "Why Some Things Are Better Done in Tandem." In *Investigating English Pronunciation*, 47–82. Springer.

Horgues, Céline, and Sylwia Scheuer. 2017."Misunderstanding as a two-way street: Communication breakdowns in native/non-native English/French tandem interactions. In *International Symposium on Monolingual and Bilingual Speech*.

Huang, Xiaoyi, Nayoung Kim, and Kiel Christianson. 2019. "Gesture and Vocabulary Learning in a Second Language." *Language Learning* 69 (1): 177–97. https://doi.org/10.1111/lang.12326.

Kimbara, Irene. 2006. "On Gestural Mimicry." *Gesture* 6 (1): 39–61.

Kosmala, Loulou, Céline Horgues, and Sylwia Scheuer. " 2023. A Multimodal Study of How Pronunciation-Induced Communication Breakdowns are Managed During Tandem Interactions." *Research in Language* 21.3: 291-312.

Maljesi, Ali Reza. 2015. "Matching gestures-Teachers' repetitions of students' gestures in second language classrooms." *Journal of Pragmatics*, 30-45.

Mondada, Lorenza, and Simona Pekarek Doehler. 2004. "Second Language Acquisition as Situated Practice: Task Accomplishment in the French Second Language Classroom." *Canadian Modern Language Review* 61 (4): 461–90.

Pekarek Doehler, Simona. 2006. "«CA for SLA»: Analyse Conversationnelle et Recherche Sur l'acquisition Des Langues." *Revue Française de Linguistique Appliquée* 11 (2): 123–37.

Pekarek Doehler, Simona, and Evelyne Pochon-Berger. 2011. "Developing 'Methods' for Interaction: A Cross-Sectional Study of Disagreement Sequences in French L2." *L2 Interactional Competence and Development* 56:206.

Rasenberg, Marlou, Asli Özyürek, and Mark Dingemanse. 2020. "Alignment in Multimodal Interaction: An Integrative Framework." *Cognitive Science* 44 (11): e12911.

Skehan, Peter. 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

Tellier, Marion, and Keli Yerian. "How to Study Pedagogical Gesture in Naturalistic Settings." In Gesture and Multimodality in Second Language Acquisition, pp. 99-123. Routledge, 2022.

Towards Automated Game-Based Early Screening for Language Disorder

Hamdan Al-Ali[1], **Elsa Soares**[2], **Gonçalo Leal**[2], **Ana Rita Valente**[3],
Nicole Agrela[2], **Alexandra Marquis**[4], **Hanan Aldarmaki**[1]
[1]MBZUAI, [2]SpeechCare Center, [3]University of Aveiro, [4]UAE University
[1]{hamdan.alali; hanan.aldarmaki}@mbzuai.ac.ae

Abstract

This paper examines the potential of gamifying early childhood language disorder screening to make the process more accessible and scalable. We provide an overview of current practices in screening and assessment, and a description of our on-going work towards automation of early screening. By integrating developmental milestones into a video game format and employing automatic speech recognition and natural language processing, this approach aims to enhance the efficiency and reach of early screening in order to identify children who need further professional assessment.

1 Introduction

Language development is a crucial aspect of early childhood development, significantly impacting future academic success and social integration (Sunderajan and Kanhere, 2019). Traditional screening methods for developmental language disorders involve one-on-one sessions that, while effective, are resource-intensive, lengthy and not easily scalable (Eriksson et al., 2010). This process, combined with the global shortage of experienced Speech-Language Pathologists (SLPs) (Squires, 2013), presents a challenge in efficiently identifying children who could benefit from early intervention on a wide scale.

Gamification, the application of game-design elements in non-game contexts, is a powerful tool that can engage and motivate children, potentially transforming the screening process into an enjoyable, playful activity. Additionally, recent advancements in technology, particularly in Automatic Speech Recognition (ASR) and Natural Language Processing (NLP), offer new potential for automating the language screening process.

We propose the development of a game that integrates these technological advancements with established developmental milestones to create a screening tool for early childhood language disorders. The proposed game aims to target children aged 3 to 4 years, a critical developmental window for identifying potential disorders (Ward, 1999). By embedding screening parameters into a game environment, we aim to gather comprehensive data on a child's language capabilities in a setting that is both natural and engaging. This approach aims to address the challenges posed by the shortage of SLPs and increase the probability of early screening by extending the reach and efficiency of screening processes from the outset.

In this paper, we describe work in progress exploring related projects at the intersection between speech-language pathology and computer science, identifying screening methodologies that are amenable to automation, and proposing game activities that have the potential to probe the target developmental milestones.

2 Communication Disorders

Communication disorders encompass a range of impairments in the ability to receive, send, process, and comprehend concepts through verbal, nonverbal, and graphic symbol systems (Fogle, 2022). A communication disorder may be evident in the processes of hearing, language, and/or speech. These disorders vary significantly in severity from mild to profound, and can be either developmental or acquired (Cooper, 2018). For the purposes of this paper, the focus will be specifically on language disorders (Owens, 2020).

Language disorders may affect different aspects of language, including phonology (the sound system of a language), morphology (the structure of words), and syntax (the arrangement of words to form sentences). They may also involve the content of language, which pertains to semantics, or the meanings of words and sentences. Furthermore, language disorders can influence the function of

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 26–31
October 14–15, 2024. ©2024 Association for Computational Linguistics

language in communication, known as pragmatics, which involves the social uses of language (ASHA, 1993). Individuals with communication disorders often face difficulties in various aspects of life (McCormack et al., 2009). Studies involving large groups of children with communication disorders have shown that they tend to have lower academic achievement, struggle more with reading, experience increased bullying, have weaker peer relationships, and encounter more psychosocial challenges than their typically developing peers (Lewis et al., 2016).

2.1 Language Disorder Screening

Early identification of language disorders in young children is crucial for timely intervention and support (Ward, 1999). This review evaluates several screening tools that, while not specifically designed exclusively for children aged 3 to 4 years, cover this critical developmental period within their scope. We examine their methodologies, effectiveness, and clinical implications based on recent studies and evaluations.

2.1.1 Northwestern Syntax Screening Test

The Northwestern Syntax Screening Test (NSST), (Lee, 1971), designed for children from 36 to 47 months, evaluates both receptive and expressive language abilities. Modifications have reduced the original test from 20 to 11 items while maintaining 95% of the variance observed in total test scores (Ratusnik et al., 1980). This revised version now requires approximately 10 minutes for administration and provides norms in six-month intervals, enhancing its sensitivity and specificity for this age group. Ratusnik et al. (1980) conducted cross-validation with a sample of 301 children, demonstrating its reliability in maintaining consistent clinical decisions across both the original and shortened versions, thus emphasizing its utility in clinical and educational settings.

2.1.2 Developmental Profile-II

The Parent Language Checklist and The Developmental Profile II (DP-II) (Alpern et al., 1980), particularly its Academic scale, serves as a parent-report tool assessing developmental milestones from birth to 7 years. The scale, when tested on 94 children between 36 and 39 months, revealed significant deficiencies in detecting developmental issues; only 21% of children with identified problems were correctly flagged. However, alternative cutoff scores suggested by Alpern et al. (1980) have shown potential in improving its diagnostic sensitivity. This tool underscores the challenges and importance of accurate parent-report measures and the need for rigorous standardization and validation to ensure reliability.

2.1.3 Minnesota Child Development Inventory

The Minnesota Child Development Inventory (MCDI), (Ireton and Thwing, 1974), offers a comprehensive assessment of various developmental domains, including expressive language and comprehension, for children from 24 to 87 months. It includes a detailed inventory that profiles eight developmental scales and provides norms based on a sample of 796 children. The results categorize development as retarded, borderline, or within normal limits, facilitating early detection of language and other developmental delays. Its extensive age range and detailed developmental scales make the MCDI a valuable tool for early childhood educators and clinicians.

2.1.4 ASHA's Developmental Milestones

American Speech-Language-Hearing Association (ASHA)'s Developmental Milestones[1] provide guidelines on expected communication and feeding skills from birth to 5 years. These milestones are intended to assist parents and professionals in identifying potential delays and initiating discussions for further assessment or referral. It is crucial for raising awareness and guiding early interventions based on observed developmental progress.

3 Use of Technology for Language Development

Various applications have made use of gamification, NLP techniques, or Machine Learning (ML) to assist with communication disorders. Some works focused on creating educational solutions, such as Sztahó et al. (2018), Bogach et al. (2021), and Prasanna and Perera (2019), which all utilized speech processing techniques and automated their evaluation processes without gamification. In contrast, work such as Lyytinen and Louleli (2023) demonstrated gamification without the use of automated evaluation or NLP techniques. Few studies focused on automating early screening for Developmental Language Disorder (DLD). For example, Rvachew et al. (2017) developed a computer-based

[1] http://www.asha.org/public/speech/development/chart/

tool for screening literacy delays but without gamification or NLP/speech processing techniques. On the other hand, the work most closely related to ours is Beccaluva et al. (2024), which introduced MARS, a web-based tool for screening DLD by engaging children in rhythmic babbling exercises to record their vocal productions, which are then analyzed using ML. They evaluated their solution on forty-seven children, 17 diagnosed with DLD and 30 with typical development (TD), collecting additional demographic information (i.e., age, gender, typicality) along with corresponding audio. After preprocessing the data, they trained models using Support Vector Machine, Random Forest, and Logistic Regression, achieving an overall accuracy of 83% in detecting DLD. Specifically, for DLD cases, they achieved 87% precision and 70% recall.

4 Proposed Method: An interactive Game for Early Language Screening

Our proposed approach aims to synthesize these technological advancements, particularly gamification, speech processing, NLP, and ML, into a comprehensive tool for early detection of language disorders. Screening for language disorders is a broad topic, encompassing various sub-categories of screening, such as semantics, morphosyntax, pragmatics, and phonology.

While earlier detection and intervention is effective[2], we focus on the age group between three and four years old, as screen use is not recommended for children younger than 24 months (American Academy of Pediatrics, 2024). Through the game, we aim to collect data that allow us to analyze the child's language performance. To do so, we base our measurement on Yang et al. (2022), which indicates that evaluating a child's language abilities can be done by assessing utterance length and complexity, as well as lexical diversity. The assessment of lexical diversity focus on the total number of different words used by the child [3], and the type-token ratio, which measures the ratio of different word types (types) to the total number of words (tokens) used, providing insight into the child's vocabulary richness and variety. More specifically, as reflected by Winters et al. (2022) and Akmeşe and Kanmaz (2021), we will analyze:

(1) measures of linguistic productivity in narratives: total number of words, utterances, and lexical diversity, **(2) global measures of narrative linguistic complexity:** average and maximum sentence length in words, **(3) measures of syntactic-semantic complexity** (Frizelle et al., 2018)**:** number and proportion of simple and complex sentences, types of complex sentences, and the diversity of adverbial clauses, **(4) maintenance of referential cohesion** (Gagarina and Bohnacker, 2022)**:** problems with nominal or verbal agreement, use of regular and/or irregular inflection, inappropriate use of tense and mood.

4.1 Transferring Requirements to a Game

Botting (2002) reflected that storytelling is one of the best ways to observe and evaluate children's pragmatic skills. Several researchers, including Akmeşe and Kanmaz (2021), Orizaba et al. (2020), and Winters et al. (2022), have analyzed language skills based on storytelling. Given these observations, we propose implementing the measurement requirements above into a game that motivates storytelling, and other side activities.

The proposed game will be level-based, with a focus on avoiding repetitive and dull levels as suggested by Lövdén et al. (2010). Each level will feature a familiar and reassuring character, which has been found effective by Vona et al. (2020). The characters will present challenges to the child (player) that require assistance. For example, in one level, the player helps a character by re-arranging story images scattered by another character. The images might include a bus, a breakfast, and an alarm. The player will sequence the images to show: the alarm rang, the student ate breakfast, and then went to school. Then, the player narrates the story and records their voice, which we process using ASR to determine the content.

In another level, a curious character asks questions such as "What is this?", "What is he/she doing?", and "Which is bigger?". The player will be tasked with answering the questions. These tasks and questions are inspired by the Speech and Language Milestone Chart [4] by LD OnLine (2024).

As the game progresses, the player will see their progress through the main menu, reflecting their performance on each level and overall progression.

[2] Ward (1999) followed up with 122 children aged between 8 to 21 months diagnosed with early language delay, and concluded that early intervention is effective at preventing language delay at 3 years old.

[3] Word categories: noun, verb, adjective, adverb, preposition, pronoun, determiner, conjunction, and interjection.

[4] https://www.ldonline.org/ld-topics/speech-language/speech-and-language-milestone-chart

To motivate the player, we will introduce stars as collectibles and other incentives. However, following American Academy of Pediatrics (2024)'s recommendation that children aged 2 to 5 only use smart devices for 1 hour per day, we will ensure that sessions do not exceed this time limit.

4.2 Evaluation Methodology

We will start by collecting data needed by SLPs to analyze the child's case. This includes a short questionnaire at the start of the game about the child's age, languages spoken, and other questions important for understanding the environmental factors that can influence language development. This information will be part of our inputs. After the child completes the game, we will retrieve and process the data using ASR and NLP techniques like Part-of-Speech Tagging and syntactic parsing, to extract additional information and gain further insight into the children's capabilities. This data will be used to measure the four key points mentioned in Section 4. By combining parent-provided data with gameplay data, we will collaborate with SLPs to identify potential signs of language disorder and label the data. In **stage one**, SLPs will perform one-on-one screening using traditional methods to create gold labels. In **stage two**, independent SLPs will assess the children using only the data collected through the game, and their performance will be compared with the gold labels to validate the game's methodology. Finally, we will train ML models to predict language delays using the collected data, focusing on high recall to improve screening coverage.

4.3 Challenges & Future Work

The journey from concept to implementation is filled with technical and operational challenges, from developing engaging and educational game content to ensuring the accuracy and reliability of the AI-driven screening tools. Effective collaboration between game developers, speech therapists, technology experts, and educational institutions, will be crucial in overcoming these difficulties.

Technical Challenges: The accuracy of the AI-driven screening model depends heavily on the quantity and quality of the data collected. Initially, gathering a sufficiently large and diverse dataset through field trials will be costly and time-consuming. The data must be carefully labeled and validated to ensure that models learn from accurate examples. In addition to screening-related data, larger data sets of children's speech will be needed to develop accurate ASR models if speech-related activities are deemed suitable for the game design. Children's speech is challenging for automatic processing due to its natural variability and shortage of data (Gerosa et al., 2009). For bilingual children, additional complexity is expected due to code-switching.

Operational Challenges: To validate the effectiveness of various aspects of the proposed game, several field trials will be needed. A sample with sufficient number of children with various developmental conditions needs to be collected for the first stage of thorough validation. This may require the administration of a large number of manual screenings to identify a sufficient number of children with language disorder. Collaboration with pre-schools and parents will be essential at this stage. Second, to provide norms for benchmarking the game's outcomes, a large number of participants from different regions and demographic segments are needed. Additional difficulties will be encountered in bilingual communities, for which both languages need to be assessed. Last but not least, collecting data involving children requires a well-defined and thorough ethical and legal framework to ensure children's protection against any potential misuse of the data.

5 Conclusion

The proposed game-based screening tool utilizes established developmental milestones to guide its design. By embedding these milestones into a game's mechanics, we ensure that each interaction within the game serves a dual purpose: to engage the child and to evaluate their language development. The use of NLP and ML methods for analyzing the data collected from these interactions aims to provide a preliminary screening that can help identify children who may require further evaluation by a specialist. This ensures that no child in need of further screening is overlooked, while maximizing the utilization of SLP time for the most likely cases of language delay. Early detection and intervention in language disorders are critical for the educational and social development of children. By providing a more accessible and appealing method for screening, we hope to increase the number of children who receive timely intervention, thereby improving long-term outcomes in their learning and communication abilities.

References

Pelin Piştav Akmeşe and Serap Kanmaz. 2021. Narrative to investigate language skills of preschool children. *International Electronic Journal of Elementary Education*, 14(1):9–22.

Gerald Alpern, Thomas Boll, and Marsha Shearer. 1980. Developmental profile ii. *J Read*, 18:287–91.

American Academy of Pediatrics. 2024. https://www.aap.org/. Accessed: 2024-05-29.

ASHA. 1993. Definitions of communication disorders and variations.

Eleonora Aida Beccaluva, Fabio Catania, Fabrizio Arosio, and Franca Garzotto. 2024. Predicting developmental language disorders using artificial intelligence and a speech data analysis tool. *Human–Computer Interaction*, 39(1-2):8–42.

Natalia Bogach, Elena Boitsova, Sergey Chernonog, Anton Lamtev, Maria Lesnichaya, Iurii Lezhenin, Andrey Novopashenny, Roman Svechnikov, Daria Tsikach, Konstantin Vasiliev, et al. 2021. Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching. *Electronics*, 10(3):235.

Nicola Botting. 2002. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child language teaching and therapy*, 18(1):1–21.

Rachel Cooper. 2018. *Diagnosing the diagnostic and statistical manual of mental disorders*. Routledge.

Mårten Eriksson, Monica Westerlund, and Carmela Miniscalco. 2010. Problems and limitations in studies on screening for language delay. *Research in Developmental Disabilities*, 31(5):943–950.

Paul T Fogle. 2022. *Essentials of communication sciences & disorders*. Jones & Bartlett Learning.

Pauline Frizelle, Paul A Thompson, David McDonald, and Dorothy VM Bishop. 2018. Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language*, 45(5):1174–1197.

Natalia Gagarina and Ute Bohnacker. 2022. A new perspective on referentiality in elicited narratives: Introduction to the special issue. *First Language*, 42(2):171–190.

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pages 1–8.

Harry Ireton and Edward Thwing. 1974. *Minnesota child development inventory*. Behavior Science Systems, Incorporated.

LD OnLine. 2024. Ld online: The educator's guide to learning disabilities and ADHD. https://www.ldonline.org/. Accessed: 2024-05-29.

Laura L Lee. 1971. Northwestern Syntax Screening Test (NSST).

Barbara A Lewis, Emily Patton, Lisa Freebairn, Jessica Tag, Sudha K Iyengar, Catherine M Stein, and H Gerry Taylor. 2016. Psychosocial co-morbidities in adolescents and adults with histories of communication disorders. *Journal of Communication Disorders*, 61:60–70.

Martin Lövdén, Lars Bäckman, Ulman Lindenberger, Sabine Schaefer, and Florian Schmiedek. 2010. A theoretical framework for the study of adult cognitive plasticity. *Psychological bulletin*, 136(4):659.

Heikki Lyytinen and Natalia Louleli. 2023. In search of finalizing and validating digital learning tools supporting all in acquiring full literacy. *Frontiers in Psychology*, 14:1142559.

Jane McCormack, Sharynne McLeod, Lindy McAllister, and Linda J Harrison. 2009. A systematic review of the association between childhood speech impairment and participation across the lifespan. *International Journal of Speech-Language Pathology*, 11(2):155–170.

Lorena Orizaba, Brenda K Gorman, Christine E Fiestas, Gary E Bingham, and Nicole Patton Terry. 2020. Examination of narrative language at microstructural and macrostructural levels in Spanish-speaking preschoolers. *Language, Speech, and Hearing Services in Schools*, 51(2):428–440.

Robert E. Owens. 2020. *Language Development: An Introduction*, 10th edition. Pearson.

V Prasanna and Indika Perera. 2019. Speakup-a mobile application to train and overcome stuttering. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, volume 250, pages 1–8. IEEE.

David L Ratusnik, Thomas M Klee, and Carol Melnick Ratusnik. 1980. Northwestern syntax screening test: a short form. *Journal of Speech and Hearing Disorders*, 45(2):200–208.

Susan Rvachew, Phaedra Royle, Laura M Gonnerman, Brigitte Stanke, Alexandra Marquis, and Alexandre Herbay. 2017. Development of a tool to screen risk of literacy delays in French-speaking children: Phophlo. *Canadian journal of speech language pathology and audiology= Revue canadienne d'orthophonie et d'audiologie*, 41(3):321–340.

Katie Squires. 2013. Addressing the shortage of speech-language pathologists in school settings. *Journal of the American Academy of Special Education Professionals*, 131:137.

Trisha Sunderajan and Sujata V Kanhere. 2019. Speech and language delay in children: Prevalence and risk factors. *Journal of family medicine and primary care*, 8(5):1642–1646.

David Sztahó, Gábor Kiss, and Klára Vicsi. 2018. Computer based speech prosody teaching system. *Computer Speech & Language*, 50:126–140.

Francesco Vona, Emanuele Torelli, Eleonora Beccaluva, and Franca Garzotto. 2020. Exploring the potential of speech-based virtual assistants in mixed reality applications for people with cognitive disabilities. In *Proceedings of the international conference on advanced visual interfaces*, pages 1–9.

Sally Ward. 1999. An investigation into the effectiveness of an early intervention method for delayed language development in young children. *International Journal of Language & Communication Disorders*, 34(3):243–264.

Katherine L Winters, Javier Jasso, James E Pustejovsky, and Courtney T Byrd. 2022. Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 65(10):3908–3929.

Ji Seung Yang, Carly Rosvold, and Nan Bernstein Ratner. 2022. Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in psychology*, 13:905789.

L2 Interactions in Heterogeneous Learner Groups during Content and Language Integrated Learning: The Experience of *Rail.lexis* and beyond

Julia Edeleva, Martin Neef, Jiaming Liu, Martin Scheidt

TU Braunschweig

y.edeleva, martin.neef, jiaming.liu, m.scheidt@tu-braunschweig.de

Abstract

Content and language integrated learning is considered a powerful tool to promote inclusion in educational settings of learners for whom the language of instruction is their additional language. Language-related difficulties of those learners have been claimed detrimental for attaining personal educational goals. Academic language places increased cognitive demands on the learning process in general due to 1) its internal complexity; 2) L2 speakers' lower proficiency; 3) their disadvantage in terms of real-time processing. Facilitators are, therefore, encouraged to integrate interactional CLIL-elements (e.g., scaffolding) during content instruction that provide the necessary pedagogical support for better understanding of disciplinary concepts and their interrelation. In the current contribution, we present the concept and first results of *Rail.lexis,* a collaborative project of the Department of German Studies and the Department of Railway Engineering at TU Brauschweig. We present and discuss several conversational arrangements (e.g., word guessing games, a differential task matrix) that were designed to engage the learners of heterogeneous linguistic backgrounds in meaningful interactions in subject-specific classes. Subject-specific tasks are gradient regarding their cognitive complexity and the background knowledge required to solve them. Therefore, the linguistic repertoire required to negotiate different task types is also differential to ensure the participation of linguistically diverse students in language-enhanced classroom interactions.

1 Introduction

A standardized language test is an essential requirement to be admitted to foreign-language study programs for learners whose preferred language deviates from the language of instruction. The language requirements for university degree studies remain quite demanding. In German universities, at least a B2 level of CEFR (Council of Europe, 2020) is required for most study programs. A B2-level language learner is described as able to "obtain information, ideas, and opinions from highly specialized sources within their field. S/he can follow the essentials of lectures, talks, and reports and other forms of academic/professional presentation which are propositionally and linguistically complex. S/he can produce clear, detailed text on a variety of subjects related to their field of interest, synthesizing and evaluating information and arguments from a number of sources" (Council of Europe, 2020). While the number of foreign applicants who fulfil the admission criteria for German-speaking study programs has been continuously increasing since 1980 (Statistisches Bundesamt, 2023), around 20 to 40% of the enrolled foreign applicants fail to attain academic goals and quit prematurely without obtaining a degree (Heublein et al., 2020). The construct of academic success is highly subjective and grounded both in individual factors and in cultural, social, and institutional integration. Language skills constitute an individual's personal profile and are subject to change over time. Wisniewski

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 32–38
October 14–15, 2024. ©2024 Association for Computational Linguistics

and colleagues (2022) point out that the actual language proficiency level of foreign students in German barely reached B2.2 when they were screened at the beginning of the study programme. With around 40% of students, the language skills fail to progress beyond their initial level during their degree studies. Simultaneously, language-related difficulties have been claimed detrimental for attaining personal educational goals. Trenkic & Warmington (2019) studied the language skills of Chinese students in relation to their content-specific academic achievement in sociology. The researchers observed that academic achievement is strongly predicted by the higher- and the lower-level linguistic processing alike. Both letter naming fluency and more complex skills such as reading comprehension accuracy were equally prognostic of academic outcomes.

Regarding the linguistic integration of L2 students, numerous preparatory and in-study courses are provided by language centers at universities. They are commonly delivered as a one-size-fit-all offer and focused on developing general academic literacy and targeted strategies for taking standardized language tests. Subject-specific vocabulary comprising basic terms and collocations for a particular field of study lies outside the language course curriculum. At the same time, it constitutes the basis of successful functioning in a technical language and ensures stable academic progress in more advanced subject-specific modules. Therefore, it appears critical to identify and implement pedagogical activities to support language growth of those learners beyond passing a standardised admissiong test. Automated dialogue systems, or collaborative conversational agents, might be practical in self-directed learrning settings (de Araujo et al., 2024). Yet the potential of collaborative conversational agents to sustain productive academic talk has been mostly restricted to operationalising talk moves that represent selected academic functions such as recapping, rephrasing, agreeing or disagreeing (de Araujo et al., 2024). The cognitive demands of the subject-specific task itself have been barely addressed to define the intervention type to be provided. Previous studies (de Araujo et al., 2024; Valle Torre et al., 2023) observed that authentic dialogue patterns may provide reliable estimations of how the learners handle the academic functions for productive discussions. The current contribution elaborates these findings adding a further dimension of cognitive task complexity. We report initial findings regarding interaction patterns delivered by students across different conversational arrangements. We argue that cognitive task complexity should be factored in to optimise the perfomance of collaborative conversational agents.

2 Language-Enhancing Tools for Content Instruction in Heterogeneous Learner Groups

Supportive methodological tools are generally beneficial to generate language-enhanced instructional settings in content-driven classes which become more inclusive for L2 learners. Yet, the design of appropriate study materials remains one of the major challenges in the implementation of content and language integrated learning (CLIL; Bouvellan, 2014). In selected CLIL-frameworks (the 4Cs Framework, Coyle, 1999, the Quadrant Matrix, Cummins, 1981; the European Framework for Teacher Education, Marsh et al., 2011; the CLIL Pyramid, Meyer, 2010) as well as independent position papers and practical guidelines (Mehisto, 2012; Morton, 2013; San Isidro et al., 2020), content represents personalized knowledge that is constructed and re-constructed in learning interaction. Further, knowledge accrual occurs in resolving cognitively complex tasks which involve higher-order cognitive processes such as thinking and reasoning. Thus, cognitive functioning represents a separate domain that undergoes gradual development in a CLIL-enhanced classroom. Importantly, it is concurrent with specific language demands required to verbalize one's reasoning patterns. Thus, the linguistic, and cognitive alignment is an important prerequisite to instantiate language-enhanced interactions during content instruction. Though context-embedded interaction is fundamental for learning to take place, researchers document low proportions of specific academic functions (e.g., hypothesizing or prediction) in classroom interactions (Dalton-Puffer, 2007).

In the current contribution, we present the concept and first results of *Rail.lexis*, a collaborative project of the Department of German Studies and the Department of Railway Engineering at TU Braunschweig (Germany). One of the goals is to produce cognitively appropriate instructional materials to instantiate meaningful classroom interactions that are conductive of knowledge (re-)construction. We also probe selected conversational arrangements as to their didactic potential to promote peer-to-peer interactions in linguistically diverse learner groups. Those goals are further developed in a follow-up project *DaF-Z mit Nachhaltigkeit* (*Sustainability in German as a Second and Foreign Language Teaching*) whose main aim is to make teacher professionalization more diversity-sensitive by providing sustainable and technology-enhanced language learning arrangements across contexts.

In the presentation, we first survey some linguistic and psychological preliminaries that generate a comprehensive framework to assess the appropriateness of classroom interactions for targeted language-enhanced content instruction. We present two different types of conversational arrangements where the interactions of students of diverse linguistic backgrounds around basic terms were instantiated during content instruction. The first one is a word guessing game and the second is a differential task matrix which incorporates activities promoting language and cognitive growth. Based on the preliminary results which will have been evaluated by the conference date we will contrast the interaction patterns of the learners in various types of word guessing games. Finally, we discuss how the interaction patterns of linguistically heterogeneous learners in subject-specific tasks of varying cognitive complexity can inform the design of collaborative conversational agents.

3 Linguistic and Psychological Preliminaries

The dialogue constitutes a core unit of language use. It represents a flexible, yet conversation-sustaining alternation between the speaker and the hearer who are cooperating in a goal-oriented way. Engaging in fruitful and high-quality peer interactions is positively associated with learning outcomes in various contexts (Asterhan & Schwarz, 2016; Stahl et al., 2014). Several design-based research attempts have emerged to identify the characteristics of effective collaborative behavior in dialogue-based activities. Thus, academic productive talk (APT, Michaels & O'Connor, 2015; Resnick et al., 2010) operates on the following accountability principles:

- The learners should build on and develop one another's ideas to sustain a goal-oriented interaction.
- The validity of the contributions should be secured via available reference materials or direct evidence.
- The learners should logically connect their arguments, evaluate their cohesion, and draw inferences.

While the learners' reasoning is prioritized over correctness, those classroom discourse frameworks do not give sufficient attention to studying specific reasoning patterns as a gateway to explicating mental models.

In Edeleva et al. (2024), we follow the procedures of cognitive task analysis (CTA, Klein & Militello, 2001). CTA is applied to work related tasks (e.g., generating a weather forecast or detecting an infection in a neonate) and represents a collection of methods to research, identify and represent the mental processes that evolve during task performance. CTA tasks are grounded in an extensive knowledge base and require complex inferences and judgements in a complex uncertain real-time environment. Proficiency-related differences of task performers will be stipulated in the strategies that they adopt to optimize their behaviour. Those differences are grounded in subject-related knowledge structures and mental models that underlie decision making and might be more elaborate in experts compared to novices.

Simultaneously, socio-cultural approaches to language ("Five Graces Group", 2009) stipulate that it is grounded in a specific socio-cultural context. Its emergence is concurrent with knowledge accrual. As knowledge is co-constructed, the learners' linguistic repertoire replenishes and becomes more diversified. They acquire disciplinary concepts as basic terms and negotiate the relations between them through academic discourse functions. Explicit reasoning in CTA-fashioned tasks will provide a window into the mental processes of students and how they employ language as a vehicle to re-organize their knowledge patterns. Those processes will be conductive of language growth proper.

4 Conversational Arrangements

4.1. Word Guessing Games

In Edeleva et al. (2024), we contrasted scaffolding patterns of L2 German students in

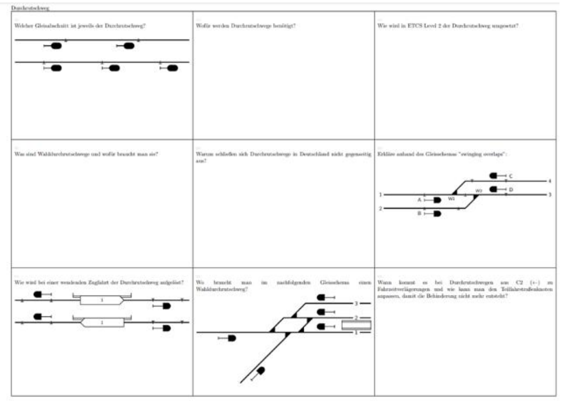
Railway Engineering in a word guessing game and a content-specific problem-solving activity. The word guessing game resembled the well-known Tabu game. The participants took turns to explain selected basic terms pertaining to the field of railway operation, albeit some intuitive explanation routes (the use of word parts or word forms, abbreviations, gestures, imitations) were eliminated by the game mechanics. The students' guessing attempts triggered meaning negotiation through linguistic adaptation. Yet, the types of scaffolds that emerged in the word guessing game differed proportionately from the strategies that emerged in a common problem-solving task. The students were less inclined to use functional and relational descriptions and embed the terms into a relevant situational context (e.g., defining initial states for a particular signal positioning). Instead, they resorted to more general factual characteristics that are contained in textbook definitions. When their initial explanation routes failed, they made use of more available prompts such as everyday meanings of the terms (e.g., *Durchrutschweg// Eng. overlap* and *rutschen// Eng. slip*). We conclude that pedagogical interventions should be equipped with supportive materials to gear the students' explanations in a more targeted way (cf. Vollmer, 2008 for similar findings).

In a follow-up study, we proceed by surveying and comparing peer-to-peer interactions in two alternative game designs. The first game is a version of a well-known "Who is the Spy?" game. The action takes place in a city where all the "citizens" receive one and the same term, the "spy" receives a related word. The "blanco" receives a blanco card without any word. Game players take turns to describe the target term. In giving their hints, they should prevent the spy from guessing the target word. After each round, the participants vote as to who they suspect to be the spy. A still other version of a word guessing game is an adaptation of "What is on my head?" where players cooperate in their word guessing attempts. The third player in a group can provide hints to steer the guessing attempts.

4.2. Adaptive Subject-Specific Tasks

We now present the differential task matrix (DTM, Figure 1) as a didactic tool that aligns the cognitive and the linguistic domain through academic discourse functions. The matrix follows the cognitive component of Bloom's taxonomy of learning (Bloom, 1956) that is originally comprised of six levels: Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation. Originally, Bloom's taxonomy was developed to rank educational objectives based on the complexity of skills and understanding. It builds on the idea that learning is ongoing and builds on prior knowledge and skills. The taxonomy ranks respective thinking skills from least to most complex along the learning trajectory. Accordingly, learning goals can be defined and learning activities can be designed. A revised taxonomy was introduced by Anderson & Krathwohl, 2001. While the original typology represents a hierarchy of educational goals, the revised typology aligns instruction, testing and assessment. It groups the cognitive operators into four knowledge dimensions:

- Knowledge of essential facts, terminology and further details that are basic to a particular discipline (factual knowledge).
- Knowledge of classification principles, theories, models, or structures pertinent to a particular discipline (conceptual knowledge).
- Knowledge of procedures and methodologies that allows the learners to modify something within a particular discipline (procedural knowledge).
- Strategic or reflective knowledge as to how to solve complex problems and tasks (metacognitive knowledge).



Durchrutschweg

Welcher Gleisabschnitt ist jeweils der Durchrutschweg?	Wofür werden Durchrutschwege benötigt?	Wie wird in ETCS Level 2 der Durchrutschweg umgesetzt?
Was sind Wahldurchrutschwege und wofür braucht man sie?	Warum schließen sich Durchrutschwege in Deutschland nicht gegenseitig aus?	Erkläre anhand des Gleisschemas "swinging overlaps":
Wie wird bei einer wendenden Zugfahrt der Durchrutschweg aufgelöst?	Wo braucht man im nachfolgenden Gleisschema einen Wahldurchrutschweg?	Wann kommt es bei Durchrutschwegen am C2 (+-) zu Fahrzeitverlängerungen und wie kann man den Teilfahrstraßenknoten anpassen, damit die Behinderung nicht mehr entsteht?

Figure 1: Example of a DTM on the topic *Overlap***.** A photocopiable verion can be found at https://zenodo.org/records/7689889

We re-defined the taxonomic relations between the knowledge dimensions and the cognitive operators to accommodate subject-specific instructional expectations and strategies and manipulated the level of thematic abstraction from individual facts and terms over structures and procedures to complex models (Greiner et al., 2019). The adjustments yielded a three-by-three matrix. Each cell contains individual tasks of varying complexity from A1 to C3. The learners have to negotiate specific problems that are framed to trigger recognition, manipulation or explication of disciplinary phenomena or states.

5 Initial Findings and Future Directions

The results of the first round of implementation (16 Civil Engineering students) show that the DTM was appraised by the students due to its practical utility for self-assessment and tracking of one's learning progress. We were interested in how the students navigate through the matrix. The learners had not been preliminarily advised about task-related differences in complexity. We observed that the hierarchy of difficulty implied in the matrix in terms of cognitive complexity and the degree of abstraction is perceived differently. In part, the preferred order in which the problems were solved was determined by their knowledge of the topic as well as subject-related competences and experiences. More expert students followed the reading direction from left to right to pick out the problem that they will be solving next. By contrast, the students with reduced subject-related proficiency were equally challenged by every problem regardless of its implied complexity level. Further on, the number of terms utilised by different learner dyads ranged from 23 to 136. The use of terms might be regarded as an approximation of the learners' available knowledge base. Thus, the DTM appears to elicit interaction patterns that discriminate between the students at different stages along their learning trajectory.

The DTM could benefit from multiple test runs and feedback loops from various learner groups to optimize relational item difficulty and achieve greater comprehensibility regarding the order in which the learners progress through the matrix. Though the primary goal of the matrix was to enhance learner interactions in content-enriched environments, particularly L2 learners whose language skills were compromised failed to engage in meaningful interactions. Those learners could be supported by additional material scaffolds (De Backer et al., 2016; Martin et al., 2019) in form of task-related prefabricated chunks, linking phrases and expressions to verbalize specific academic discourse functions. Linguistic scaffolds can also be integrated as part of the conversational agents' discourse repertoire to enhance L2 learners' linguistic development. Thus, the study delivers further compelling evidence on how technology-enhanced collaborative learning should be designed to ensure academically productive talk across different conversational arrangements. Since the DTM follows the revised taxonomy which aligns learning and assessment, the interaction patterns can also be used to develop technology-enhanced assessment tools and procedures.

Acknowledgments

The project *Rail.lexis* was financially supported from 2022 till 2023 by the Lower Saxony Ministry of Science and Culture under the "Innovation plus" funding scheme. Ongoing research is being carried out at TU Braunschweig within the project *DaF-Z mit Nachhaltigkeit* which has been granted to the first author by the Innovation in Higher Education funding agency under the Open Space funding scheme (Grant number FR-546/2023).

References

Lori W. Anderson, and David R. Krathwohl (Eds.). 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Boston, MA: Allyn & Bacon (Pearson Education Group).

Christa S.C. Asterhan, and Baruch B. Schwarz. 2016. Argumentation for learning: Well-trodden paths and unexplored territories. Educational Psychologist, 51(2): 164-187. https://doi.org/10.1080/00461520.2016.1155458

Benamin S. Bloom (Ed.). 1956. *Taxonomy of Educational Objectives: Handbook I: The Cognitive Domain.* New York: David McKay Co Inc..

Eveliina Bouvellan, E. 2014. *Teachers' beliefs about learning and language as reflected in their views of teaching materials for content and language integrated learning (CLIL).* Juväskylä Studies of Humanities. [Unpublished PhD manuscript, Juväskylä University].

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Strasbourg: Council of Europe Publishing.

Adelson de Araujo, Pantelis M. Papadopoulos, Susan McKenney, and Ton de Jong. 2024. A learrning analytics-based collaborative conversational agent to foster productive dialogue in inquiry learning. *Journal of Computer Assissted Learning*, 2024:1-15. https://doi.org/10.1111/jcal.13007

Do Coyle. 1999. Theory and planning for effective classrooms: Supporting students in content and language integrated learning contexts. In John Masih (Ed.). *Learning Through a Foreign Language: Models, Methods, Outcomes.* Lancaster: CILT, pages 46-62.

James Cummins. 1981. The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.). *Schooling and Language Minority Students: A Theoretical Framework.* National Dissemination and Assessment Center, pages 3-49.

Christiane Dalton-Puffer. 2007. *Discourse in Content and Language Integrated Learning (CLIL) Classrooms.* Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.20

Liesje De Backer, Hilde Van Keer, and Martin Valcke. 2016. Eliciting reciprocal peer-tutoring groups' metacognitive regulation through structuring and problematizing scaffolds. *The Journal of Experimental Education, 84* (4): 804-828. https://doi.org/10.1080/00220973.2015.1134419

Julia Edeleva, Martin Neef, Martin Scheidt, Gina Do Manh, and Yu Xu. 2024. Using a word guessing board game to elicit peer-scaffolds in content and language settings in Germany. In Alexiou Thomai, and Athanasios Karasimos (Eds.). *Board Games in the CLIL Classroom: New Trends in Content and Language Integrated Learning.* Berlin, Boston: De Gruyter Mouton, pages 243-262.

Ulrich Heublein, Johanna Richter, and Robert Schmelzer. 2020. *Die Entwicklung der Studienabbruchquoten in Deutschland.* (DZHW Brief 3|2020). Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) / *The development of dropout rates in Germany* (DZHW Report 3|2020). Hannover: German Centre for Higher Education and Science Research. https://doi.org/10.34878/2020.03.dzhw_brief

Five Graces Group.2009. Language is a complex adaptive system. Position paper. *Language Learning, 59* (s1): 1-26.

Franziska Greiner, Nicole Kämpfe, Dorit Weber-Liel, Bärbel Kracke, and Julia Dietrich. 2019. Flexibles Lernen in der Hochschule mit Digitalen Differenzierungsmatrizen. *Zeitschrift für Hochschulentwicklung* //Arranging flexible learning in higher education via digital differential matrices. *Journal of Higher Education Development, 14*(3): 287-302. https://doi.org/10.3217/zfhe-14-03/17

Garry Klein, and Laura Militello. 2001. Some guidelines for conducting a cognitive task analysis. In Eduardo Salas (Ed.). *Advances in Human Performance and Cognitive Engineering Research. Volume 1*. Leeds: Emerald Publishing Ltd, pages 163-199.

David Marsh, Peeter Mehisto, Dieter Wolff, and María JesúsFrigols Martín, M.J. 2011. *European framework for CLIL teacher education: Framework for the professional development of CLIL teachers*. European Centre for Modern Languages.

Nicole D. Martin, Catherine Dornfeld Tissenbaum, Dana Gnesdilow, and Sadhana Puntambekar. 2019. Fading distributed scaffolds: The importance of complementarity between teacher and material scaffolds. *Instructional Science 47*(1): 69–98. https://doi.org/10.1007/s11251-018-9474-0

Peeter Mehisto. 2012. Criteria for producing CLIL learning material. *Encuentro, 21*: 15-33.

Oliver Meyer. 2010. Towards quality-CLIL: successful planning and teaching strategies. Pulso. Revista de educación, 33: 11-29. https://doi.org/10.58265/pulso.5002

Sarah Michaels, and Catherine O'Connor. 2015. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. In Lauren B. Resnick, Christa S.C. Asterhan, Sherice N. Clarke (Eds.). *Socializing Intelligence through Academic Talk and Dialogue*. Washington DC: American Educational Research Association, pages 333-347.

Tom Morton. 2013. Critically evaluating materials for CLIL: Practitioners' practices and perspectives. In John Gray (Ed.). *Critical perspectives on language teaching materials.* London: Palgrave Macmillan, pages 111-136. https://doi.org/10.1057/9781137384263_6

Lauren B. Resnick, Sarah Michaels, and Catherine O'Connor. 2010. How (well structured) talk builds the mind. In Robert J. Sternberg, and David D. Preiss (Eds.). *From genes to context: new discoveries about learning from educational research and their applications*. New York: Springer, pages 163-194.

Xabier San Isidro-Smith, Do Coyle, and Sulushash I. Kerimkulova. 2020. *CLIL classroom practices in multilingual education in Kazakhstan*. Astana:

Nazarbayev University Graduate School of Education.

Gerry Stahl, Ulrike Cress, Sten Ludvigsen, and Nancy Law. 2014. Dialogic foundations of CSCL. *International Journal of Computer-Supported Collaborative Learning, 9*(2): 117-125. https://doi.org/10.1007/s11412-014-9194-7

Statistisches Bundesamt. 2023. Statistical report on the number of foreign students in Germany by type of educational institution and country of origin.

Danijela Trenkic, and Meeha Warmington. 2019. Language and literacy skills of home and international university students: How different are they, and does it matter? *Bilingualism: Language and Cognition, 22* (2): 349-365. https://doi.org/10.1017/S136672891700075X

Manuel Valle Torre, Catharine Oertel, and Marcus Specht. 2023. The sequence matters in learning - a systematic literature review. *LAK '24: Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 263-272. https://doi.org/10.1145/3636555.3636880

Helmut Johannes Vollmer. 2008. Constructing tasks for content and language integrated learning and assessment. In Johannes Eckerth, and Sabine Siekmann (Eds.). *Task-Based Language Learning and Teaching. Theoretical, Methodological, and Pedagogical Perspectives.* Frankfurt/M.: Peter Lang, pages 227-290.

Katrin Wisniewski, Wolfgang Lenhard, Leonore Spiegel, and Jupp Möhring (Eds.). 2022. *Sprache und Studienerfolg bei Bildungsausländerinnen und Bildungsausländern /Language and Study Success in German-as-an-Additional-Language Students.* Münster: Waxmann.

Fifty shapes of BLiMP: syntactic learning curves in language models are not uniform, but sometimes unruly

Bastian Bunzeck and **Sina Zarrieß**
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de

Abstract

Syntactic learning curves in LMs are usually reported as relatively stable and power law-shaped. By analyzing the learning curves of different LMs on various syntactic phenomena using both small self-trained llama models and larger pre-trained pythia models, we show that while many phenomena do follow typical power law curves, others exhibit S-shaped, U-shaped, or erratic patterns. Certain syntactic paradigms remain challenging even for large models, resulting in persistent preference for ungrammatical sentences. Most phenomena show similar curves for their paradigms, but the existence of diverging patterns and oscillations indicates that average curves mask important developments, underscoring the need for more detailed analyses of individual learning trajectories.

1 Introduction

The training goal of modern neural language models is simple: optimizing the prediction of the next (or a masked) token. During optimization over enormous numbers of such tokens, complex linguistic knowledge emerges as a "side effect". But how is this knowledge and its learning trajectory characterized? Existing empirical evidence seems to suggest that morphological, syntactic and basic semantic knowledge in language models is acquired quite early during pre-training, normally with a power-law like increase over the first 5-15% of the first training epoch (*inter alia* Chiang et al., 2020; Liu et al., 2021; Saphra, 2021; Müller-Eberstein et al., 2023).

However, evaluation protocols that assess concrete learning trajectories of LMs are only beginning to emerge. Current probing approaches often mask developmental difficulties by reporting averaged scores over large and varied evaluation data sets, although, as Ritter and Schooler (2001) note, "[a]veraging can mask important aspects of learning". The learning curves – plots of task performance over the training period – are frequently assessed in a purely qualitative way, with little common best practices as to which training phases and how many epochs are to be described (Viering and Loog, 2023).

In this paper, we take first steps towards a more systematic analysis of the concrete learning curves for a variety of linguistic phenomena. We train a suite of small LMs, checkpointing them logarithmically during their first training epoch, test them on the BLiMP probing suite (Warstadt et al., 2020) and compare them to recent larger LMs that provide similarly-spaced checkpoints. We analyze the resulting learning curves qualitatively (in more detail than previous research), but also quantitatively (by categorizing and clustering shapes). In doing so, we are able to discern which phenomena are easier to learn and how trajectories differ between smaller and larger language models. Moreover, we investigate whether similar phenomena also exhibit similar trajectories or whether averaged learning curves obstruct some of the underlying trade-offs and instabilities of linguistic learning in LMs.

We find that when looking at the individual BLiMP paradigms and their learning curves, a more nuanced picture of how they are (not) learned emerges. While most curves do follow the prototypical power law, completely stable curves and to a lesser degree S- and U-shaped curves are also frequent. However, many paradigms also feature ill-behaved curves that never converge to stable performance or decrease over training. Inside the broader phenomenon sets, we find sheaves of curves for those mastered earlier, whereas the curves for hardly mastered phenomena exhibit strong differences. Moreover, larger models generally converge towards more power-law curves. As such, our study puts some previous results into question – certain syntactic phenomena seem to be hardly learnable even by large language models trained on massive

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 39–55
October 14–15, 2024. ©2024 Association for Computational Linguistics

amounts of data, and even good performance at early training stages can deteriorate again after the model is confronted with more linguistic data.

2 Related work

Learning curves in ML and humans Every ML training run has a learning curve (target function or loss function over time), but these curves have not received much scrutiny and are often assumed to follow a power law, despite varying significantly depending on the task (Shalev-Shwartz and Ben-David, 2014; Viering and Loog, 2023). Viering and Loog (2023) review the variety of learning curve shapes, identifying both well-behaved, steadily increasing curves and ill-behaved curves that show degrading performance or oscillation. Well-behaved, monotonic curves are the common targets of ML research (Viering et al., 2019), often categorized using power law or exponential functions. In reality, not every learning curve is monotonically increasing. Exceptions include phase transitions with sudden performance boosts (Viering and Loog, 2023), peaks (Nakkiran, 2019), dips (Loog and Duin, 2012), and curves that oscillate through several maxima and plateaus (Sollich, 2001). Thus, the space of possible curve shapes is empirically much wider than often theoretically assumed.

Human learning can also be characterized by learning curves, abstracted to measurable performance on a task, with most empirical studies showing that human learning typically follows power laws (Ritter and Schooler, 2001). In language acquisition, some phenomena deviate from typical patterns. For example, past tense acquisition often follows a U-shaped curve: initially, children correctly produce high-frequency irregular and regular past forms item-based (Tomasello, 2000). As they abstract rules, they overregularize, applying regular rules to irregular verbs previously produced correctly (Saxton, 2017). Performance then gradually recovers to adult-like levels. Another common shape is the S-shaped logistic curve, where slow initial learning is followed by a rapid onset and then slow final gains. Examples are, e.g., vocabulary acquisition (Murre, 2014) or the production frequency of non-finite sentences (Hulk and Müller, 2000). However, evidence on the prevalence of and the complex trade-offs between such curves is still rather meagre. Due to a lack of empirical data, combined with small sample sizes, limited cross-linguistic studies, and the study of very narrowly defined phenomena, some scholars argue that these effects are much weaker than assumed (e.g. Marcus et al., 1992).

Learning trajectories in LMs In their seminal paper on neural networks learning the English past tense, Rumelhart and McClelland (1986) report U-shaped learning over one epoch of training (although this development was mostly caused by their specific re-ordering of training instances, cf. Pinker and Prince, 1988). In a modern follow-up, Kirov and Cotterell (2018) find a more oscillating pattern in their LSTM-model for past-tense acquisition, although they report scores across several epochs, which hinders comparability.

Shifting the attention to the current standard in NLP, language models, it becomes apparent that investigations into learning curves are not (yet) standard practice in evaluating language models, (primarily due to the need for fine-grained checkpointing, which only few LMs provide). The more general practice of probing over time, however, is somewhat established. Chiang et al. (2020) and Liu et al. (2021) show that when comparing a variety of probing benchmarks on masked language models, syntactic information is generally acquired earlier than semantic, pragmatic and commonsense knowledge (cf. also Saphra, 2021; Teehan et al., 2022). Besides, syntactic information is also commonly located in earlier layers of LLMs (Tenney et al., 2019). Müller-Eberstein et al. (2023) analyze multiple checkpoints of the MultiBERT LMs (Sellam et al., 2022). They also find that morphological and syntactic structure is acquired very early by the models (after ~10% of the first training epoch), whereas semantic, pragmatic and general world knowledge emerge later. Their logarithmically-scaled curves still exhibit interesting, mostly S-shaped curves with a rapid take-off after a period of little learning. This is also in line with Chen et al. (2024), who find a sudden drop in training loss in masked LM training which aligns with the emergence of syntactic attention structure in attention heads.

Turning to the focus of our experiments, minimal pair tests, several additional empirical studies can be reported. Huebner et al. (2021) derive their own "Zorro" benchmark from BLiMP by excluding phenomena not found in child-directed speech. They test an extremely small (5M parameters) masked LM and show that, generally, scores improve across

	Param.	Train. tokens	Hddn. layers	Attn. heads	Embed. size	BLiMP score
baby_llama	2.97M	10M	8	8	128	64%
teenie_llama	2.97M	100M	8	8	128	67%
weenie_llama	11.44M	10M	16	16	256	67%
tweenie_llama	11.44M	100M	16	16	256	71%
pythia-14m	14M	300B	6	4	512	65%
pythia-70m	70M	300B	6	8	512	75%
pythia-160m	160M	300B	12	12	768	79%
pythia-410m	410M	300B	24	16	1024	82%
pythia-1b	1B	300B	16	8	2048	82%
pythia-1.4b	1.4B	300B	24	16	2048	82%

Table 1: Model hyperparameters of our self-trained llama models and the compared pythia models

training. They mostly show power law-like development, with the greatest improvements occurring in early stages of training. Yet, this does not apply to all included phenomena – some are never learned well (e.g. island effects or anaphor agreement). These show diminishing accuracy after early performance peaks – a fact not further discussed.

Liu et al. (2021) also examine BLiMP development during the training of a masked LM and find that their curves, which categorize phenomena more coarsely, converge to stable performance quickly, approximating power-law curves after about 20% of pre-training. Morphological and short-distance syntactic phenomena are mastered fastest, while more complex syntactic aspects, like island effects, take longer. This pattern holds for other linguistic probes, but benchmarks testing common sense or reasoning exhibit unstable behavior with oscillating curves and performance dips. Choshen et al. (2022) take a similar approach with autoregressive LMs (GPT-2, TransformerXL). They find that grammatical phenomena are acquired in a stable order along classical linguistic layers. However, not all curves show monotonic improvement; some syntax and morphology paradigms never reach stable performance and deteriorate over training. This behavior is consistent across different initializations of both architectures but does not apply to phenomena involving semantic knowledge.

3 Methods

3.1 Investigated models

We analyze two different model architectures, four self-trained llama models (Touvron et al., 2023a) and six models from the pythia family (Biderman et al., 2023).

Data We train our models on the BabyLM data set (Warstadt et al., 2023). It features written and spoken source corpora that span a wide range of registers – child-directed speech/text, adult conversations, movie dialogue, and data from Wikipedia and Project Gutenberg. Before training our models, we clean the data from artefacts, adapting scripts by Timiryasov and Tastet (2023). The pythia models are trained on The Pile (Gao et al., 2020), a 300B token corpus sourced from the internet, academic literature, code from GitHub and, to a lesser degree, spoken language, which makes it more comparable to regular LLM training corpora.

Models and training hyperparameters We use the `transformers` library (Wolf et al., 2020) to train four different llama models[1] (Touvron et al., 2023b). Our smallest model we call `baby_llama`. The larger models are differentiated by more **t**okens (100M instead of 10M for `teenie_llama`), more **w**eights (11.44M instead of 2.97M for `weenie_llama`) or both in the case of `tweenie_llama`. As training hyperparameters, we chose a batch size of 16, 200 warmup steps, and a learning rate set to 3e-4 in accordance with Touvron et al. (2023a). From the pythia suite of GPT-NeoX models (Andonian et al., 2023), we take the six smallest models, ranging from 14M to 1.4B parameters. They were all trained on the same data, but with different model hyperparameters (cf. Table 1). Our models were trained on a single NVIDIA RTX A4000 GPU, contrasting with the pythia models trained on clusters of 32–64 GPUs.

3.2 BLiMP performance

We test BLiMP performance with `lm-eval-harness` (Gao et al., 2022). By calculating perplexity for the sentences in each

[1] Available at https://huggingface.co/bbunzeck

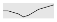
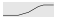
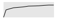
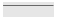
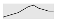
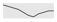
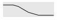
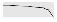
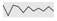
	Shape	Graphical	Description
Well-behaved	U		Medium performance followed by a dip, then rapid improvement and stabilization
	S		Initially no learning, then rapid onset and finally stabilization
	Pow		Rapid early learning, followed by stabilization and no further gains
	Stable		No change in performance across training (standard deviation < 0.2)
Ill-behaved	InvU		Inverse U-shape, stabilization after a performance peak and subsequent decrease
	RevU		Dip in performance, stabilization on lower level than before dip
	RevS		Reversed S-curve, early performance is good, but then diminishes rapidly and never recovers
	RevPow		Reverse power-relationship – performance degradation at end of training
	Osc		Performance never stabilizes and jumps between better and worse scores

Table 2: Overview of proposed curve shapes

pair, BLiMP can be used to discern whether a grammatical sentence is preferred by an LM (less perplex): an accuracy of 50% equals the random baseline. BLiMP covers 12 different linguistic phenomena from morphology, syntax and semantics (or their interfaces), with 67 included paradigms (individual data sets). We deliberately chose BLiMP due to the its widespread use and the wealth of previous results, although it suffers from problems like semantically implausible sentences (see (Vazquez Martinez et al., 2023) for more criticism and alternative data sets).

3.3 Analyzing learning curves

In line with Viering and Loog (2023), our learning curves are based on performance changes over precisely one training epoch. This choice allows us to capture the learning potential from the data upon initial exposure and observe trajectories as models encounter new data continuously. Recognizing that many linguistic phenomena are acquired early in training, we look at logarithmically spaced evaluation checkpoints: 10 checkpoints within the first 10% of training and 9 additional checkpoints until the epoch's completion.

Assessing the shapes of learning curves systematically is a complex task. We qualitatively assign shapes, aided by fitting fifth-degree polynomials to each curve. Our categorization includes well-behaved curves, such as S-shaped, U-shaped, and power law curves, as observed in the acquisition literature. We also identify ill-behaved curves by their inverted (mirrored on the x-axis) and reversed (mirrored on the y-axis) variants. Additionally, curves that remain stable from the earliest training steps (standard deviation < 0.02) are considered well-behaved, whereas curves that oscillate continuously without stabilizing are deemed ill-behaved. A summary of our systematization is provided in Table 2.

To further examine similarities between models and paradigms, we define a feature vector for each model-paradigm combination by computing the performance differences between all successive pairs of training steps across all BLiMP paradigms. This allows us to represent each model-paradigm combination as a point in a high-dimensional vector space.

4 Results

BLiMP After one training epoch, our `baby_llama` achieves a general BLiMP accuracy of 64%, improving to 67% with more data (`teenie_llama`) or more parameters (`weenie_llama`). The combination of both (`tweenie_llama`) reaches 71%. The smallest pythia model (14M parameters), despite being trained on much larger datasets, only achieves 65%, increasing to 75% for the 70M model and 79% for the 160M model. The largest pythias (410M, 1B, 1.4B) all reach 82%, close to peak BLiMP performance reported in the original BLiMP paper (83% by GPT-2), the highest score on the HELM evaluation database (84%, Liang et al., 2023), and the best BabyLM model (86%, Warstadt et al., 2023). Therefore, our results are comparable to even larger models. As an ablation, an untrained llama model performed similarly to the random baseline, scoring 51%.

Variation in phenomenon-averaged curves In the spirit of earlier analyses, we first consider the averaged curve shapes (over the phenomenon sets in BLiMP) from a qualitative viewpoint (see Figure 1, which contrasts the smallest and largest models investigated). For the smallest llama model, the learning curves exhibit a range of shapes, including power-law curves, S-shaped curves and U-shaped curves. Many curves do not show any improvements over the training epoch. The first 10% of training is marked by the highest degree of variation, but many performance gains also happen later than that. Here it already becomes apparent that

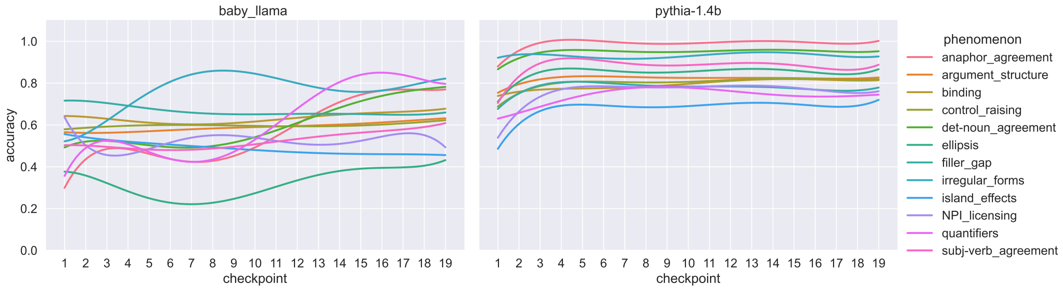


Figure 1: Learning curves over one epoch for the smallest llama and largest pythia model, averaged across BLiMP phenomena (for both models, the first ten checkpoints correspond to the first 10% of training, the following nine to remaining 90%)

more training on natural language data does not linearly improve performance on linguistic probing tasks. The largest pythia model, in contrast, displays learning curves that mostly resemble power law curves. Improvements are concentrated within the first 4-5% of training, after which the performance remains relatively stable across all phenomena (although minor performance tops and dips are observable).

We present a more detailed visualization of individual curves, categorized by phenomenon and model, in Figure 2.

Individual curves are frequently ill-behaved The first striking observation is found in the many ill-behaved curves. For the llama models, more than a quarter of the learning curves are ill-behaved, while for the pythia models, this is true for more than one fifth (distributions found in Table 3). While larger models generally do distinguish more minimal pair paradigms effectively, those phenomena that exhibit unstable and erratic curves in smaller models frequently continue to do so in larger models. Apart from that, smaller models have a higher number of curves that remain well below 50%, indicating that for some phenomena, these models actively prefer the ungrammatical variant. This issue occurs only sporadically in larger models, for example with selected paradigms concerning quantifiers or filler-gap phenomena.

Patterns inside phenomenon sets: sheaves and divergence Another striking aspect visible in Figure 2 is that sheaves of curves – curve sets that are close across training and show very similar shapes – are found across all models for different phenomena (e.g. argument structure and determiner-noun agreement). They become more power-law-like as the models increase in size. Apart from sheaves, we can also find diverging patterns, where some curves inside one phenomenon show strong improvements and other curves exhibit deteriorating performance over one training epoch, for example with subject-verb agreement and filler-gap phenomena. Such diverging patterns are more prevalent in smaller models, where they often appear as almost perfectly mirrored curves. In larger models, divergent patterns are less pronounced, but for phenomena prone to divergence, some curves still tend to worsen in the largest models.

The effects of model and data size The relationship between model size and performance is not straightforward. Our llama models scale in both parameters and dataset size, while the pythia models only scale in parameters but are trained on significantly more tokens. This increased amount of data results in less granularity in our analysis. However, the smallest pythia model, with few parameters but a large amount of training data, exhibits many S-shaped curves across several phenomena (binding, determiner-noun agreement, filler-gap, etc.). Its curves show a pronounced sudden take-off in BLiMP performance after being trained on many more tokens compared to the llama models. Thus, the amount of training data alone does not correlate with good performance after relatively few training steps.

	llama models	pythia models
Ill-behaved	27.24%	22.39%
Power law	33.21%	45.77%
S-shaped	12.32%	13.18%
Stable	14.93%	14.67%
U-shaped	12.32%	3.98%

Table 3: Percentage of curve types for both model families

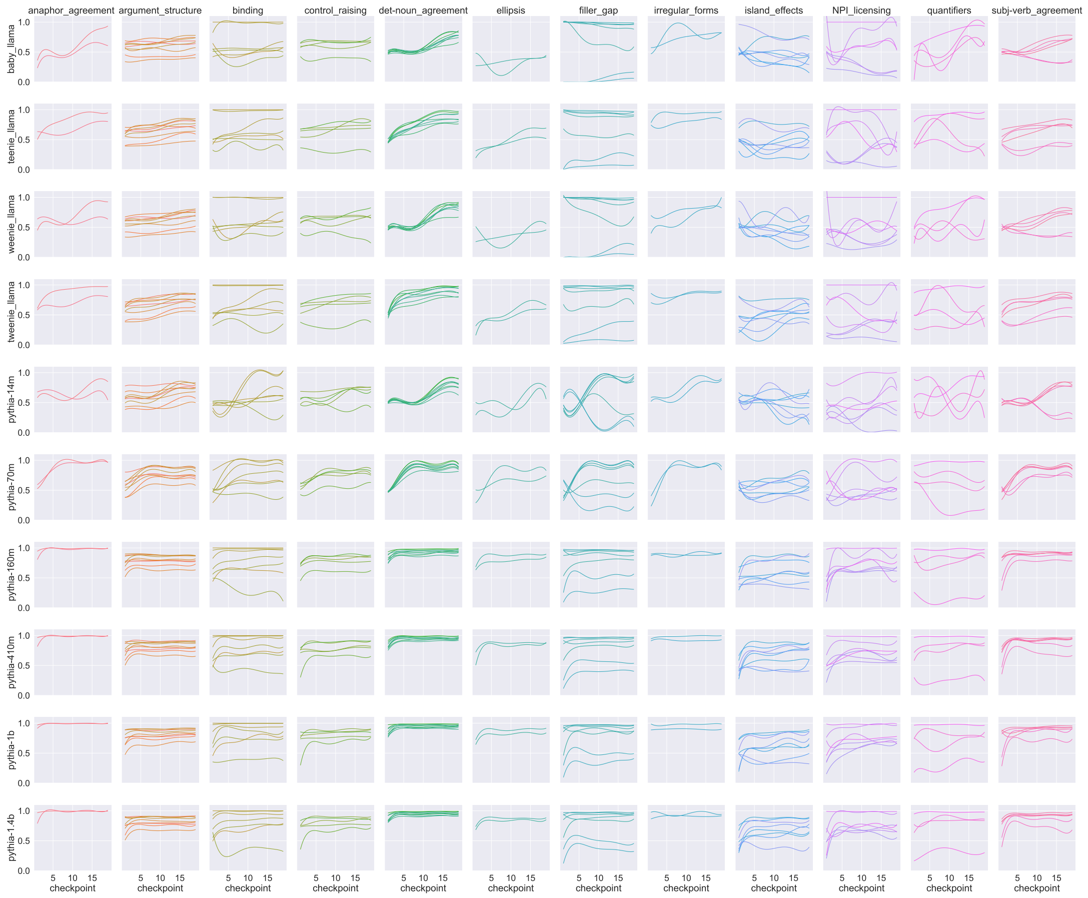


Figure 2: Learning curves for all paradigms in BLiMP, separated for models (rows) and phenomenon sets (columns)

When quantitatively comparing the distributions of curves among all investigated models, a clear division emerges between llama and pythia models. This is visualized in Figure 3. The division is particularly pronounced in larger pythia models with 160 million or more parameters. Llama models and the two smaller pythia models have a higher proportion of U-shaped and S-shaped curves, with the pythia-14M showing up to 47.8% S-shaped curves. These models also display more ill-shaped patterns, ranging from 25% to 40%. Llama models trained on larger datasets have over 40% power-law curves, whereas the smallest pythia model shows no power-law curves. Larger pythia models have over half of their learning curves following a prototypical power-law pattern, with a significant number of stable curves (20-25%), few U-shaped developments, and no S-shaped developments. Additionally, they exhibit fewer ill-behaved curves (15-20%) compared to smaller models. The four largest pythia models show very little variation, further highlighting this distinction.

Clustering training trajectories To assess further commonalities between paradigms or models, we visualize the developmental trajectory vectors, reduced in dimensionality using t-SNE (van der Maaten and Hinton, 2008), in a scatter plot (Figure 4) and visually examine whether they form specific clusters. Initially, the plot presents a messy picture with little visible structure. Clustering effects for different models or model architectures appear rather weak. However, clustering effects are more pronounced for BLiMP phenomena. We observe clusters for argument structure, determiner-noun agreement, and subject-verb agreement—phenomena that typically form sheaves. Additionally, NPI licensing, binding, and filler-gap phenomena also cluster, even though their curve shapes are quite varied. Conversely, there are no discernible patterns for phenomena like quantifiers or irregular forms.

Turning points across training The diverging mirrored curves described earlier in Section 4 also indicate another pattern: the minima for many paradigms coincide with the maxima for others.

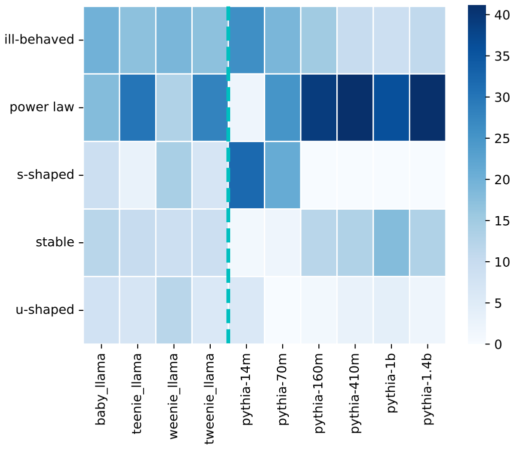


Figure 3: Co-occurrence frequencies between models and curve shapes (darker rectangles indicate higher frequency), the dashed line separates llama and pythia models

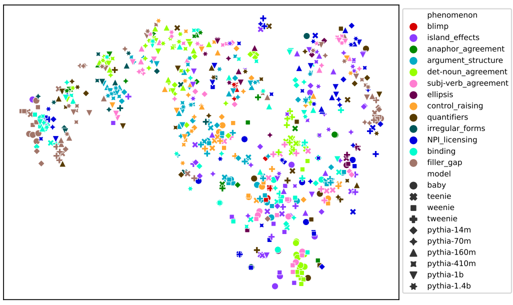


Figure 4: Dimensionality-reduced scatterplot of curve development for each model-paradigm combination

The point plots in Appendix B show checkpoint-wise deviations from mean performance, revealing particularly strong positive and negative deviations at certain checkpoints. The effects are especially pronounced in Figure 10, where almost all NPI paradigms show their maximum performance in the last few checkpoints, except for `only_npi_licensor_present`, which has deteriorated from earlier maxima in the first 10% of training

Key results From our qualitative and quantitative analyses, the most striking observations can be summarized as follows:

- Ill-behaved curves occur across all models, though they are less frequent in larger models with more internal parameters. When looking at non-averaged curves, these ill-behaved developments are much more pronounced-

- For many phenomenon-model combinations, the curves for related paradigms emerge as similarly shaped sheaves of individual curves. This is particularly true for, e.g., argument structure or determiner-noun agreement.

- In contrast to the aforementioned sheaves also diverging patterns are observed within phenomena. Some paradigms within the same phenomenon have mirrored learning trajectories, where improvement in one paradigm is directly correlated with diminishing performance in another. This divergence is particularly pronounced for filler-gap phenomena, as well as in subject-verb agreement and binding.

- Shape-wise similarities are more pronounced for phenomena across different models, whereas (especially for the smaller models) there is high variation within models.

5 Discussion

Our results indicate that larger models perform better, exhibiting higher BLiMP scores, fewer ill-behaved curves, and more power-law curves, aligning with existing literature on scaling dynamics (Warstadt et al., 2020, 2023). In our self-trained llama models, improvements are seen both with increased parameters and more data, with the combination leading to even greater enhancement. Interestingly, the smallest pythia model, despite being trained on significantly more tokens compared to the llama models, performs worse and has the most S-shaped curves. This suggests that in the very small pythia model, the real learning of linguistic features only begins after a large number of tokens are seen, whereas in our smaller llama models, this learning occurs much earlier. A possible explanation for this discrepancy could be the higher quality of the datasets used to train our llama models (BabyLM 10M and 100M), which offer a wider variety of genres and registers, compared to the web-sourced "The Pile" dataset used for all pythia models.

Our findings largely confirm, but also revise and expand upon, earlier reports of rapid syntax learning in language models. Many phenomena are acquired quickly (as in Liu et al., 2021, also Müller-Eberstein et al., 2023), yet some BLiMP paradigms are never fully mastered as in Huebner et al., 2021 for Zorro or Choshen et al., 2022 for BLiMP). The learning trajectories are non-linear; more tokens do not necessarily improve performance. Phenomena exhibit various curve shapes – some start strong, dip, and stabilize, while others oscillate indefinitely. Even in the largest models, certain phenomena remain unlearned, showing a persistent preference for ungrammatical sentences. This aligns with literature identifying these phenomena as difficult to learn, often displaying unusual learning curve patterns. Easily learned phenomena have organized sheaves of curves, while hard-to-learn phenomena exhibit scattered individual curves, suggesting that phenomena based on similar linguistic features are not uniformly grounded in the same ML features.

Some peculiarities found in our analyses might be caused by BLiMP itself. For example, the `principle_A_case_1` paradigm, which exhibits almost only stable and perfect learning curves, always features a possessive pronoun (e.g. *her*) in the grammatical sentence and a reflexive (e.g. *herself*) in its ungrammatical counterpart. However, possessives are much more frequent in language than reflexives (e.g. *her*: 1.517.948 tokens vs. *herself* 56.741 tokens in ukWaC, Baroni et al., 2009), so it is reasonable to assume that a sentence containing a reflexive always has a higher perplexity. For a randomly shuffled training corpus that is representative and balanced (in the sense of Stefanowitsch, 2020, 28), these patterns should be learned very quickly from little data and thus have such a stable learning curve, whereas other phenomena that are less tied to frequence differences might not use such easy surface heuristics. Similar criticisms, e.g. about problems with the quality of example sentences, have been put forward by, *inter alia*, Vazquez Martinez et al. (2023).

An ML-based explanation for such peculiarities is that models pick up orthogonal features – features that improve performance on some paradigms within a phenomenon but degrade performance on others – during the learning process (Choshen et al., 2022). It remains open whether ML features must necessarily correspond to those considered important in linguistic theory. The presence of mirrored curves/turning points also supports the hypothesis of orthogonal features.

Finally, BLiMP's choice of target phenomena is heavily influenced by generative, syntax-centric linguistics. Other contemporary linguistic theories (e.g. usage-based linguistics, construction grammar) might not find these phenomena particularly meaningful. In construction grammar, argument structure is determined by constructional patterns, allowing verbs to take new arguments and convey new meanings (Goldberg, 2013). Therefore, perfect performance on BLiMP may not necessarily be a desirable goal, as it might not reflect the flexible and creative language use characteristic of humans. Additionally, grammaticality is a contested notion, difficult to measure, often gradient, and strongly influenced by socio-cultural factors (Vogel, 2018, 2019). Consequently, stable curves might only be desirable for phenomena that exhibit less gradience in human evaluation, whereas worse scores and eternal oscillations might entail better linguistic generalizations for less clear-cut paradigms.

6 Conclusion

Our study set out to characterize linguistic learning in language models through an analysis of learning curves. We conclude that while the rapid syntax learning assumption from earlier studies generally holds, it also needs revision. When averaging across many phenomena and paradigms, performance gains appear to follow a prototypical power law. However, this is not true when examining individual phenomena, many of which exhibit ill-behaved curves. Stability in BLiMP performance is often an illusion; stable average curves are based on oscillating and heavily changing minimal pair paradigms within them. With larger models and more data, there is a general shift towards greater stability and more power law curves, but even in very large models, not everything works perfectly.

On a meta-level, our study demonstrates that analyzing learning curves is a powerful tool for better characterizing learning processes. Many benchmarks include systematically organized subphenomena, and our methodology can illuminate specific performance developments and complex trade-offs during the learning process. This highlights the need for the community to develop best practices for reporting learning curves, categorizing their shapes, and determining the appropriate granularity for analysis across one or several epochs. Researchers should be cautious with their interpretations, as the complexity and variety of learning curves suggest a more nuanced approach is necessary.

Future work could expand on our findings by exploring how controlling for distributions of linguistic data, like Wei et al. (2021) describe, changes the curves and learning success, which would further enhance our understanding of language model learning dynamics in a more restricted setting.

Acknowledgements

We thank Simeon Junker and Henrik Voigt for their helpful comments on earlier drafts of this paper.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A02.

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. GPT-NeoX: Large scale autoregressive language modeling in PyTorch.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Preprint*, arxiv:2304.01373.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained Language Model Embryology: The Birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *Preprint*, arxiv:2101.00027.

Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Fazz, Niklas Muennighoff, Thomas Wang, Sdtblck, Tttyuntian, Researcher2, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, Kkawamu1, Xagi-Dev, Uyhcire, Andy Zou, Ben Wang, Jordan Clive, Igor0, Kevin Wang, Nicholas Kross, Fabrizio Milo, and Silentv0x. 2022. EleutherAI/lm-evaluation-harness: V0.3.0. Zenodo.

Adele E. Goldberg. 2013. Argument Structure Constructions versus Lexical Rules or Derivational Verb Templates. *Mind & Language*, 28(4):435–465.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Aafke Hulk and Natascha Müller. 2000. Bilingual first language acquisition at the interface between syntax and pragmatics. *Bilingualism: Language and Cognition*, 3(3):227–244.

Christo Kirov and Ryan Cotterell. 2018. Recurrent Neural Networks in Linguistic Theory: Revisiting Pinker and Prince (1988) and the Past Tense Debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Preprint*, arxiv:2211.09110.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing Across Time: What Does RoBERTa Know and When? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Loog and Robert P. W. Duin. 2012. The Dipping Phenomenon. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Georgy Gimel'farb, Edwin Hancock, Atsushi Imiya, Arjan Kuijper, Mineichi Kudo, Shinichiro Omachi, Terry Windeatt, and Keiji Yamada, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626, pages 310–317. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gary F. Marcus, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu, and Harald Clahsen. 1992. Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development*, 57(4):i.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace chronicles: How linguistic information emerges, shifts and interacts during language model training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.

Jaap M. J. Murre. 2014. S-shaped learning curves. *Psychonomic Bulletin & Review*, 21(2):344–356.

Preetum Nakkiran. 2019. More Data Can Hurt for Linear Regression: Sample-wise Double Descent. *Preprint*, arxiv:1912.07242.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

F.E. Ritter and L.J. Schooler. 2001. Learning Curve, The. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 8602–8605. Elsevier.

David E. Rumelhart and James L. McClelland. 1986. On Learning the Past Tenses of English Verbs. In *Parallel Distributed Processing*, volume 2, pages 535–551. MIT Press, Cambridge, MA.

Naomi Saphra. 2021. Training dynamics of neural language models.

Matthew Saxton. 2017. *Child Language: Acquisition and Development*, 2nd edition edition. SAGE, Los Angeles.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2022. The MultiBERTs: BERT Reproductions for Robustness Analysis. *Preprint*, arxiv:2106.16163.

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*, 1 edition. Cambridge University Press.

Peter Sollich. 2001. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Anatol Stefanowitsch. 2020. *Corpus Linguistics: A Guide to the Methodology*. Language Science Press, Berlin.

Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. Emergent Structures and Training Dynamics in Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 146–159, virtual+Dublin. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby Llama: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 251–261, Singapore. Association for Computational Linguistics.

Michael Tomasello. 2000. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4(4).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arxiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arxiv:2307.09288.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Hector Javier Vazquez Martinez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating Neural Language Models as Cognitive Models of Language Acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Tom Viering and Marco Loog. 2023. The Shape of Learning Curves: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819.

Tom Viering, Alexander Mey, and Marco Loog. 2019. Open problem: Monotonicity of learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3198–3201. PMLR.

Ralf Vogel. 2018. Sociocultural determinants of grammatical taboos in German. In Liudmila Liashchova, editor, *The Explicit and the Implicit in Language and Speech*, pages 116–153. Cambridge Scholars Publishing.

Ralf Vogel. 2019. Grammatical taboos: An investigation on the impact of prescription in acceptability judgement experiments. *Zeitschrift für Sprachwissenschaft*, 38(1):37–79.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–6, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency Effects on Syntactic Rule Learning in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Learning curves for all paradigms

In consideration of legibility and brevity, detailed plots in the appendix are provided as downsized vector graphics. Interested readers may zoom in for finer detail and further examination.

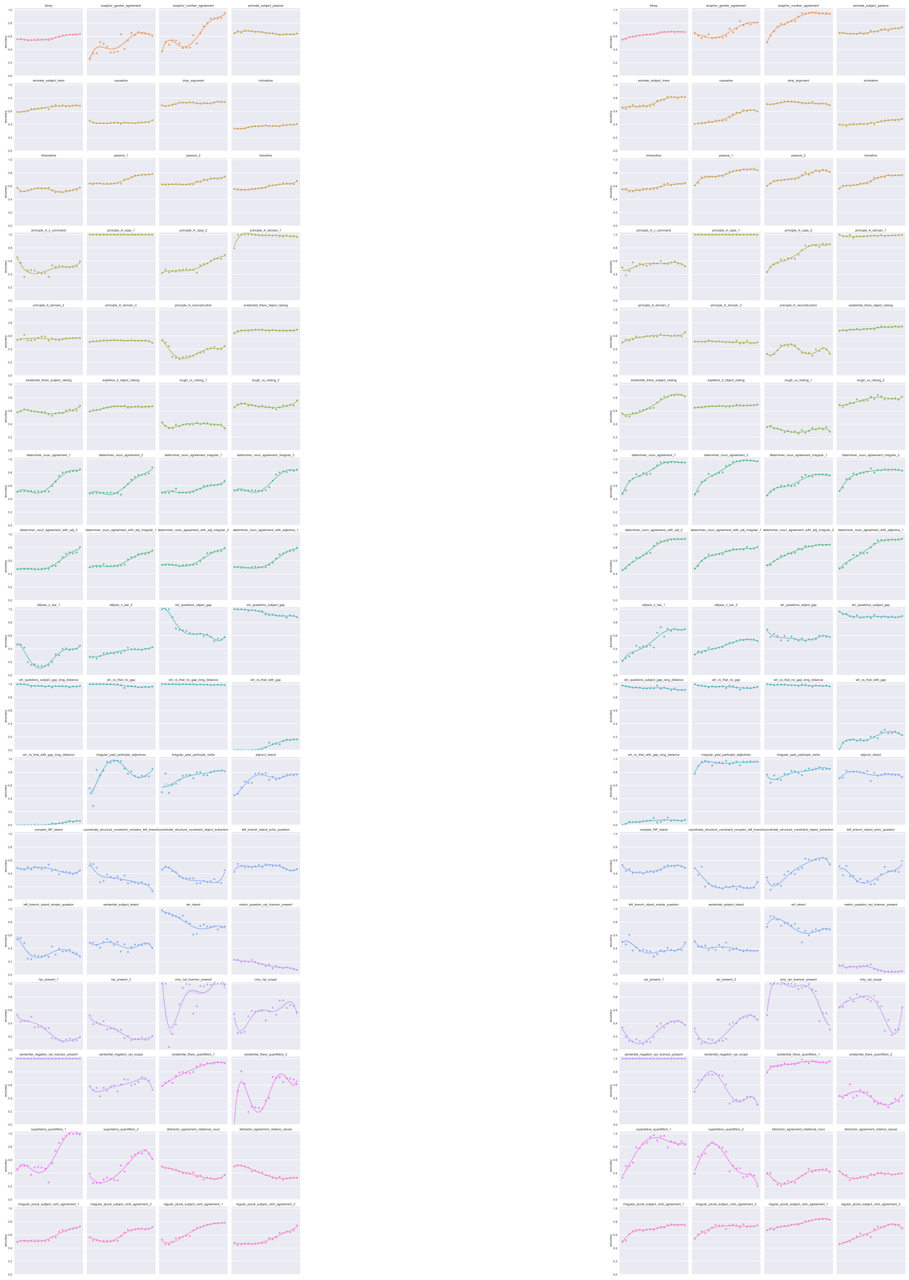


Figure 5: Learning curves for baby_llama and teenie_llama

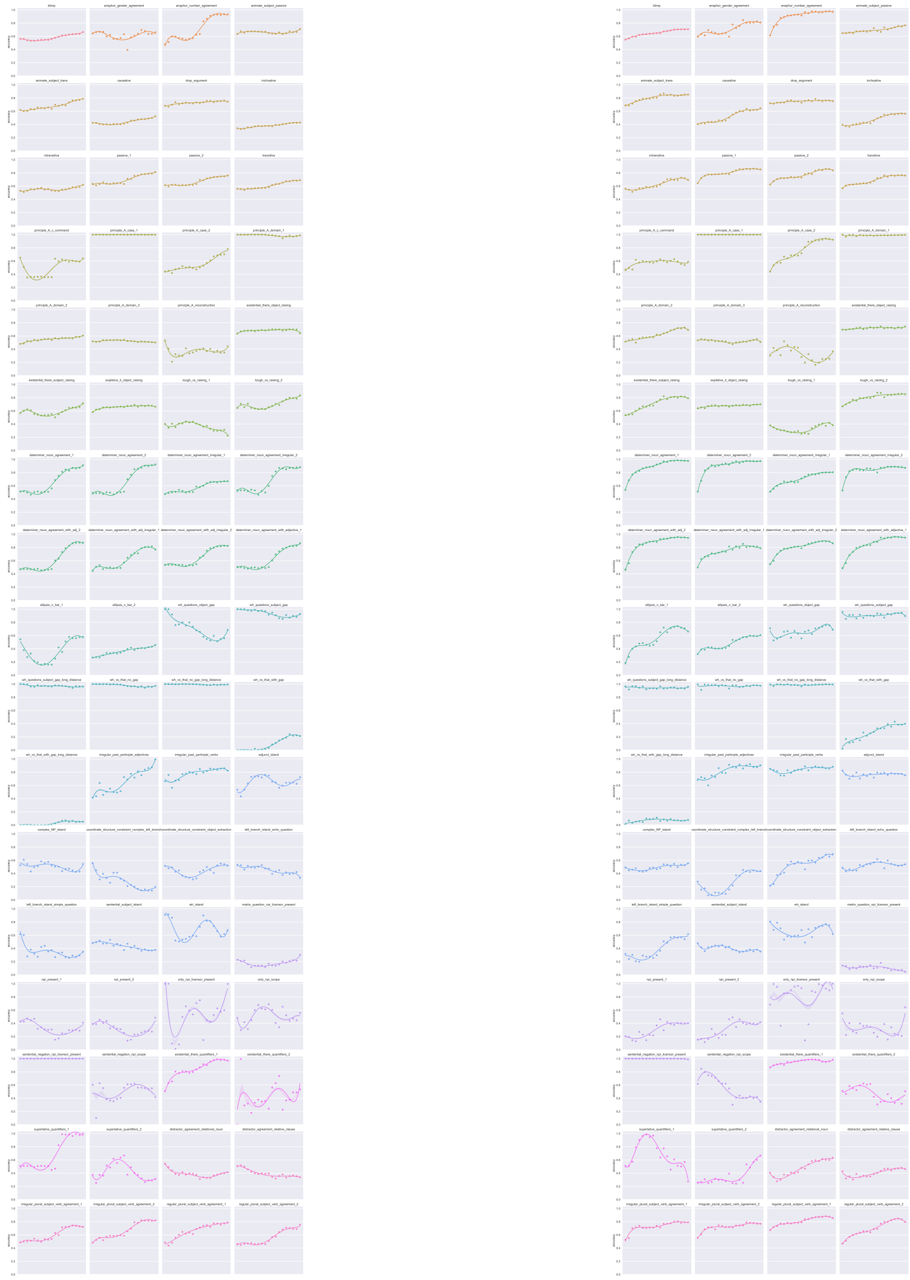


Figure 6: Learning curves for `weenie_llama` and `tweenie_llama`

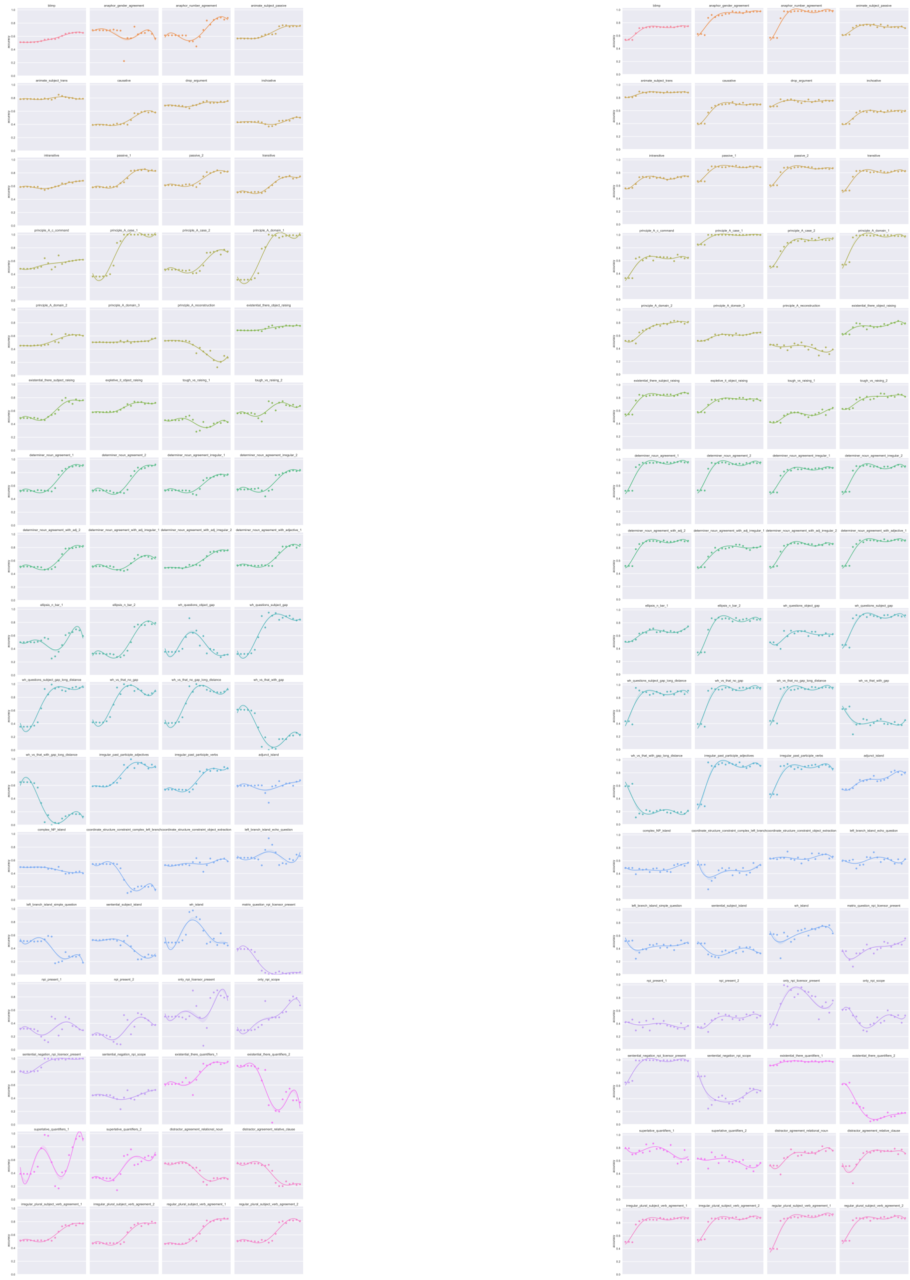


Figure 7: Learning curves for `pythia-14m` and `pythia-70m`

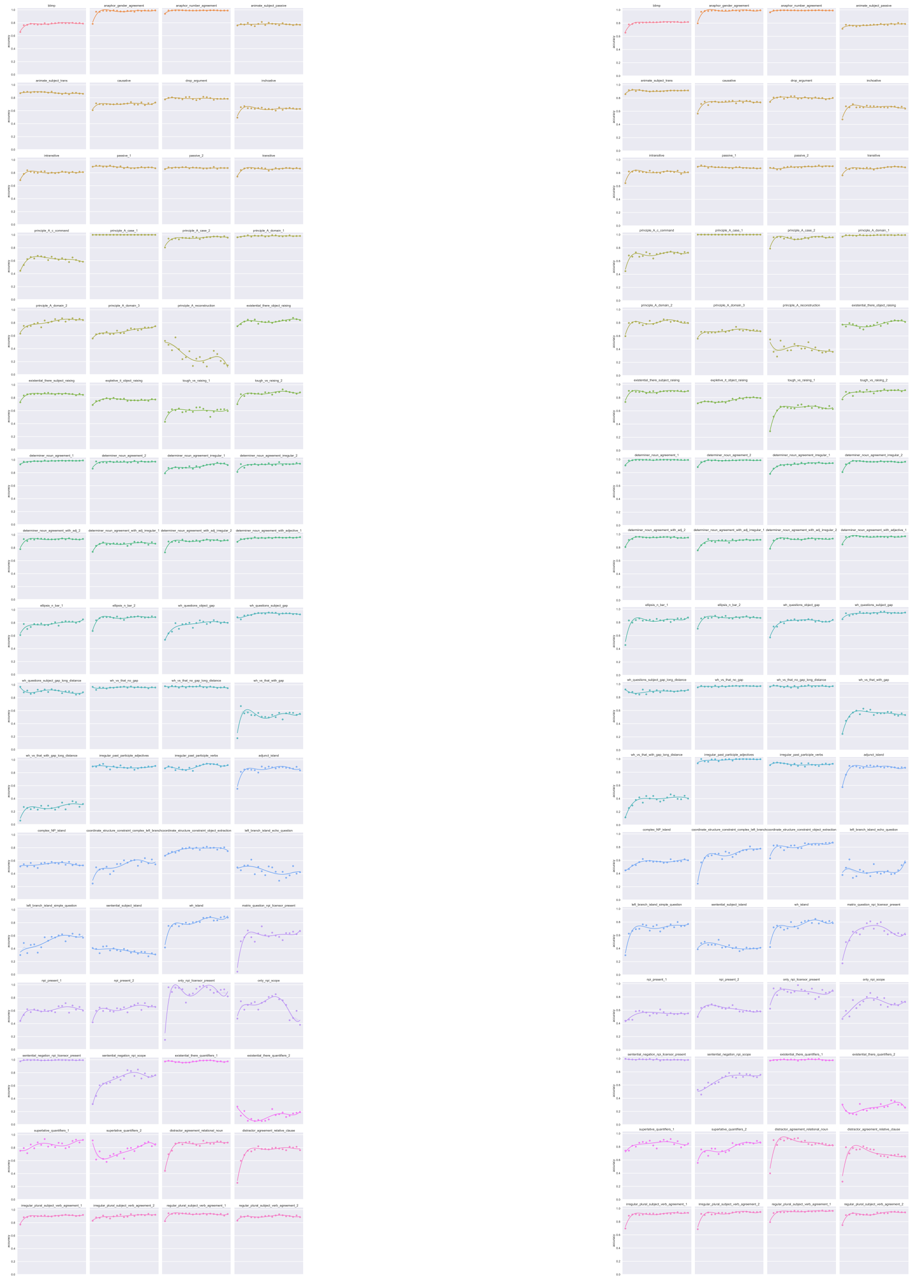


Figure 8: Learning curves for `pythia-160m` and `pythia-410m`

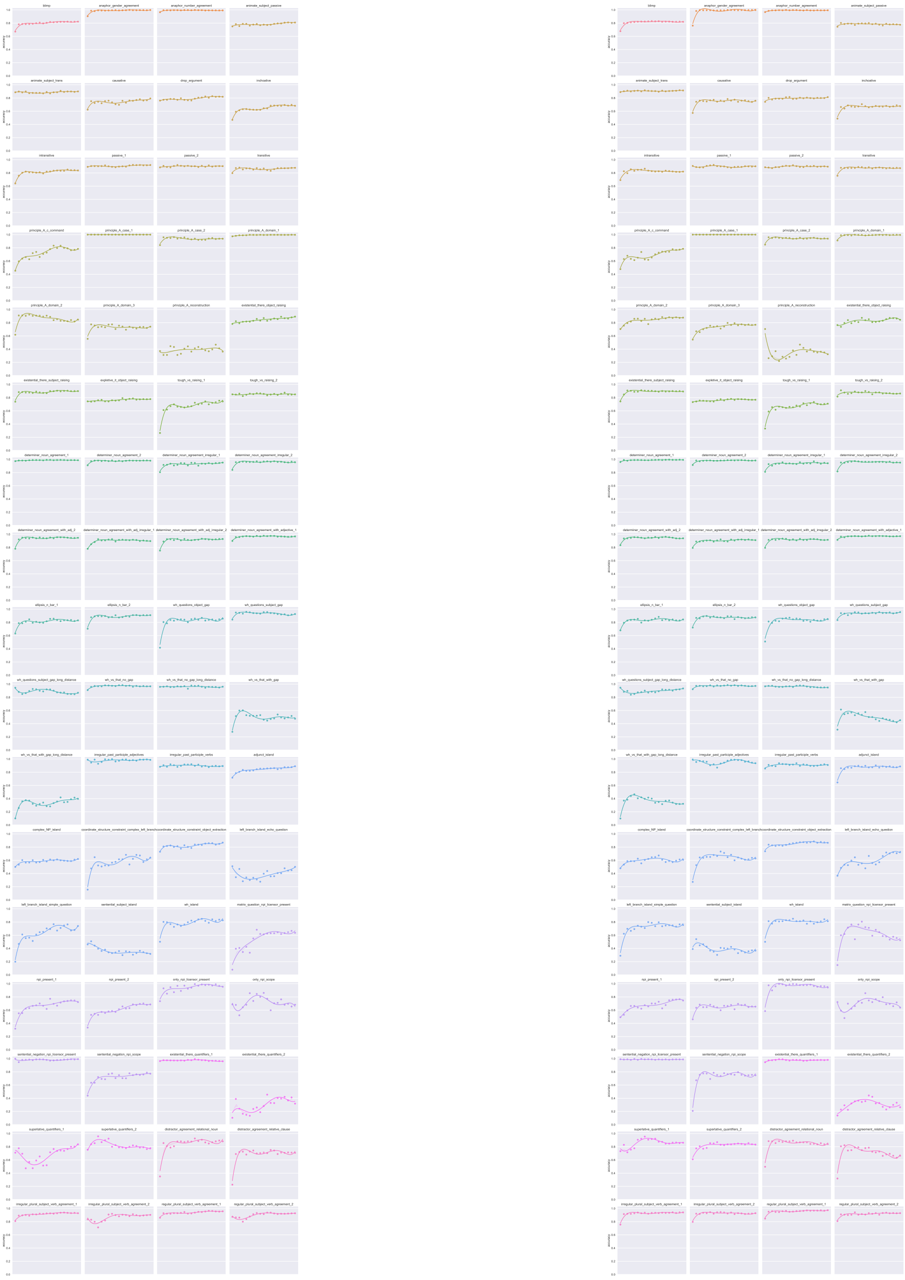


Figure 9: Learning curves for `pythia-1b` and `pythia-1.4b`

B Point plots for distance to mean performance

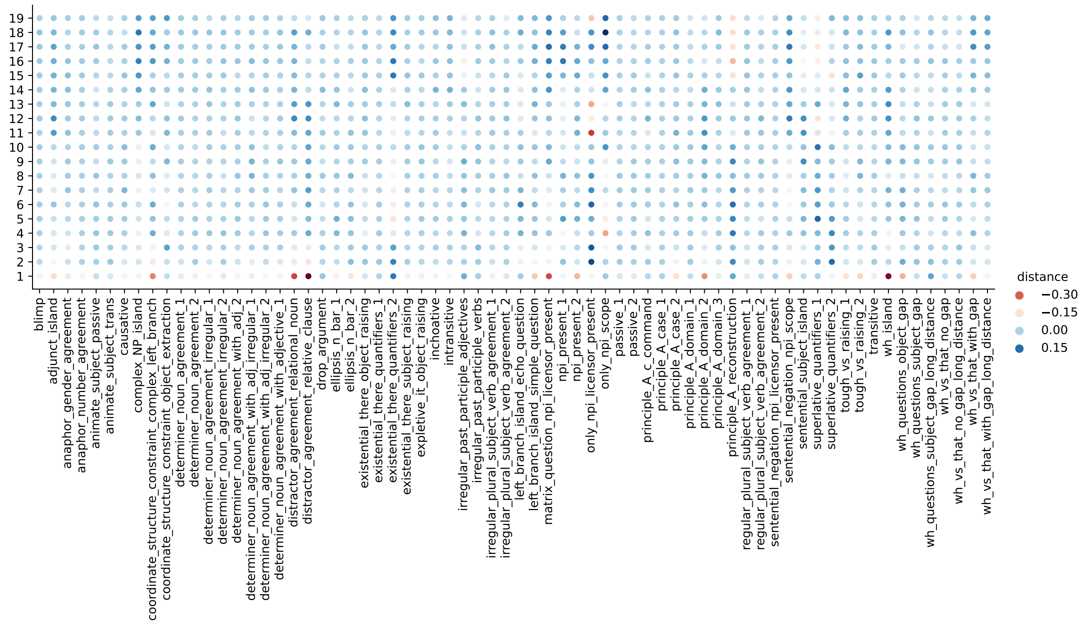
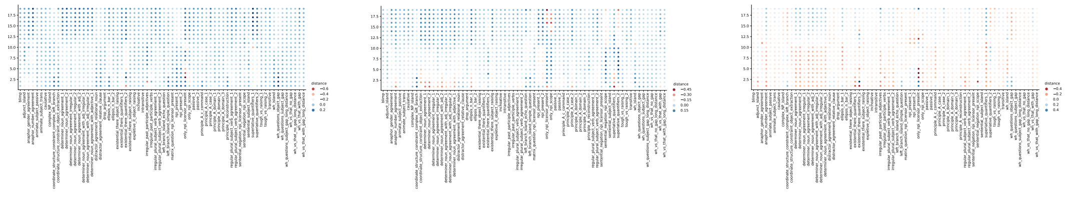


Figure 10: Paradigm-wise distances to mean paradigm performance for `pythia-70m` model

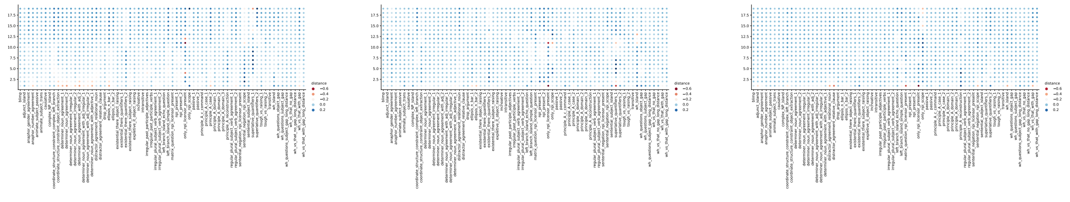


(a) `baby_llama`

(b) `teenie_llama`

(c) `weenie_llama`

Figure 11: Distance to mean for `baby_llama`, `teenie_llama` and `weenie_llama`

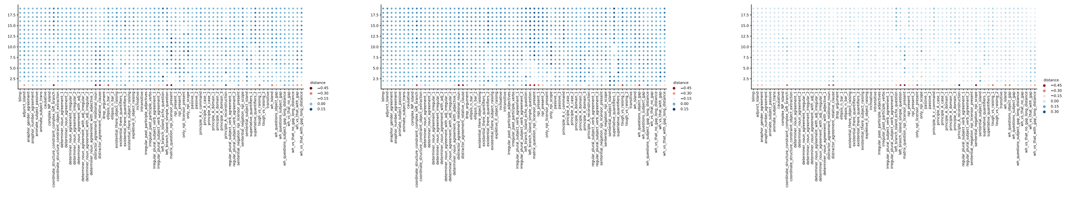


(a) `tweenie_llama`

(b) `pythia-14m`

(c) `pythia-160m`

Figure 12: Distance to mean for `tweenie_llama`, `pythia-14m` and `pythia-160m`



(a) `pythia-410m`

(b) `pythia-1b`

(c) `pythia-1.4b`

Figure 13: Distance to mean for `pythia-410m`, `pythia-1b` and `pythia-1.4b`

Not Just Semantics: Word Meaning Negotiation in Social Media and Spoken Interaction

Staffan Larsson
Department of Philosophy, Linguistics
and Theory of Science
Gothenburg University, Sweden
staffan.larsson@ling.gu.se

Jenny Myrendal
Department of Education,
Communication and Learning
Gothenburg University, Sweden
jenny.myrendal@gu.se

Abstract

This paper outlines an ongoing research project with the goal of of investigating how meanings of words (and phrases) are interactively negotiated in social media and in spoken interaction. This project will contribute to a comprehensive theory of word meaning negotiation.

1 Introduction

This paper outlines the project *Not Just Semantics: Word Meaning Negotiation in Social Media and Spoken Interaction* (VR 2022-02125), a project funded by the Swedish Research Council, that started in 2023 and currently planned to continue until 2026. The goal of the project is to investigate how meanings of words (and phrases) are interactively negotiated in social media and in spoken interaction. This project will contribute to a comprehensive theory of word meaning negotiation, which will characterise the phenomenon empirically and provide rigorous quantitative and qualitative analysis (including formalisation).

2 Purpose and aim of the project

While we may often take the meanings of our words as a given, the meanings of words (and phrases) are in fact frequently interactively negotiated by participants in linguistic interaction (Ludlow, 2014; Myrendal, 2015). Such Word Meaning Negotiations (WMNs) can be used in resolving misunderstandings, but can also be used rhetorically by interlocutors, to advance their view on some (possibly controversial) matter and to make their claims more plausible. Currently, WMN is an underexplored area, and we believe there is an opportunity for groundbreaking research with far-reaching scientific and practical benefits that this project will seize on.

The excerpts below are taken (and translated into English) from a Swedish online discussion forum. The posts are made by different participants in the discussion. The discussion concerns whether or not piercing the ears of young children is morally acceptable, or if it constitutes (child) abuse (sv. "(barn)misshandel").

1. Piercing the ears of young children (...) is abuse towards another human being! (...)
2. It isn't child abuse to pierce someone's ears. (...)
3. Of course it is abuse when you subject the child to unnecessary pain that they haven't asked for.
4. Clearly ABUSE to pierce the ears of young children! (...) - you inflict pain upon the child and a physical change which the child herself has not chosen and which cannot be made undone.

In addition to arguing for or against ear piercing in young children, participants are debating the meaning of 'child abuse', arguing about what the phrase means in order to support their overall claim for or against ear piercing. They do this by discussing whether or not "ear piercing" should count as a case of "(child) abuse" (1, 2 and 4 above), by offering full or partial definitions of "child abuse" (3 and 4 above), and in various other ways. WMNs can also concern politically charged phrases such as 'climate denier' (Sw. 'klimatförnekare'):

1. What do you mean by denier? Do you mean people who deny that we have an acute climate crisis (...)?
2. To be critical against alarmists is not the same as being a denier

We refer to discussions like these, where meanings are more or less explicitly negotiated, as Word Meaning Negotiations (WMNs). WMN occurs in many types of linguistic interaction, including everyday spoken conversation and social media interaction (Myrendal, 2015, 2019).

Understanding word meaning negotiation will contribute to our understanding of the social and normative nature of meaning, and the interactive processes involved in establishing shared meanings. Knowing more about WMN would also help us understand the role it plays in everyday life, as well as in politically or emotionally charged disputes in social media. The project will contribute towards a comprehensive theory of word meaning negotiation, which will characterise the phenomenon empirically and provide rigorous quantitative and qualitative analysis (including formalisation).

The project focuses on the following fundamental questions:

- What are the conversational strategies used in WMNs?
- How can we model how meanings are modified in WMNs?
- How are WMNs connected to arguments on controversial issues?
- What differences are there between social media and spoken interaction w.r.t. WMN?

3 State-of-the-art

Work in psycholinguistics has shown that speakers negotiate word choices and domain-specific meanings; see e.g. Clark and Gerrig (1983), Brennan and Clark (1996), Healey (1997), Pickering and Garrod (2004) and Mills and Healey (2008). Researchers in Conversation Analysis (CA) have studied phenomena such as disagreement and repair in conversation (Sacks, 1973; Kitzinger, 2012). Repair has also been studied from a computational perspective Purver et al. (2003).

Research on second language acquisition (Nakahama et al., 2001; Pica, 1994; Varonis and Gass, 1985) has identified a type of meaning negotiation that occurs when there is insufficient understanding between interlocutors regarding the meanings of particular words. This type of 'meaning negotiation' mostly refers to conversational repair.

Some types of Discourse Analysis, such as Critical Discourse Analysis (Fairclough, 2013), investigate how societal power relations are established and reinforced through language use. Walton (2001) discusses the role of definitions in argumentation, focusing on legal and political contexts.

Work in the philosophy of language (Ludlow, 2014) describes how discussions about the precise meanings of words like "planet", "person" and "rape" have recently entered the media spotlight. Often, different positions on controversial topics are aligned with views about the meanings of words. Ludlow mostly studies monological texts published in traditional media, but stresses the significance of studying how meaning is negotiated in interaction.

Work within computational linguistics related to social media and spoken interaction has addressed a vast array of topics, including lexical semantic change change (Tahmasebi et al., 2018), argument mining for online interactions (Ghosh et al., 2014), automatic detection of disagreement in online dialogue (Misra and Walker, 2017; Allen et al., 2014), automatic detection of emotions like sarcasm or nastiness in online conversation (Justo et al., 2014; Lukin and Walker, 2017) as well as classification of stance in online interaction (Sridhar et al., 2014; Walker et al., 2012). However, none of these studies have focused on the role played by meaning negotiation in relation to argumentation or disagreement in online communication.

All the research cited above is relevant and will be used to inform the approach developed in the present project. However, none of the approaches listed above have focused precisely on WMN as defined here and studied it using the combination of methods that we propose.

4 Significance and scientific novelty

Mainstream work on empirical and formal studies of dialogue and meaning has not, until recently, taken meaning negotiation seriously. This may in part be connected to an (explicit or implicit) assumption that meanings of words can be treated as static. As a result, there is to date very little work on quantitative and qualitative studies of naturally occurring WMNs.

In historical linguistics, semantic change has been studied "from a distance", focusing on slow, long-term and widespread changes. However, this cannot be the whole story. In the end, negotiation of meanings must take place in concrete instances of interaction (spoken or written) within a language community.

We also believe that understanding the process of WMN is essential to understanding the social nature of linguistic meaning. This touches on longstanding debates in linguistics and philosophy of language, such as the possibility of a private language, the normativity of language, the limits of meaning variation in language, and to what extent

WMN category	count	%
WMN: non-understanding	121	4,2
WMN: disagreement	6	0,2
WMN without trigger	40	1,4
WMN: other	7	0,2
Non-pursued WMN	2	0,07
Non-WMN clarification req.	23	0,8
Reference/named entity	23	0,8
Unclear	7	0,2
Multiple categories	9	0,2
No WMN	2619	91,7

Table 1: Preliminary annotation of 2857 potential WMNs (found using search expressions such as "what do you mean by") from the spoken section of the BNC

language is to be regarded as a mathematical, psychological or social entity. In WMNs, we can observe the social-normative dimension of language and meaning being played out in plain sight. By developing methods for finding WMNs we enable an empirical and data-driven approach to the study of this dimension of linguistic meaning. By developing and formalising a theory of WMNs we aim to give a precise account of the interactive dynamics involved in the emergence, perpetuation and variation of linguistic meaning in a speaker community over time.

Apart from theoretical and empirical work, this project will also develop automatic methods of detecting and analysing WMNs. Being able to detect, analyse and understand WMN has a range of potential applications. By better understanding the role of word meaning negotiations in discussions and controversies, we may gain insight into how opinions of individuals and communities influence (and are influenced by) the meanings we ascribe to words and expressions, and how opinions and word meanings interact over time.

5 Preliminary and previous results

Myrendal (2015, 2019) is the main starting point for the present project, and describes how word meanings are negotiated in social media, especially focusing on online discussion forum communication. Online discussion forums offer a particularly suitable material for studying naturally-occurring WMNs. These discussions typically take place between strangers who discuss a wide variety of more or less controversial topics, such as abortion, gender roles, and immigration policies.

Myrendal concludes that a WMN occurs when a discussion participant remarks on a word choice of another participant, thus initiating a meta-linguistic discussion in which a particular word is openly questioned and its meaning is up for negotiation. Myrendal distinguishes two main types of WMNs. NONs (non-understanding WMNs) comprise WMN sequences that are caused by insufficient understanding of a particular word. The second type, called DINs (disagreement WMNs), encompass sequences that originate in disagreement between participants regarding the meaning of a word and the way it is used in the discussion context.

Both NONs and DINs typically start off as a series of turns following a specific interaction pattern. Initially, a word is used by a participant which is remarked upon by another participant in a later turn, indicating that there is some kind of problem with regards to the meaning and/or use of the word. From that point in the interaction, the meaning of the word is up for negotiation and subsequent turns devote their attention to negotiation of word meaning.

P1: I'm anti-sexist, which means that I'm against sexism in society. Ask me anything!
P2: What do you mean by the concept of ”sexism”?
P1: That people are treated differently because of their gender.

Note that there needs to be a meta-linguistic shift that turns the focus of the conversation from being on topic to being on language in order for any conversation to turn into a WMN sequence. This shift is invited in the second turn, and the shift occurs in the third turn. On this basis, Myrendal (2015) develops a taxonomy of dialogue acts utilised by participants in WMN sequences. (Only selected parts of the taxonomy are presented here.)

- **Explicification** is a dialogue act used to introduce a definition-like component to the negotiated trigger word.
- **Exemplification** is a dialogue act that provides examples of what the trigger word can mean, or usually means, in a situation other than the current discussed situation.
- **Contrasting** is a dialogue act that positions the trigger word against another word, typically highlighting a similarity or difference between the two contrasted words.

- **Meta-linguistic clarification requests** are used to elicit more information about the perceived meaning of the trigger word.

Myrendal offers some quantitative results about the frequency of WMNs in online discussion forums, but these results are conditioned by the specific methods used to detect WMNs, and limited to Swedish discussion forum data. In the present project, we wish to take a more comprehensive approach to finding and classifying WMN sequences, thus enabling stronger quantitative claims.

We have previously have worked on formalisation of the WMN strategies identified by Myrendal, describing how they relate to updates of speakers' takes on meanings (Larsson and Myrendal, 2017; Noble et al., 2019). The present project will provide a large scale formal description of WMNs and their effect on word meanings.

6 Project Description

We take a dialogical perspective on language and communication in which linguistic meaning is viewed as a collaborative and interactive accomplishment between interlocutors (Clark, 1996; Linell, 2009). From this perspective, words possess flexible semantic qualities ("meaning potentials") that can be used in and across contexts to create situated meaning (Norén and Linell, 2007).

Methodologically, we will use a mix of methods to capture the complexities of the WMN phenomenon: corpus linguistics, quantitative analysis, and qualitative analysis (including formalisation). More precisely, we will use the following methods:

- Collecting a corpus of relevant social media and spoken interactions
- Identifying, classifying and annotating WMNs
- Developing automatic methods for detecting and classifying WMNs
- Quantitative analysis of WMNs
- In-depth qualitative analysis

The project is divided into four work packages, as follows:

WP1: Corpus collection: We will select relevant data of Swedish and English in interactive settings. Potential WMN sequences will be identified and retrieved using search expressions identified in previous research. While the precise search expressions will of course differ between languages, we expect similar overall patterns to occur in both Swedish and English. Table 1 presents preliminary results categorising dialogues retrieved from the spoken BNC with this method.

WP2: Detection and annotation Based on the analysis of dialogue acts involved in WMN in Myrendal (2015) and Noble et al. (2019), we are currently developing an annotation schema along the lines of Allen and Core (1997). The schema will go through a cycle of reliability testing and adjustment until satisfactory levels of reliability and depth of analysis have been reached. Data will be annotated by students who have received a brief explanation of the coding schema.

We will also develop and evaluate new techniques for detection and classification of WMN sequences. This work will build on Myrendal (2015) and on work on automatic detection of miscommunication related phenomena in dialogue (Purver et al., 2018). We will also be trying out LLMs on the task of detecting and classifyinng WMNs.

WP3: Quantitative analysis This WP aims to answer fundamental questions such as how common WMNs are. Using annotated corpus materials, we will investigate the overall frequency of WMNs, the relative frequency of the different negotiation strategies, and the dependence of these frequencies on contextual factors such as the type of social media platform and the general orientation of the forum. Some very preliminary frequency results from the BNC are shown in 1.

WP4: Qualitative analysis For the analysis of WMN strategies, we will use qualitative methods of interaction analysis influenced by and adapted from CA Hutchby and Wooffitt (2008).We will also continue work on formalisation of how various WMN strategies relate to updates of speakers' takes on meanings, using TTR (Cooper, 2023) which enables capturing rich dynamic meanings. In addition to providing detailed analysis, formalisation is a first step towards implementation of WMN capabilities in artificial agents (Schlangen, 2016).

Finally, we will explore how WMNs are connected to rhetorical argumentation, by examining to what extent and in which ways topoi play a role in WMNs (Breitholtz, 2020). Breitholtz suggests that the interpretation of word meaning is closely connected to reasoning where participants draw on topoi, rules of thumb for reasoning. We will explore the idea that topoi provide a link between WMNs and argumentation.

Acknowledgements

This work was supported by the Swedish Research Council (VR) grant 2022-02125 *Not Just Semantics: Word Meaning Negotiation in Social Media and Spoken Interaction*, and VR grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers.

Kelsey Allen, Giuseppe Carenini, and Raymond Ng. 2014. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1169–1180.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: the use of common sense reasoning in conversation*. Brill.

S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22:482–493.

H. H. Clark and R. J. Gerrig. 1983. Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior*, 22:591–608.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Robin Cooper. 2023. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press. Open access: https://global.oup.com/academic/product/from-perception-to-communication-9780192871312.

Norman Fairclough. 2013. Critical discourse analysis. In *The Routledge handbook of discourse analysis*, pages 9–20. Routledge.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the first workshop on argumentation mining*, pages 39–48.

P.G.T. Healey. 1997. Expertise or expertese?: The emergence of task-oriented sub-languages. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 301–306.

Ian Hutchby and Robin Wooffitt. 2008. *Conversation analysis*. Polity.

Raquel Justo, Thomas Corcoran, Stephanie M Lukin, Marilyn Walker, and M Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Celia Kitzinger. 2012. Repair. *The handbook of conversation analysis*, pages 229–256.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue acts and updates for semantic coordination. In *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue*, page 59.

Per Linell. 2009. *Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making*. IAP.

Peter Ludlow. 2014. *Living Words: Meaning Underdetermination and the Dynamic Lexicon*. Oxford University Press.

Stephanie Lukin and Marilyn Walker. 2017. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. *arXiv preprint arXiv:1708.08572*.

Gregory J. Mills and Patrick G. T. Healey. 2008. Semantic negotiation in dialogue: The mechanisms of alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, SIGdial '08, pages 46–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amita Misra and Marilyn Walker. 2017. Topic independent identification of agreement and disagreement in social media dialogue. *arXiv preprint arXiv:1709.00661*.

Jenny Myrendal. 2015. *Word Meaning Negotiation in Online Discussion Forum Communication*. Ph.D. thesis, University of Gothenburg.

Jenny Myrendal. 2019. Negotiating meanings online: disagreements about word meaning in discussion forum communication. *Discourse Studies*, 21(3):1–23.

Yuko Nakahama, Andrea Tyler, and Leo Van Lier. 2001. Negotiation of meaning in conversational and information gap activities: A comparative discourse analysis. *TESOL quarterly*, 35(3):377–405.

Bill Noble, Asad Sayeed, and Staffan Larsson. 2019. Towards a formal model of word meaning negotiation. In *Proceedings of the 23rd SemDial*, pages 210–212.

Kerstin Norén and Per Linell. 2007. Meaning potentials and the interaction between lexis and contexts: An empirical substantiation. *Pragmatics*, 17(3):387.

Teresa Pica. 1994. Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language learning*, 44(3):493–527.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02):169–226.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in cognitive science*, 10(2):425–451.

Harvey Sacks. 1973. The preference for agreement in natural conversation. *Linguistic Institute, Ann Arbor, Michigan*.

David Schlangen. 2016. Grounding, justification, adaptation: Towards machines that mean what they say. In *Proceedings of the 20th Workshop on the Semantics and Pragmatics of Dialogue (JerSem)*.

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. 2014. Collective stance classification of posts in online debate forums. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Evangeline Marlos Varonis and Susan Gass. 1985. Non-native/non-native conversations: A model for negotiation of meaning. *Applied linguistics*, 6(1):71–90.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012. That's your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Douglas Walton. 2001. Persuasive definitions and public policy arguments. *Argumentation and Advocacy*, 37(3):117–132.

Toward Real Time Word Based Prosody Recognition

Alex Tilson and **Frank Förster**
Robotics Research Group
School of Physics, Engineering and Computer Science
University of Hertfordshire
AL10 9AB, United Kingdom
{a.tilson, f.foerster}@herts.ac.uk

Abstract

Prosodic salience is a heuristic based on word-level prosody in child-directed speech that is thought to serve as a cue for attentional focus. It has been used in the context of robotic language acquisition to extract the contextually most relevant words from a human tutor's speech to ground them in a robot's sensorimotor data. However, the pipeline for performing word-based prosody-recognition operated in a semi-automatic manner and required substantial manual effort. We describe our efforts to automate the existing pipeline by including real time prosody recognition, and a modern speech recognition and forced alignment model. The intention is to enable its use in real time for human-in-the-loop robotic language acquisition and other socially driven forms of online learning.

1 Introduction

Prosodic salience is a measure calculated from a speech signal's pitch, energy, and duration features, and can be used to identify the most relevant words of an utterance produced by caregivers in child-directed speech.

This heuristic has demonstrated use in robotic language acquisition (Saunders et al., 2011, 2012), and can facilitate a more effective language learning process for robots, drawing on insights from how human children acquire language.

It has been used as part of the ITALK project (Broz et al., 2014) to learn the names of, and interactions with, objects based on human tutors' linguistically unconstrained speech when trying to teach the robot the names of various objects after having been told to speak to the robot as if it were a 2-year old child.

Further research was performed by (Förster et al., 2011; Förster et al., 2019) demonstrating that negation words, such as "no", are prosodically salient which may explain why it is typically amongst the first 10 words in English-speaking children's early active vocabularies (Fenson et al., 1994).

While the previous work shows prosodic salience to be useful for word-level language acquisition in developmental robotics, it may have a wider potential in speech interfaces. For instance, it might be used within dialogue systems in generating different responses depending on whether some word was produced meekly or with strong intonation - think of the difference between a meekly uttered "no" and a vehemently shouted one.

However, for word-based prosody recognition, features must be aligned accurately to the correct segment of the speech signal that is representative of the word. For prosodic salience, this meant that a large quantity of manual effort was required in the past from human transcribers marking word boundaries (e.g. Saunders et al., 2011, 2012; Förster et al., 2011; Förster et al., 2019). Hence for any meaningfully large corpus to be processed by this method, it would need to be scaled up by automating the alignment process with speech processing methods and speech recognition models.

The present paper describes our efforts to automate the existing semi-automatic prosody-processing pipeline.

2 Background

2.1 Child-Directed Speech and Language Acquisition

Child-directed and infant-directed speech (CDS and IDS respectively) are marked out by a number of modifications compared to adult-directed speech, that have been hypothesized to be conducive to human language acquisition. While not all of these modifications are present in all languages, they typically include an overall higher pitch, exaggerated intonation contours, a focus on topics relating to physically co-present objects or events, and important words being placed at an utterance-final

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 62–67
October 14–15, 2024. ©2024 Association for Computational Linguistics

position (Clark, 2009, chap. 2). Moreover, objects words have been observed to be pronounced relatively loudly (Saxton, 2017, chap. 4). While most of these observations characterise CDS and IDS on a general level, Soderstrom (2007) hypothesises that in IDS some of these acoustic modifications are performed on a word-level to aid the infant in both segmenting an utterance and singling out a target word.

In the context of robotic language acquisition and for the purpose of symbol grounding, based on the aforementioned features of CDS, Saunders, Lehmann, Sato, and Nehaniv (2011), operationalised word-based prosodic salience (cf. section 3). This was done to identify and extract prosodically salient words from an utterance produced by a human tutor when speaking to a child-like humanoid robot. Here, the prosodic salience of a word is identified as the product of a word's normalised pitch, energy, and duration values.

2.2 Human-in-the-Loop Real-time Reinforcement Learning

Senft et al. (2019) created an implementation for a reinforcement learning agent that learns from social feedback in an education setting. The reward signal the robot learned from was in the form of corrective feedback via a human manually pressing buttons to reward, punish, or manually initiate, actions. Because the teacher must consciously provide explicit feedback to the robot, their workload did not sufficiently decrease over time.

Belpaeme et al. (2018) express that the use of explicit signals in these cases acts as a proxy for naturally expressed implicit social signals. As some of these signals are typically embedded within speech, they contend that speech processing technology presented a bottleneck in their study preventing them from using such implicit speech-based social signals.

Prosodic salience is an example of such implicit social signals and similarly suffers from this bottleneck because of its reliance on the temporal alignment between the lexical level and the audio signal.

2.3 Forced Aligners

Forced alignment (FA) is the process of aligning a transcript to an audio signal. The traditional approach to forced alignment makes use of Hidden Markov Model (HMM) based automatic speech recognition pipelines, where statistical methods are used to model the probability distributions of phonetic units and to align the audio with the text.

Whilst traditional speech recognition models have largely been surpassed by attention based models such as (Baevski et al., 2020) and (Radford et al., 2023), attention based forced aligners haven't improved performance as significantly. For instance, NeuFA (Li et al., 2022) only marginally improves on the HMM based Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) and WhisperX (Bain et al., 2023) simply performs worse.

3 Methods

3.1 Prosodic Salience Pipeline

The prosodic salience estimation pipeline (Saunders et al., 2011) is as follows:

1. Transcribe the speech signal
2. Align the transcription
3. Split words into groups of utterances
4. Estimate the mean pitch and mean energy features of each spoken word
5. Estimate salience
6. Create a lexicon using the most salient words of each utterance
7. Ground sensorimotor experience with lexical units.

Originally steps 1, 2, and 3 were performed semi-automatically with human correction, with transcription alignment comprising the majority of manual effort from human transcribers. To automate these processes, we used Deepgram's Nova model (Deepgram, 2024) through an API call, which automatically produces a transcript, word boundaries, and utterance boundaries.

Additionally, step 4 relied on the Prosodic Feature Extraction Tool (PFET) (Huang et al., 2006), to calculate the relevant pitch, duration and energy features. However, as it was built on top of Praat (Boersma and Weenink, 2024), it was designed as an analysis tool, and has limited capabilities for full automation and real time execution.

To automate this process, we use OpenSMILE (Eyben et al., 2010) which is a highly configurable open source toolkit for signal processing and audio feature extraction. Comparitively, it has real time execution capabilities, and can be fully automated.

The calculation for estimating prosodic salience (or step 5), which is independent of the pipeline, is as follows: for a given word W_i of an utterance U, with $U = [W_1, \dots W_n]$, the word-based mean pitch, energy, and duration are scaled with respect to the maximum word-based mean value of the respective measure within U (p = pitch, e = energy, d = duration, s = prosodic saliency).

$$\hat{p}(U) = max(\{p(W_i) \,|\, W_i \in U\}) \quad (1)$$

$$\hat{e}(U) = max(\{e(W_i) \,|\, W_i \in U\}) \quad (2)$$

$$\hat{d}(U) = max(\{d(W_i) \,|\, W_i \in U\}) \quad (3)$$

$$s(W_i) = \frac{p(W_i)}{\hat{p}(U)} \times \frac{e(W_i)}{\hat{e}(U)} \times \frac{d(W_i)}{\hat{d}(U)} \quad (4)$$

For single word utterances, if the word's pitch, energy, or duration are larger than the first standard deviation of pitch, energy, and duration for the whole interaction session, then the word is also marked as salient.

3.2 Analysis

Step 1: Test of partially-modified pipeline using OpenSMILE To test the suitability of OpenSMILE to act as replacement for PFET, the modified pipeline using OpenSMILE was compared against the original pipeline using PFET by running them on robot-directed speech (RDS) corpus of Förster et al. (2019). For this corpus both manual transcriptions and word boundary time stamps are available, such that no additional method to detect word boundaries such as Deepgram is needed. Executing both pipelines yielded two sets of prosodically salient words whose frequencies we subsequently compared using Kendall rank correlation test.

Step 2: Test of fully-automated pipeline using both OpenSMILE and Deepgram The fully-automated pipeline was tested using Deepgram's Nova model which can generate both speech transcripts and word boundary time stamps. Using this pipeline, we generated a speech aligned transcript for the Newman-Ratner CDS corpus (Newman et al., 2016). This corpus was chosen due to the similarity of the two scenarios within which both the Förster and Newman-Ratner corpora were recorded. For reasons of data protection we were not allowed to upload the Förster corpus into the cloud-based Deepgram, hence the need for the Newman-Ratner corpus. The utterance boundaries and word alignments generated by Deepgram were then used with the prosodic features generated by OpenSMILE to calculate the duration, mean pitch, and mean energy values for each word, followed by calculating their prosodic salience. Subsequently the prosodically most salient words, one per utterance, were extracted for this corpus, and table of word frequencies created (cf. section 4). Upon reviewing the extracted words, we noticed an unexpected absence of object labels, which are known to occur frequently in child-directed speech and we would expect to be prosodically salient, as seen in the Förster corpus. This necessitated additional analyses to investigate the cause of the dissimilarity between the two corpora.

Follow-up Analysis: Forced Aligners on RDS Listening to the selected section of the audio recordings of the Newman-Ratner corpus made it clear that the fully-automated pipeline had failed to pick out the prosodically most salient words. After verifying the correctness of the speech transcripts generated by Deepgram, two potential error sources were identified: (1) a failure of OpenSMILE to correctly calculate the different prosodic feature values, for example due to noise or poor audio quality, and (2) a failure of Deepgram to correctly determine the word boundaries, leading to a misalignment of transcript and audio recording.

Hence Deepgram's alignment accuracy was tested using a test audio file from the Förster corpus and by comparing Deepgram's word boundaries to the human-generated baseline. The file was 193 seconds long, consisting of 181 words, of which only 141 were used, as they were a part of an utterance which contained a saliently predicted word and therefore the most likely to affect the results. To account for cases where more than one word was produced by Deepgram, word alignments were paired based on the closest match of start and end timestamps. Algorithm 1 was used to quantify the degree of misalignment.

Algorithm 1 *Overlap Function* a and b are time intervals under comparison, specifying word boundaries as tuples, with a: ground-truth, and b: other boundaries (here: generated by Deepgram).

```
function OVERLAP(a, b)
    $a\_len \leftarrow |a[1] - a[0]|$
    $b\_len \leftarrow |b[1] - b[0]|$
    $overlap \leftarrow \min(a[1], b[1]) - \max(a[0], b[0])$
    $missing \leftarrow a\_len - overlap$
    $extra \leftarrow b\_len - overlap$
    return $(overlap, missing, extra)$
end function
```

The `overlap` function calculates the overlap, missing length, and extra length between two intervals a and b.

4 Results

4.1 Prosodic Feature Extraction

Step 1 The outcome of the analysis performed in Step 1 is depicted in Fig. 1. Shown are the relative frequencies of the top 10 most frequent prosodically salient words of the Förster corpus for both pipelines. The Kendall rank correlation test yielded a $\tau_B = 0.86$ ($p <= .001$), indicating a large correlation.

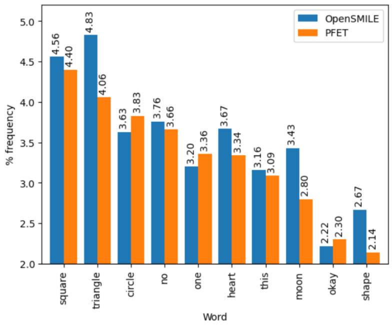


Figure 1: Frequency of the 10 most frequent prosodically salient words by pipeline. The pipeline is identical besides the prosodic feature extraction method. Methods compared are the originally used Chen and Harper's prosodic feature extraction tool (PFET) vs OpenSMILE.

Step 2 Table 1 depicts the 10 prosodically most salient words as detected by the fully-automated pipeline. Objects labels, dominating the list of most-frequent prosodically salient words in Förster's RDS corpus are suspiciously missing here, indicating a problem with the prosody detection.

Word	Freq.	Word	Freq.
oh	1367	baby	457
yeah	1032	look	436
you	700	that	339
okay	669	what	240
no	665	see	238

Table 1: 10 most frequent prosodically most salient words of the Newman-Ratner corpus by frequency of occurrence over all participants as output by the fully-automated pipeline.

4.2 Forced Aligners on Robot Directed Speech

Table 2 shows the total overlap, missing, and extra sections between the baseline word alignment and the one generated by Deepgram when run on the test audio file from the RDS corpus. Numerically, the missing and extra parts of audio account for nearly the same portion as overlap. This is catastrophic for prosodic salience estimation, as word level prosody data can change frequently, and nearly half of it is either erroneous or missing.

Category	Time (seconds)
Overlap	34.021
Missing	14.366
Extra	15.678

Table 2: Totals for Overlap, Missing, and Extra, segments of audio for Deepgram's prediction compared to human aligned. Overlap represents the total time in seconds where the time ranges agree. Missing represents the portions of audio where the prediction undershoots the word. Extra represents the portions of audio where the predicted region overshoots the word.

5 Discussion

Our results indicate that word boundary detection as performed by forced aligners, still remains an open problem when applied to child-directed speech and with respect to word-based prosody detection. We observed that once a boundary detection error occurs within an utterance, this type of error frequently propagates to the boundaries of subsequent words in that utterance. This subsequently renders word-based prosody detection difficult to impossible. However, given a correct set of correct word boundaries, current automatic prosody feature extraction tools such as OpenSMILE appear to perform sufficiently well when compared semi-automatic prosody processing methods involving tools such as PRAAT. Because traditional FA performs poorly in non-standard domains, settling for a hybrid usage of HMM and attention models for speech alignment appears to be insufficient. Purely attention based forced alignment models hold some promise for improvement.

Future Work Elsner and Ito (2017) posit that forced aligners perform poorly on CDS due to its atypical phonetics, resulting in what they call "catastrophically aligned words". In their work, a Kaldi forced aligner was adapted to CDS by treating it as a domain adaptation problem. We hence intend to tune NeuFA (Li et al., 2022) and similar attention-based aligners to chosen CDS and RDS corpora to adequately adapt it to the respective domains.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics*, 3(21):eaat5954.

Paul Boersma and David Weenink. 2024. Praat: doing phonetics by computer [computer program]. Retrieved 31st May 2024 from http://www.praat.org/.

Frank Broz, Chrystopher L. Nehaniv, Tony Belpaeme, Ambra Bisio, Kerstin Dautenhahn, Luciano Fadiga, Tomassino Ferrauto, Kerstin Fischer, Frank Förster, Onofrio Gigliotta, Sascha Griffiths, Hagen Lehmann, Katrin S. Lohan, Caroline Lyon, Davide Marocco, Gianluca Massera, Giorgio Metta, Vishwanathan Mohan, Anthony Morse, Stefano Nolfi, Francesco Nori, Martin Peniak, Karola Pitsch, Katharina J. Rohlfing, Gerhard Sagerer, Yo Sato, Joe Saunders, Lars Schillingmann, Alessandra Sciutti, Vadim Tikhanoff, Britta Wrede, Arne Zeschel, and Angelo Cangelosi. 2014. The ITALK Project: A Developmental Robotics Approach to the Study of Individual, Social, and Linguistic Learning. *Topics in Cognitive Science*, 6(3):534–544.

Eve V. Clark. 2009. *First Language Acquisition*. Cambridge University Press, Cambridge, UK.

Deepgram. 2024. Deepgram Voice AI: Text to Speech + Speech to Text APIs. Accessed: 2024-10-06.

Micha Elsner and Kiwako Ito. 2017. An Automatically Aligned Corpus of Child-Directed Speech. In *Proc. Interspeech 2017*, pages 1736–1740.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Larry Fenson, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Frank Förster, Chrystopher L. Nehaniv, and Joe Saunders. 2011. Robots that say 'no'. In *Advances in Artificial Life. Darwin Meets von Neumann*, pages 158–166, Berlin, Heidelberg. Springer Berlin Heidelberg.

Frank Förster, Joe Saunders, Hagen Lehmann, and Chrystopher L. Nehaniv. 2019. Robots learning to say "no": Prohibition and rejective mechanisms in acquisition of linguistic negation. *J. Hum.-Robot Interact.*, 8(4).

Zhongqiang Huang, Lei Chen, and Mary Harper. 2006. An open source prosodic feature extraction tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jingbei Li, Yi Meng, Zhiyong Wu, Helen Meng, Qiao Tian, Yuping Wang, and Yuxuan Wang. 2022. Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8007–8011. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Rochelle S. Newman, Meredith L. Rowe, and Nan Bernstein Ratner. 2016. Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, 43(5):1158–1173. Epub 2015 Aug 24.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Joe Saunders, Hagen Lehmann, Frank Förster, and Chrystopher L. Nehaniv. 2012. Robot acquisition of lexical meaning - moving towards the two-word stage. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7.

Joe Saunders, Hagen Lehmann, Yo Sato, and Chrystopher L. Nehaniv. 2011. Towards using prosody to scaffold lexical meaning in robots. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–7.

Matthew Saxton. 2017. *Child language: Acquisition and Development*. Sage Publications Ltd, London, UK.

Emmanuel Senft, Séverin Lemaignan, Paul E. Baxter, Madeleine Bartlett, and Tony Belpaeme. 2019. Teaching robots social autonomy from in situ human guidance. *Science Robotics*, 4(35):eaat1186.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Author Index