Extending the BabyLM Initiative: Promoting Diversity in Datasets and Metrics through High-Quality Linguistic Corpora

Laurent Prévot
CNRS & MEAE
CEFC
Taipei, Taiwan
laurent.prevot@cnrs.fr

Sheng-Fu Wang
Academia Sinica
Institute of Linguistics
Taipei, Taiwan
sftwang@gate.sinica.edu.tw

Jou-An Chi
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
r11142005@ntu.edu.tw

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
shukaihsieh@ntu.edu.tw

Abstract

BABYLM initiative paves the way for a range of experiments aimed at better understanding language models (LMs) and the differences and similarities between human and artificial language learning. However, the current framework is limited to the English language and a range of evaluation metrics, focused on syntax, semantics, and pragmatics. In this paper, we propose some steps towards extending the framework to other languages, like French, leveraging existing linguistic resources for these languages. Additionally, we advocate for greater exploration of genre variations within subcorpora for training LMs, as well as for the adoption of additional evaluation metrics with different underlying principles. Our proposal consists of using high-quality spontaneous speech corpora as a source for extracting production-related variables, which the models are then fine-tuned to predict. We hypothesize that these production-related features offer insights into the language processing mechanisms underlying the data and that cognitively sensitive models should outperform others in predicting these features. Specifically, we propose focusing on the prediction of phenomena such as speech reductions, prosodic prominences, sequences co-occurring with listeners' backchannels, and disfluencies. To illustrate our approach, we present an example involving the prediction of speech reductions and prosodic prominences in spontaneous speech in two different languages (French and English), using models trained on 10 million tokens from different data source mixtures.

1 Introduction

The BABYLM initiative is built on three interrelated aspects: (i) data sets for training language models, (ii) evaluation metrics designed to capture cognitive and linguistic skills and their development, and (iii) models that are either more cognitively plausible and/or capable of learning efficiently from "small" datasets. This initiative represents a strategic and timely effort to better understand the differences between artificial and human language learners.

While the 2023 edition focus was primarily on models, the 2024 call expands the scope to include investigations into both datasets and evaluation metrics—a crucial step, as we will argue in this position paper. Specifically, we propose concrete directions for expanding language model training datasets and exploring new evaluation metrics to deepen the linguistic and cognitive relevance of the BABYLM evaluation framework. Regarding evaluation metrics, we advocate for a novel approach that leverages existing high-quality spontaneous speech corpora.

One observation about the BABYLM initiative to date is that the datasets used are in English. While this is a natural starting point, it represents a significant limitation. Expanding the scope to include more languages is not only about better representing linguistic communities or potential model users. Achieving comparable, contrastive results across different languages within the BABYLM framework could offer valuable insights into both the learning models and the underlying learning processes.

While the original BABYLM initiative argues convincingly for a mix of data sources including transcripts of child-caregiver conversations, everyday conversations, subtitles, and simple texts, different mixtures can be explored. Due to data

The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, pages 147–158

scarcity, it is still impossible to gather a 100M data set based on real spoken conversational data but the 10M is accessible for a few languages like English, French and Mandarin and a few others. Conversational speech is the genre within which humans acquire their basic language skills. It is a genre quite distant from the usual written or web content on which LMs are trained, increasing the risk of biases for LMs produced. Moreover, it has been argued that it is a genre of high relevance to language emergence (Levinson, 2020; Christiansen and Chater, 2022). How could a purely interactional dataset, including both child-directed and general conversation transcripts, be compared to more balanced mixtures? This opens the door for testing various hypotheses. For instance, does including more encyclopedic knowledge help with higher-level commonsense tasks, while a purely conversational training set provides a model with better communicative and conversational abilities?

In this context, current evaluation metrics, while a good starting point, appear biased in two ways: they tend to favor canonical written forms and prioritize syntactic, semantic, and commonsense pragmatics. However, language and communicative competence include many other dimensions. Although the initiative clearly emphasizes the importance of using speech transcripts, both child-directed and everyday conversations, as training data, to our knowledge, none of the evaluation metrics employed address explicitly the specificities of spontaneous speech.

To summarize, we argue that, in line with the directions proposed in this year's new call, training datasets, and evaluation metrics are just as crucial as models for understanding the computational learning of language structure. We propose evaluation metrics based on spontaneous speech data and demonstrate how we can build such metrics from different aspects of the speech signal and transcripts obtained from high-quality spontaneous speech corpora.

2 Related Work

Since the emergence of large language models, there has been strong interest from the computational linguistics community in understanding why they are so successful. Warstadt et al. (2020b) explore the conditions (e.g., the amount of training data) under which RoBERTa develops and leverages linguistic features, such as part of speech (POS) and morphology, as opposed to relying on simpler surface-level features like simple position-based or length-based features. More recently, several studies have probed LLMs to better characterize their performance across various domains, particularly with regard to their linguistic competence versus commonsense reasoning. These studies have also examined the relationship between model performance and the amount of training data required for different tasks. In particular, Zhang et al. (2021) used training sets of varying sizes, 1M, 10M, 100M, and 1B tokens, to show that syntactic and semantic competence becomes robust in the 10M-100M range, whereas larger datasets are needed to achieve strong results in pragmatic and commonsense reasoning tasks.

More broadly, there have been proposals for evaluating the performance of LLMs on diverse linguistic tasks. Warstadt et al. (2019b) leveraged a substantial body of generative syntax-semantics literature to develop benchmarks based on acceptability judgments, coming either the linguistic literature like the CoLA benchmark further extended by exploiting more sources and data augmentation methods in BLIMP (Warstadt et al., 2020a). In addition to these binary decision tasks, Zhang et al. (2021) combined three other types of evaluation metrics: *classifier probing* (following (Ettinger et al., 2016; Adi et al., 2017)), which includes tasks from POS tagging to coreference resolution; *information-theoretic* probing based on the minimum description length (MDL) principle; and *fine-tuning on higher-level tasks* such as those in the SuperGLUE benchmark.

Most of the benchmarks have been proposed for English. However, BLiMP Warstadt et al. (2019a) has inspired a series of language-specific benchmarks, such as CLIMP for Mandarin Chinese (Xiang et al., 2021), as well as benchmarks for other languages like Japanese (Someya and Oseki, 2023), Dutch (Suijkerbuijk et al.), and Russian (Taktasheva et al., 2024). These are important additions to the evaluation landscape. While these benchmarks represent important extensions to the general evaluation framework, they all rely on syntax-semantics structures derived from introspection and textbook data, as will be discussed in the next section. In parallel to these efforts, monolingual language models have been developed using large amounts of data (Chang et al., 2024), as well as experiments involving varied data quantities (Micheli et al., 2020).

In another line of research, several studies have tested the ability of large language models (LLMs) to perform tasks inspired by cognitive science, particularly in the domains of semantics and pragmatics (Ettinger, 2020; Binz and Schulz, 2023).

Our approach of using actual speech data to extract production-based metrics can be related to studies that use behavioral or neurophysiological data linked with linguistic datasets. Specifically, there has been significant work focusing on textual datasets combined with eye-tracking (Hollenstein et al., 2021) or neurophysiological (Bingel et al., 2016; Hollenstein et al., 2018) measures. Additionally, datasets from passive listening tasks, linked to fMRI, have been released for various languages (e.g., French, Mandarin, and English) (Li et al., 2022). These datasets have been used, for instance, to study the impact of training parameters on a language model's ability to predict neurophysiological data (Pasquiou et al., 2022). Focusing on spontaneous speech, (Rauchbauer et al., 2019; Hmamouche et al., 2024) examined the predictability of fMRI-derived signals from conversational variables, including lexical information.

In terms of specialized language models, (Cabiddu et al., 2025) developed LMs based on child-directed speech transcripts and evaluated them on word-sense disambiguation tasks. They concluded that word acquisition trajectories could be better captured by multimodal models that incorporate acoustic features, among other aspects. Regarding more specifically tokenizers, Beinborn and Pinter (2023) proposed an evaluation paradigm focusing on the cognitive plausibility of subword tokenization. They compared `BPE`, `WordPiece`, and `UnigramLM` and revealed a lower "cognitive correlation" for the latter. Lastly, in the most recent BabyLM edition, (Martinez et al., 2023) introduced an interesting learning curriculum that constrained vocabulary in the early stages to simulate more cognitively plausible learning curves. Although this approach did not yield consistent overall results, marginal gains were observed in selected tasks.

3 A proposal for a new source of metrics

All initiatives mentioned are grounded in text-based and/or handcrafted paradigms, potentially coupled with behavioral and /or physiological lab measures. In contrast, we propose using actual spontaneous conversational transcripts to build complementary benchmarks that test not only the syntax-semantics dimensions but also real-world language use. These metrics will remain fundamentally linguistic in nature rather than focusing on task-specific or end-to-end evaluation.

Language is acquired, especially in its early stages, within spontaneous, conversational environments. While conversational language shares grammatical structures with other genres, its unique characteristics suggest that simply listing syntactic "errors" or semantic incongruities does not fully capture linguistic competence. Furthermore, in a conversational context, what may be considered a production error from a formal grammatical perspective is often perfectly acceptable and successfully achieves its communicative purpose. Therefore, we aim to develop a complementary approach that provides a broader set of metrics for evaluating language models from both cognitive and communicative perspectives when combined with existing benchmarks.

Specifically, we propose using spontaneous speech corpora, as they offer insights into human language processing through various observable production phenomena. Our approach is a kind of *classifier probing* (Ettinger et al., 2016; Adi et al., 2017; Warstadt et al., 2019b), but rather than focusing on meta-linguistic tasks (e.g., predicting syntactic categories), we aim to predict phenomena that serve as partial indicators of language processing. We propose a preliminary set of potential metrics, which remains open for further development. These metrics include *speech reductions, listener's backchannel signals, prosodic prominences*, and *disfluencies*. The common point among these metrics is that they are all grounded in spontaneous speech production, and each has been the subject of extensive research.

3.1 Speech reductions

Speech reductions have been studied across a range of linguistic levels, from phonetics to semantics, especially when considering the issue of signal information density. In spontaneous speech, some chunks of speech are produced in a reduced manner, both in terms of duration and articulatory amplitude. The location of these reductions is not random. For example, studies have suggested that speakers tend to smooth the information density of their speech signal over time, with reductions serving as a mechanism to achieve this smoothing effect (Aylett and Turk, 2004).

The relationship between information density and speech reduction has led to research developments on this topic with various approaches. These approaches may differ in the probabilistic measures used to predict reductions, such as lexical frequency, contextual probability, and informativity (Aylett and Turk, 2004; Gahl, 2008; Cohen Priva, 2012; Seyfarth, 2014). They also differ in terms of the linguistic level at which reductions occur, whether at the phoneme-, syllable-, word-level, or in terms of overall speech rate. Many of these studies include and compare different types of probabilistic measurements (e.g., lexical frequency and contextual probability) within a single study (e.g., (Seyfarth, 2014; Cohen Priva and Jaeger, 2018)) and some of them also compare probabilistic measurements calculated at different linguistic levels (e.g., segment- and syllable-levels in Van Son et al. (1999), segment- and word/-level measurements in Van Son and Pols (2003), syllable- and word-level measurements in Wang (2022)). Inclusion and comparison of reductions or phonetic variability across various linguistic levels in the same study have also been done (e.g., individual segments and prefixes as a whole in Pluymaekers et al. (2005); morphemes and words in Tang and Bennett (2018)), albeit less frequently.

These studies show that phonetic reduction can be predicted to varying degrees on the basis of the statistical distribution of linguistic units, and the prediction has been repeatedly found with varying types of measurements at various levels of linguistic units. This motivates the development of a reduction-labeling task for evaluating language models.

3.2 Prosodic Prominences

Prosodic prominence refers to the emphasis placed on certain units, often demarcated at the level of words or syllables, within a spoken utterance. This emphasis can be measured through (and perceived based on) acoustic cues such as movements in fundamental frequencies, duration, intensity, and segmental properties such as the formant structure of vowels. Recent work by Wolf et al. (2023) has shown a significant degree of redundancy between the representations encoded from tokens alone and those derived from acoustic-prosodic information. Acoustic-prosodic features such as word-level energy, fundamental frequency, duration, pause, and composite measurements derived using a wavelet-based algorithm (Suni et al., 2017) were used to quantify this redundancy. Their findings suggest that prosodic information can be predicted, to some extent, from the word itself and its surrounding context.

Furthermore, Kakouros and O'Mahony (2023) suggests that language models (in their study, BERT) use syntax-semantics layers to predict prosodic aspects. While we do not argue that text alone can fully predict prosodic prominence (as also noted by Wolf et al. (2023)), we remark that part of prosodic prominence can indeed be predicted by a language model. In the case of spontaneous speech, this prosodic information reflects an additional layer of language processing. Therefore, language models that better capture human language processing should have an advantage over models trained exclusively on raw written text, particularly concerning prosodic prediction.

3.3 Listeners' Signals

Although not directly linked to the speaker's production, backchanneling (Yngve, 1970) offers another perspective on language processing. Backchannels do not occur randomly; they are frequent in casual conversations and closely related to turn-ending prediction (Skantze, 2021). There has been an ongoing debate about whether predicting the exact location of a turn-ending is a matter of lexical and syntactic completion prediction (De Ruiter et al., 2006) or based on prosodic cues from the main speaker (Bögels and Torreira, 2015). Finer-grained experiments by Riest et al. (2015) identified semantic completion as a crucial source of information for predicting turn-endings.

Our position is that if a listener can anticipate when it is appropriate to produce a backchannel, and even if part of this decision is based on prosodic cues from the main speaker, language models should be capable of predicting these moments to some extent.

3.4 Disfluencies

Directly predicting disfluencies (Shriberg, 1994), as discussed earlier, is challenging because disfluencies are explicit in the token stream. Removing them and labeling sequences where they originally appeared is cumbersome and potentially problematic. A more effective approach might be to adopt a well-established evaluation method: comparing "acceptable" versus "unacceptable" sequences. While disfluencies exhibit various sub-

tleties, most follow a few simple patterns. We could compare actual utterances from high-quality linguistic corpora that do include detailed disfluency transcription with artificially generated utterances where disfluent patterns have been injected, similar to the syntactic acceptability approach used in Wagner et al. (2009) and Warstadt et al. (2019b).

4 Data

4.1 Pre-training data for creating LLMs

We have trained several language models. For the French experiment, we trained one model on 10M tokens from conversational datasets inspired by the original BABYLM data mix (ORFEO[1] (Benzitoun et al., 2016) and CHILDES-FR[2] (MacWhinney, 2014; Rose and MacWhinney, 2014)) and another on 10M tokens from Wikipedia. The training process used standard parameters (a BPE tokenizer with a 10K vocabulary size[3], a minimum token frequency of 2, and training for 3 epochs), with implementations from the HUGGINGFACE packages.

Similarly, three English models were trained on three size-matched datasets containing 9M tokens from the following sources: a subset of the BABYLM 10M training data, "spoken" data that included BNC and Switchboard subsets from the BABYLM 100M training data, and a subset of Simple Wikipedia data from the BABYLM 100M training data. Subsets of the corresponding validation data from BABYLM were also used to create 0.9M-word validation sets for early stopping in LM training (maximum epochs = 100, early stopping patience = 3).

We included ROBERTA models in the fine-tuning experiments to serve as a topline for this task. The purpose of using ROBERTA models, which do not fit any of the BABYLM tracks, was to better contextualize our proposed metrics as a form of sanity check. The underlying idea is that if full-fledged LMs like ROBERTA fail to perform the task, it is likely that the task cannot be achieved given the provided data.

4.2 Benchmarks

For these experiments, we used two sources to build benchmarks: the Corpus of Interactional Data (CID) for French[4] (Blache et al., 2017) and the Buckeye Corpus for English[5] (Pitt et al., 2005). CID is an 8-hour corpus of 1-hour conversations between friends (16 speakers). It features fiercely spontaneous conversational speech. Buckeye is a corpus with 38.1 hours of spontaneous speech (40 speakers) recorded in an interview format.

The main reason for the choice of these corpora is the high quality of their speech transcript alignment, down to the syllable or even the segment level. These spontaneous datasets have also been used in various phonetic studies (Raymond et al., 2006; Meunier and Espesser, 2011).

5 Experiments

The experiments evaluated different pre-trained models for our set of tasks.[6] More precisely, we fine-tuned the pretrained models separately on a token classification task to predict which tokens were labeled (reduced / prominent / backchannelled) and which were not. A simple cross-validation was conducted across groups of speakers to maximize diversity across the folds.

5.1 Speech Reduction

There are several methods to determine whether a portion of speech is reduced. Following approaches in the literature, we first derived ratios of every word token's actual duration and its expected duration. For the French benchmark, we leveraged annotations of syllable boundaries in the corpus and developed a model that predicts syllable duration based on the segment it contains, similar to Wang (2022). A model is trained on one-half of the corpus and then applied to estimate the expected token duration in the remaining half of the corpus. For the English benchmark, we calculate words' expected duration from their component phonemes' mean duration in the corpus (Bell et al., 2009; Gahl et al., 2012; Seyfarth, 2014).

In both cases, we then converted the ratios into binary labels by applying a threshold of 0.7 (i.e., a reduction of at least 30%). This threshold resulted in labelling 33% of the tokens as reduced in the French benchmark and about 35% of the tokens in the English benchmark. These labels were then encoded in a BIO format.

[1] https://hdl.handle.net/11403/cefc-orfeo

[2] https://phon.talkbank.org/access/French/

[3] We tested vocabulary sizes of various sizes. Although the scores varied, they did not affect the performance hierarchy between the models.

[4] https://hdl.handle.net/11403/sldr000720

[5] https://buckeyecorpus.osu.edu/

[6] Notebooks for pretraining LMs and performing the experiment can be accessed at https://github.com/prevotlaurent/babyLM_TW_FR.

The main results for French and English are presented in Figures 1 and 2 respectively (see more detailed results in appendix). The results confirm that these models can predict speech reductions to some extent in a spontaneous speech corpus. Then, the "conversational" and "spoken" data models appear to have some advantages over the Wikipedia-based ones, even though the differences were not statistically significant.[7] Finally, the topline performance of RoBERTa is clear for the French results.

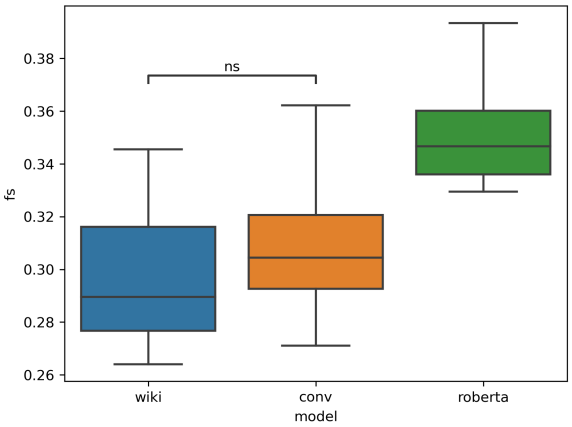

Figure 1: F-score comparing training data for predicting Speech reduction on CID corpus (RoBERTa as a top line). The significance between `wiki` and `conv` is not tested to be significant.

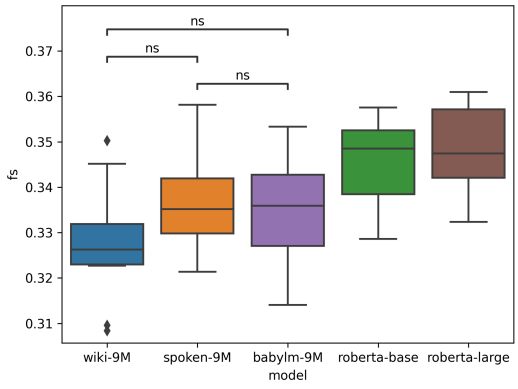

Figure 2: F-score comparing training data for predicting Speech reduction on the Buckeye corpus (RoBERTa models as top lines). The differences between models trained on 9M words were not significant.

5.2 Prosodic Prominences

To detect prosodically prominent tokens we used Suni et al.'s (2017) method based on wavelet that combines various acoustic features for determining prominence at the token level. One of the reasons for this tool choice is that it had been used already in the LMs literature (Wolf et al., 2023) to quantify the amount of redundancy between textual and prosodic levels. We used the default configuration of this tool and used a threshold score of 1.25 (See figure 9 in the appendix for details on the score values distribution). In the French data, this threshold amounted to 13.8% of the tokens labeled as prosodically prominent. In the English data, this threshold amounted to 14.7% of prosodically prominent tokens.

In both French and English experiments, the `conversational` and `spoken` models are significantly better than the `wiki` counterparts. RoBERTa models' topline performance is also clearer for both languages in this task.

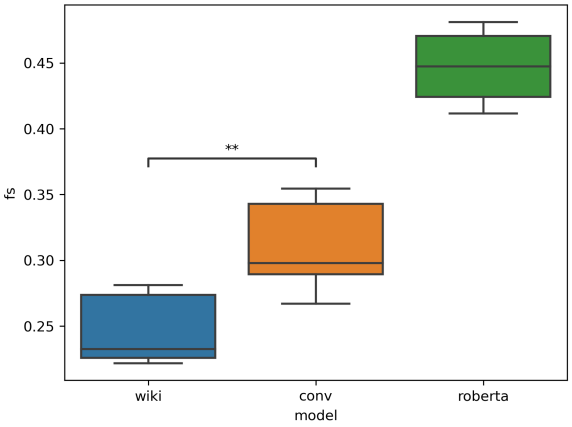

Figure 3: F-score comparing training data for predicting prosodically prominent tokens on CID corpus (RoBERTa as a sanity top line). **: p <= 0.01

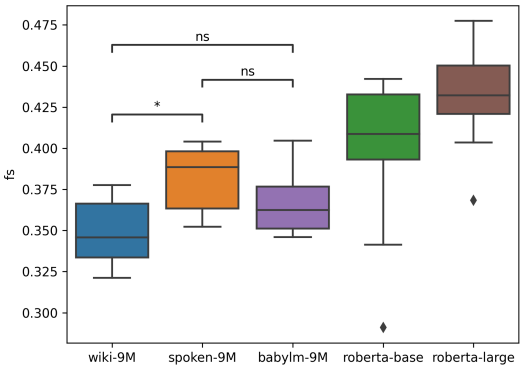

Figure 4: F-score comparing training data for predicting prosodically prominent tokens on the Buckeye corpus (RoBERTa models as top lines). *: p <= 0.05

[7] All statistical significances have been tested through a Mann-Whitney-Wilcoxon two-sided test.

5.3 Backchannels

We also designed, for the French benchmark only,[8] a task for predicting tokens around which a backchannel had been produced by the listener. To detect those, we used a simple list of tokens, eg. for French (*'mh', 'ouais', '@', 'ah', 'oui', 'bon', 'voilà', 'putain', 'accord', 'ben', 'oh', 'hum', 'eh', 'uh', 'OK'*). For each token of the target participant, we checked whether the other participant had produced one of these backchannel tokens in a time frame of $250ms$ before the beginning of the target token and $250ms$ after the end of the target token. This resulted in labeling 7.73% of tokens as being in the temporal vicinity of the listener's backchannels.

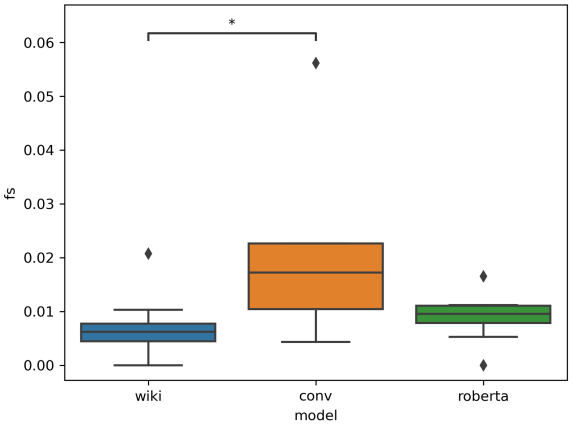

Figure 5: F-score comparing training data for predicting tokens overlapping a listener's backchannel in the CID corpus. *: p <= 0.05

As seen in figure 5, LLMs do not manage to solve this task with the data we gave them. While there is a statistically significant benefit for conversational pre-training (and in this case even over the bigger RoBERTa) the overall score does not go over 5% of f-score suggesting that none of these LLMs are getting close to modeling this phenomena. This is likely due to the nature of backchannelling: The literature points toward the contribution of lexico-syntactic cues to predict the end of turns, but the dominant cues remain prosodic ones, which these LLMs had no direct access to in their training data.

5.4 Testing models on BabyLM's zero-shot tasks

To examine whether models trained on spoken data can also be competitive in tasks that are not apparently tied to spoken language, we ran the English LMs[9] on the zero-shot classification tasks in BabyLM, i.e., BLiMP (Warstadt et al., 2020a) and EWoK (Ivanova et al., 2024), shown in Table 1. While the model trained on spoken data loses its advantage from our proposed reduction and prominence classification task and ranks the worst in the BLiMP supplement task, it is still competitive with other small models in filtered BLiMP and EWoK. Furthermore, the model trained on the BabyLM data, with a mixture of spoken and written materials, has the trend of outperforming the model trained on Simple Wikipedia both in our proposed tasks but also in BLiMP.

	BLiMP		EWoK
	supp.	filtered	filtered
RoBERTa-Large	71.9	73.9	65.5
RoBERTa-Base	70.3	74.3	62.9
BabyLM-9M	59.1	59.9	68.0
Wiki-9M	57.3	58.9	67.8
Spoken-9M	55.9	59.2	68.7

Table 1: English Models' performances in BLiMP & EWok

6 Potential shortcomings and Limitations

Information-centric nature. Our metrics are related to information-theoretic notions such as information density, entropy, and predictability. There is a substantial body of literature that demonstrates that these concepts can at least partially explain the phenomena discussed in the previous sections. This reminds us that information-theoretic measures, such as perplexity (a common LLM evaluation metric), are inherently connected to the variables we aim to predict. One potential limitation is that the models may only capture the information-theoretic contribution to our tasks. However, the prediction of these phenomena cannot be reduced to information-theoretic explanations alone. Each metric introduces its own set of subtleties related to language processing, and our goal is to evaluate LLMs in terms of their ability to grasp these subtleties.

Text-only. The phenomena we propose for probing the models are inherently related to speech processing, which goes beyond what

[8] This metric requires a truly conversational corpus with both parties accurately transcribed which is not the case of the English corpus used here.

[9] At the moment, we still lack similar benchmarks for French to do the same with our French LMs.

can be achieved with a text-only approach. Beyond the acoustic modality, the visual channel also plays a role, especially in contributing to backchannels. However, it is possible to limit multimodality to just text and speech by excluding face-to-face corpora from the benchmark. Our goal in proposing these metrics is not to achieve state-of-the-art performance in predicting these phenomena. Rather, we aim to treat them as *"traces"* of human language processing visible at the surface level, and to test which models are better at predicting these traces from text-only input.

Surface level shortcuts. A concern related to the previous point is the risk that models rely on surface-level elements as shortcuts to predict the variables we are targeting. While we do not have a definitive solution to this issue, since the nature of our metrics involves performance details observable in surface forms, we believe it is still worth pursuing this line of investigation. If the approach behaves consistently across our range of proposed metrics and languages, it may provide valuable information for language model evaluation. The next step will be to build controlled evaluation sets, similar to those developed in McCoy et al. (2019), that allow the exclusion of surface-level confounds in a principled way.

Triviality of the main result. From a machine-learning perspective, it might be seen as a trivial result that models trained on data similar to test sets perform better than models trained on other types of data. First of all, it is worth emphasizing that pretraining datasets and benchmarks in our experiments are completely independent as they do not come from the same raw corpora. Also, the pretraining datasets and corpora for building benchmarks have been curated by different teams and transcribed with different conventions. Nevertheless, we cannot deny that the conversational datasets are by all aspects (sentence length distribution, lexical frequencies, etc) more similar to benchmarks than Wikipedia datasets are.

As trivial as it seems, it may be one of our main points: to produce models more closely related to human cognition, one should use data sets made of spontaneous speech (and not generic textual / web content). The fact that ROBERTA outperforms all models does not change this fact since ROBERTA is trained on a dataset several orders of magnitude bigger.

7 Conclusion and Roadmap

In this position paper, we advocate for advancing the BABYLM initiative in several key areas. First, expanding beyond English is both necessary and feasible, given the initiative's design centered on "small-scale" data sets. Here we used French as an example, but we have also built the Mandarin equivalent datasets[10], emphasizing the importance of multilingual perspectives. Our proposal focuses on using training data composed entirely of spontaneous speech transcripts, which offers insights into language learning processes. It will be crucial to explore more nuanced variations in training data, such as balancing conversational speech, child-directed speech, and simple texts. Equally important is the development of complementary evaluation metrics. We propose using spontaneous speech data to benchmark models and assess linguistic phenomena, such as speech reductions, prosodic prominences, and backchannel responses, as key indicators of human language processing.

For the time being, we have English, Mandarin, and French training datasets with different data mixtures. The next steps involve systematizing the pilot experiments on speech reductions conducted here for the Mandarin dataset. Then, we will extract all the other proposed metrics for the benchmark datasets. Through this expanded set of experiments, we aim to demonstrate the value of the proposed approach and generalize it to other linguistic phenomena. In a broader perspective, we hope to show that benchmarks like BLIMP that require a significant amount of expert and naive human input to build, can be complemented with benchmarks derived from the numerous existing high-quality linguistic corpora, without additional human efforts.

Acknowledgments

We would like to thank Zoe Naud and Ri-Sheng for discussions in the context of this work.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of

[10]For Mandarin, we used Gutenberg, The NCCU Spoken Corpus of Spoken Taiwan Mandarin (Chui and Lai, 2008), Taiwan Corpus of Child Mandarin[11], Chinese Wikipedia and Open Subtitles.

sentence embeddings using auxiliary prediction tasks. In *Proceeddings of ICLR Conference Track*, Toulon, France.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1):31–56.

Lisa Beinborn and Yuval Pinter. 2023. Analyzing cognitive plausibility of subword tokenization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alan Bell, Jason M Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1):92–111.

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fmri-with an application to pos induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. 2017. The corpus of interactional data: A large multimodal annotated resource. *Handbook of linguistic annotation*, pages 1323–1356.

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Francesco Cabiddu, Mitja Nikolaus, and Abdellah Fourtassi. 2025. Comparing children and large language models in word sense disambiguation: Insights and challenges. *Language Development Research*, 5(1).

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*.

Morten H Christiansen and Nick Chater. 2022. *The language game: How improvisation created language and changed the world*. Random House.

Kawai Chui and Huei-ling Lai. 2008. The nccu corpus of spoken chinese: Mandarin, hakka, and southern min. *Taiwan Journal of Linguistics*, 6(2).

Uriel Cohen Priva. 2012. *Sign and signal: Deriving linguistic generalizations from information utility*. Ph.D. thesis, Stanford University.

Uriel Cohen Priva and T Florian Jaeger. 2018. The interdependence of frequency, predictability, and informativity in the segmental domain. *Linguistics Vanguard*, 4(s2).

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp*, pages 134–139.

Susanne Gahl. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3):474–496.

Susanne Gahl, Yao Yao, and Keith Johnson. 2012. Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language*, 66(4):789–806.

Youssef Hmamouche, Magalie Ochs, Laurent Prévot, and Thierry Chaminade. 2024. Interpretable prediction of brain activity during conversations from multimodal behavioral signals. *Plos one*, 19(3):e0284342.

Nora Hollenstein, Emmanuele Chersoni, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. Cmcl 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Sofoklis Kakouros and Johannah O'Mahony. 2023. What does bert learn about prosody? In *20th International Congress of Phonetic Sciences ICPhS*. International Phonetics Association.

Stephen C Levinson. 2020. On the human "interaction engine". In *Roots of human sociality*, pages 39–69. Routledge.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le petit prince multilingual naturalistic fmri corpus. *Scientific data*, 9(1):530.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. Climb–curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Christine Meunier and Robert Espesser. 2011. Vowel reduction in conversational speech in french: The role of lexical factors. *Journal of Phonetics*, 39(3):271–278.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

Alexandre Pasquiou, Yair Lakretz, John T Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*, pages 17499–17516. PMLR.

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

Mark Pluymaekers, Mirjam Ernestus, and R Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

Birgit Rauchbauer, Bruno Nazarian, Morgane Bourhis, Magalie Ochs, Laurent Prévot, and Thierry Chaminade. 2019. Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B*, 374(1771):20180033.

William D Raymond, Robin Dautricourt, and Elizabeth Hume. 2006. Word-internal/t, d/deletion in spontaneous speech: Modeling the effects of extralinguistic, lexical, and phonological factors. *Language variation and change*, 18(1):55–97.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89.

Yvan Rose and Brian MacWhinney. 2014. The phonbank project: Data and software-assisted methods for the study of phonology and phonological development.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L Frank. Blimp-nl.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, and Ekaterina Artemova. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*.

Kevin Tang and Ryan Bennett. 2018. Contextual predictability influences word and morpheme duration in a morphologically complex language (kaqchikel mayan). *The Journal of the Acoustical Society of America*, 144(2):997–1017.

RJJH Van Son and Louis CW Pols. 2003. How efficient is speech. In *Proceedings of the institute of phonetic sciences*, volume 25, pages 171–184.

RJJH Van Son, Louis CW Pols, et al. 1999. Effects of stress and lexical structure on speech efficiency. In *EUROSPEECH*.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *Calico Journal*, 26(3):474–490.

Sheng-Fu Wang. 2022. The interaction between predictability and pre-boundary lengthening on syllable duration in taiwan southern min. *Phonetica*, 79(4):315–352.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng Fu Wang, and Samuel R. Bowman. 2019a. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

A Variables to predict distributions and thresholds

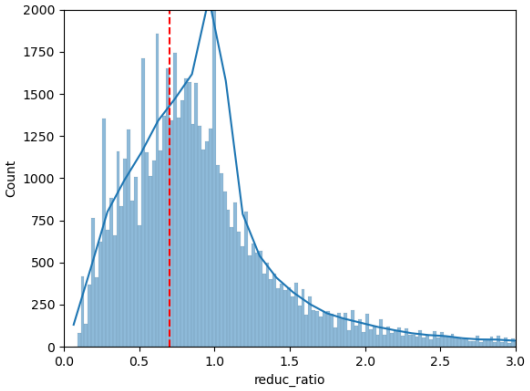

Figure 6: Distribution reduction ratios as calculated in the French Dataset and the threshold selected.

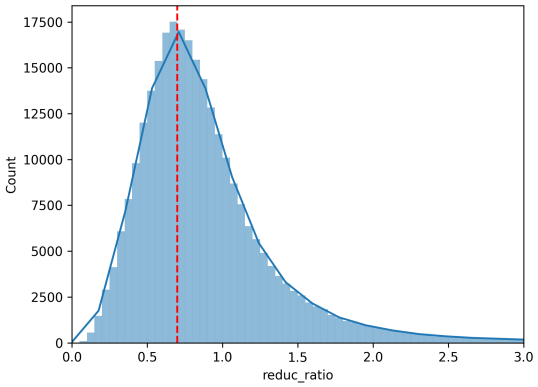

Figure 7: Distribution reduction ratios as calculated in the English Dataset and the threshold selected.

B Complete score tables

Language	Task	Model	F1	Precision	Recall
French	reduction	wiki-10M	.298 (.030)	.299 (.035)	.298 (.029)
		conv-10M	.310 (.029)	.300 (.033)	.321 (.030)
		XLM-Roberta-Base	.352 (.023)	.342 (.027)	.363 (.020)
	prominence	wiki-10M	.246 (.026)	.282 (.028)	.219 (.027)
		conv-10M	.311 (.033)	.356 (.039)	.277 (.033)
		XLM-Roberta-Base	.446 (.029	.503 (.040)	.403 (.033)
	backchannel	wiki-10M	.007 (.006)	.004 (.004)	.040 (.023)
		conv-10M	.020 (.016)	.014 (.014)	.057 (.025)
		XLM-Roberta-Base	.009 (.005)	.006 (.004)	.024 (.017)
English	reduction	Wiki-9M	.327 (.013)	.322 (.014)	.334 (.022)
		Spoken-9M	.336 (.012)	.333 (.011)	.340 (.019)
		BabyLM-9M	.335 (.012)	.331 (.013)	.340 (.020)
		Roberta-Base	.345 (.010)	.345 (.014)	.345 (.016)
		Roberta-Large	.349 (.009)	.343 (.011)	.355 (.015)
	prominence	Wiki-9M	.349 (.019)	.405 (.041)	.311 (.035)
		Spoken-9M	.382 (.020)	.453 (.046)	.333 (.028)
		BabyLM-9M	.366 (.018)	.437 (.045)	.318 (.029)
		Roberta-Base	.398 (.049)	.499 (.044)	.336 (.060)
		Roberta-Large	.431 (.030)	.488 (.057)	.392 (.046)

Table 2: Full results on the proposed speech-based benchmarks

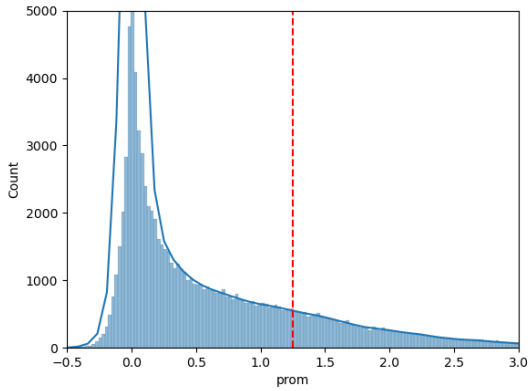

Figure 8: Distribution of prominence score as calculated in the French Dataset and the threshold selected.

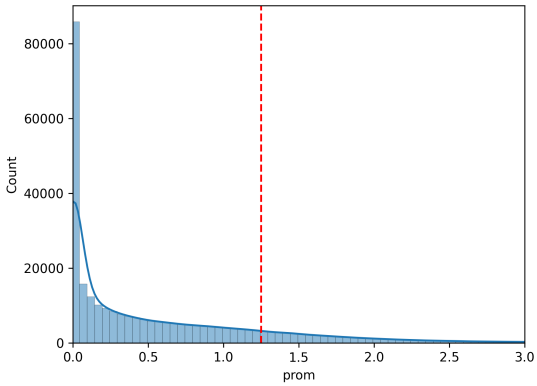

Figure 9: Distribution of prominence score as calculated in the English Dataset and the threshold selected.