BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation

Matilde Barbini[5,2,*], Maria Letizia Piccini Bianchessi[2,†], Veronica Bressan[3,2,†], Achille Fusco[4,2,†], Sofia Neri[1,2,†], Sarah Rossi[1,2,†], Tommaso Sgrizzi[1,2,†] and Cristiano Chesi[1,2]

[1]*University School for Advanced Studies IUSS Pavia, Palazzo del Broletto. Piazza della Vittoria, 15 - 27100 Pavia*

[2]*NeTS Lab, IUSS Pavia, Palazzo del Broletto. Piazza della Vittoria, 15 - 27100 Pavia*

[3]*Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Fondamenta Tofetti 1075, 30123 Venice*

[4]*University of Florence*

[5]*EPFL Lausanne Doctoral Program Digital Humanities EDDH - Social Computing Group*

Abstract

In this work we introduce the automatically generated dataset in BLiMP-IT, a novel benchmark for evaluating Italian language models based on minimal pairs (i.e. sentence pairs that differ only in a critical morphosyntactic aspect). Drawing inspiration from the success of BLiMP for English, BLiMP-IT combines and adapts several existing resources—including COnVERSA, AcCompl-it, and BLiMP—to construct a high-quality evaluation dataset for Italian. We present an automatic methodology for generating the evaluation's items by leveraging a large Italian corpus for lexicon extraction, POS tagging, and animacy annotations. Our approach not only ensures coverage of diverse morphosyntactic phenomena (e.g., agreement and inflection, verb class, non-local dependencies) but also scales the creation of minimal pairs to automatically expand the items for the evaluation benchmark. BLiMP-IT demonstrates that an automated pipeline for generating minimal pairs to evaluate LMs is both feasible and effective, ensuring comprehensive coverage of diverse morphosyntactic phenomena in Italian while reducing reliance on manual annotation.

Keywords

Computational Linguistics, Automatic Sentence Generation, Language Model Evaluation, Linguistic Benchmarks

1. Introduction

The development of benchmarks and datasets for the linguistic evaluation of Language Models (LMs) in a specific language is essential for a systematic assessment of their ability to handle its morphosyntactic structures. Given cross-linguistic variation in inflectional morphology, syntactic configurations, agreement mechanisms, and word order flexibility, language models often exhibit differential performance depending on the structural properties of the target language. A dedicated evaluation framework allows for rigorous analysis of morphosyntactic accuracy, including the handling of inflectional paradigms, syntactic dependencies, agreement constraints, and constituent ordering, providing a comprehensive assessment of a model's grammatical competence. In linguistic theory acceptability judgments have been often defined as the main empirical method used to access human linguistic competence and language acquisition [1, 2]. This methodology has also been proved to be a classical and reliable tool for assessing the linguistic capabilities of LMs across various linguistic phenomena [3, 4, 5, 6]. A common methodology is the employment of minimal pairs, couples of sentences differing minimally in their structure, with one being grammatically acceptable and the other one being unacceptable. An effective LM should assign higher probabilities to grammatically acceptable sentences than to their unacceptable counterparts. Alternatively, it can be evaluated by presenting a series of sentences—both grammatical and ungrammatical—and requiring the model to perform a binary acceptability classification. While benchmarks such as BLiMP have provided valuable insights for English, the lack of analogous resources for Italian poses a challenge for multilingual NLP and for an effective and comprehensive evaluation of these models. We address this gap by introducing BLiMP-IT [1], a benchmark specifically designed for Italian. Our contributions are twofold:

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

*Corresponding author.

†These authors contributed equally.

matilde.barbini@iusspavia.it (M. Barbini); letizia.piccinibianchessi@iusspavia.it (M. L. P. Bianchessi); veronica.bressan@iusspavia.it (V. Bressan); achille.fusco@iusspavia.it (A. Fusco); sofia.neri@iusspavia.it (S. Neri); sarah.rossi@iusspavia.it (S. Rossi); tommaso.sgrizzi@iusspavia.it (T. Sgrizzi); cristiano.chesi@iusspavia.it (C. Chesi)

0009-0007-7986-2365 (M. Barbini); 0009-0005-8116-3358 (M. L. P. Bianchessi); 0000-0003-3072-7967 (V. Bressan); 0000-0002-5389-8884 (A. Fusco); 0009-0006-7830-6669 (S. Neri); 0009-0007-2525-2457 (S. Rossi); 0000-0003-1375-1359 (T. Sgrizzi); 0000-0003-1935-1348 (C. Chesi)

[1]Forthcoming in *Proceedings of GLOW 47*. The resources for BLiMP-IT can be found at https://nets-lab.github.io/blimpit/

- **Resource Adaptation and Assembly:** We construct BLiMP-IT by integrating and adapting existing Italian and English datasets and benchmarks for the linguistic evaluation of LMs, within a minimal pairs' framework.
- **Automatic Minimal Pair Generation:** We develop an automated pipeline for generating minimal pairs by extracting a detailed lexicon from a large Italian corpus, tagging it with linguistic information (e.g., POS, UPOS, animacy), and systematically mapping various linguistic phenomena to unique sequence tags, to produce both grammatical and ungrammatical sentence pairs (i.e. minimal pairs) [2].

In this work, we focus on the automatic pipeline component of the BLiMP-IT resource, providing a comprehensive description of its operational workflow.

2. Related work

Large Language Models (LLMs) have sparked an ongoing debate about whether they develop genuine linguistic competence or rely primarily on spurious statistical generalizations [7, 8]. This fundamental question is complicated by LLMs' opacity in processing language patterns and their tendency to conflate world knowledge with morphosyntactic competence [9]. While some interpret LLMs' performance with complex grammatical configurations as evidence against the Poverty of Stimulus hypothesis [10], critics note that such results depend on dramatically oversized training data compared to child language acquisition [11]. Moreover, higher performance on increasingly specific tasks does not always correspond to genuine gains in linguistic understanding [12], suggesting that standard performance metrics may inadequately capture linguistic competence [1]. Within this context, developing linguistically informed benchmarks has become crucial for evaluating model performance and the nature of their competence [13]. The evaluation of language models via acceptability judgments and minimal pairs has a long-standing history in theoretical and computational linguistics. Recent benchmarks such as BLiMP [14] and CLiMP [15] have demonstrated the value of this approach, while recent shared tasks have highlighted how small-sized training regimes (10-100M tokens) can achieve relatively good results on various linguistic benchmarks including BLiMP and CoLA [16, 14]. However, the most performant architectures that show improvement with additional training often yield diminishing returns in psycholinguistic terms [17]. Recent work capitalizing on the BabyLM Challenge in English [18] and similar tasks in Italian [19] has stressed the importance of adopting appropriate linguistic benchmarks to meaningfully challenge the Poverty of Stimulus hypothesis. For Italian specifically, resources like Laccolith [20] and AcCompl-it [21] have targeted acceptability judgments through binary and rating-based methods. However, there remains a need for a comprehensive Italian benchmark that harnesses the minimal pairs framework, a gap that BLiMP-IT aims to fill.

3. BLiMP-IT Dataset Construction

3.1. Minimal Pairs Framework

The minimal pairs framework adopted in BLiMP-IT centers on constructing sentence pairs that differ only in a critical grammatical feature. One sentence in the pair is grammatically acceptable, while the other violates a specific morphosyntactic rule. This approach builds on previous work in linguistic evaluation, notably the BLiMP benchmark for English (e.g., [14]) and provides a fine-grained measure of a language model's sensitivity to subtle grammatical contrasts. Minimal pairs serve as a fine-grained diagnostic tool: by presenting a model with two sentences that are identical except for one grammatical feature, researchers can assess whether the model is sensitive to the relevant linguistic distinction. For example, in the case of subject-verb agreement, a model should consistently assign higher probability or acceptability to the correct agreement form (e.g., "La ragazza mangia la mela" vs. "La ragazza mangiano la mela") [3]. This controlled setup eliminates confounding variables and allows precise measurement of model performance on particular phenomena. To ensure interpretability and reproducibility, BLiMP-IT constructs minimal pairs based on abstract tag templates that encode both grammatical and ungrammatical structures. These templates are manually designed and systematically mapped to lexical entries drawn from a linguistically annotated corpus. The use of tag-based generation not only facilitates large-scale pair creation but also guarantees that the only difference between the sentences in a pair is the grammatical target under investigation. The minimal pairs are organized around four major categories of morphosyntactic phenomena: Agreement and Inflection, Verb Class and Argument Structure, Pronouns, and Non-local Dependencies. Each pair is associated with a specific sub-phenomenon (e.g., determiner-noun agreement, reflexive clitic placement, long-distance wh-dependencies), enabling detailed evaluation across diverse syntactic domains. In design-

[2] The automatically generated resources, as well as a flowchart describing the process, can be accessed at https://nets-lab.github.io/blimpit-generation/. Please note that these data are provisional and subject to ongoing generation and refinement.

[3] "The girl eats the apple" vs. *"The girl eat the apple"

ing these pairs, particular attention was paid to structural symmetry, lexical consistency, and plausibility. Sentences were constructed to be semantically neutral where possible, to avoid introducing biases unrelated to the grammatical phenomenon. This was especially important for more complex structures, such as those involving coordination or wh-movement, where maintaining interpretability across grammatical and ungrammatical variants can be challenging. Finally, minimal pair evaluation supports both probabilistic scoring (e.g., comparing log-likelihoods assigned by a language model) and binary classification tasks, such as acceptability judgments. This flexibility allows BLiMP-IT to be used with a wide range of language models and evaluation metrics, aligning with the goals of interpretability and cross-model comparability.

3.2. BLiMP-IT: Integrated Resources

BLiMP-IT encompasses 78 morphosyntactic phenomena, which are categorized into four main groups: Agreement and Inflection (including phenomena such as noun-determiner and subject-verb agreement), Verb Class and Argument Structure (addressing issues like auxiliary selection and θ-role assignment), Pronouns (focusing on clitics, reflexives, and person agreement), and Non-local Dependencies (encompassing long-distance dependencies and island effects).

The dataset is constructed by integrating multiple existing Italian linguistic resources (and English resources in the case of BLiMP) while also incorporating newly created minimal pairs. Our sources include:

- COnVERSA: A battery designed for assessing grammaticality through minimal pairs [22].
- AcCompl-it: An evaluation campaign component focused on acceptability and complexity judgments [21].
- BLiMP: a test set for evaluating the grammatical knowledge of English LLMs, featuring 67 minimal pair paradigms across 12 categories [14].
- New phenomena: a set of new linguistic phenomena such as ATB [23] and parasitic gaps (inspired by [24]).

The adaptation process involved selecting phenomena that are central to Italian grammar (e.g., noun-determiner agreement, subject-verb agreement, verb argument structure, clitic usage, and non-local dependencies) and reformulating the examples to align with the minimal pairs methodology. For instance, items from English BLiMP, if compatible with and relevant for Italian morphosyntax, were carefully translated and restructured to account for Italian-specific syntactic and morphological features.

4. BLiMP-IT: automated generation

4.1. Corpus Creation for Lexicon Extraction

A fundamental component of our automatic generation pipeline is the creation of a large high-quality Italian dataset, initially developed to take part to the BabyLM challenge [25], which consists of approximately 3 million tokens sourced from diverse resources and serves as the foundation for lexicon extraction. It is divided into five sections: child-directed speech (CHILDES Italian section), child movie subtitles (from OpenSubtitles), child songs (from the Zecchino D'Oro repository), telephone conversations (VoLIP corpus, [26], and fairy tales (from copyright-expired sources). After a cleaning process that removed metalinguistic annotations and children's productions, the corpus contains 2,431,038 tokens with an overall Type-Token Ratio (TTR) of 0.03. The distribution of tokens across sections is as follows: CHILDES (346,155 tokens, TTR = 0.03), SUBTITLES (700,729 tokens, TTR = 0.05), CONVERSATIONS (58,039 tokens, TTR = 0.11), SONGS (222,572 tokens, TTR = 0.08), and FAIRY TALES (1,287,826 tokens, TTR = 0.05). Statistical analysis of the corpus ensures sufficient lexical diversity and coverage of the linguistic phenomena under investigation.

4.2. Lexicon Extraction and Linguistic Tagging

We extract a lexicon from the corpus that captures key linguistic attributes for each word. First, we annotate words with both POS and UPOS tags using state-of-the-art taggers (spaCy). In addition, we manually labeled nouns with animacy information to address semantic nuances. This lexicon forms the basis for selecting appropriate words when generating minimal pairs.

4.3. The pipeline for minimal pairs generation

Our automatic minimal pair generation process follows a structured and modular pipeline designed to produce large-scale, linguistically controlled sentence pairs. This section details each stage of the pipeline, emphasizing both the design rationale and the implementation steps.

- **Resource loading:** The process begins with the loading of two key components: (i) a lexicon extracted from the Italian corpus, enriched with linguistic annotations such as part-of-speech (POS), universal POS (UPOS), animacy, and morphological features; and (ii) a set of tag sequences, each defining the structure of a sentence in terms of

syntactic categories. These tag sequences are constructed in minimal pairs, where each pair consists of a grammatical and an ungrammatical variant. The ungrammatical variant introduces a targeted morphosyntactic violation (e.g., a mismatched subject-verb agreement or incorrect determiner-noun concord), ensuring that the only difference between the two sequences is the critical grammatical contrast under investigation. This design supports a controlled evaluation of model sensitivity to specific phenomena.

- **Tag Matching and Word Selection:** Once the tag templates are loaded, the system proceeds to match each tag in a sequence with a suitable word from the lexicon. Word selection is guided by the required grammatical features encoded in the tag (e.g., number, gender, animacy, tense). To prevent repetition and encourage lexical diversity, a tracking mechanism records previously selected tokens and prioritizes less frequently used words when possible. Special handling is applied to verbs, which require agreement features to be matched precisely with their subject counterparts. The system identifies verb roots and selects appropriate inflected forms based on number and person. Additionally, animacy plays a role in selecting nouns and pronouns, especially in structures where semantic compatibility influences grammaticality (e.g., reflexive pronouns or clitic constructions). If a matching lexical item cannot be found for a given tag within the constraints, the system either retries with an alternative lexeme or skips the current sequence to maintain sentence well-formedness and overall dataset quality.

- **Sentence Construction:** With the tag-to-word mappings established, the system constructs sentence pairs by linearizing the selected tokens according to their tag sequence order. Minimal surface normalization is performed at this stage, including the insertion of appropriate punctuation, handling of elisions and contractions, and capitalization of the sentence-initial token. Each sentence is generated in parallel with its minimal counterpart, ensuring that both share identical lexical items and structure, differing only in the targeted morphosyntactic element. This parallelism ensures the interpretability and diagnostic value of each pair.

- **Iterative Generation and Quality Control:** To ensure dataset diversity and minimize redundancy, the pipeline includes a control mechanism to detect and filter out duplicate or near-duplicate sentence pairs. Duplicates are identified not only by surface form but also by underlying tag structure, preventing syntactically redundant examples from being overrepresented. The generation process is iterative: multiple passes are performed over the tag templates and lexicon, dynamically adjusting word choices based on availability and prior usage. When generation fails (e.g., no valid word found for a required combination), the system logs the instance and skips the pair to avoid compromising the grammatical precision of the dataset. Internally, each generated (good-sentence, bad-sentence) tuple is stored in a Python set and tested for membership in O(1) time: any exact surface-form repeat is skipped. To prevent an endless loop when unique pairs run out, the loop also caps the total number of attempts (e.g., 10× the target) and logs a warning if it cannot reach the requested count.

- **Quality check:** We employ a human-in-the-loop strategy, where a team of linguistic experts meticulously reviews the generated minimal pairs to ensure grammatical accuracy and naturalness. Each pair is independently rated by at least two reviewers and any doubts trigger a discussion session to reach consensus and to establish if the pair must be removed. Experts also log error types and provide targeted feedback on problematic tag sequences or lexicon entries.Their expertise not only enhances the overall quality of our evaluation tool but also ensures inter-rater reliability, fostering consistency and objectivity in the assessment process.

4.4. Methodological challenges

While the automatic generation pipeline described above enables scalable creation of minimal pairs, its implementation also revealed several methodological challenges that required careful consideration. First, the process of animacy annotation introduced a bottleneck due to the need for manual labeling. Although part-of-speech (POS) and universal POS (UPOS) tags could be obtained using existing NLP tools such as spaCy, the classification of nouns and pronouns based on animacy required human intervention. This task is particularly sensitive in Italian, where animacy can influence grammaticality judgments, especially in constructions involving clitics, reflexives, or subject-verb agreement. Ensuring consistent annotation across the lexicon was essential to preserve the validity of minimal pairs involving semantically conditioned structures. Second, the construction of sequence tags—representing grammatical and ungrammatical syntactic structures—proved complex. Tag se-

quences must encode subtle contrasts in grammaticality while remaining compatible with the lexicon and word selection rules. Designing these templates required extensive linguistic knowledge and iterative refinement. In some cases, identifying minimal but meaningful structural contrasts demanded revisiting the theoretical underpinnings of the targeted phenomenon. Another critical challenge was matching lexical items to abstract tag templates. While the lexicon provides detailed linguistic annotations, finding appropriate word combinations that meet all morphological and syntactic constraints was nontrivial. This was especially true for verbs, where selecting appropriate inflected forms (e.g., singular/plural, tense, auxiliary selection) required tracking agreement features and root compatibility. Additionally, ensuring lexical diversity while avoiding repetitive or unnatural constructions added further complexity to word selection. The generation process also involved quality control mechanisms to filter out low-quality or duplicate pairs. Despite automated checks, certain errors—such as overly rigid or implausible sentences—could only be caught through manual review. This underscores the continued importance of human-in-the-loop validation, particularly for capturing edge cases that automatic systems may overlook. Finally, the reliance on a corpus of child-directed speech and simplified texts (developed for the BabyLM Challenge) had implications for lexical diversity. While the corpus offered controlled and well-annotated input data, its domain-specific nature may limit coverage of more formal or idiomatic constructions. Addressing this limitation requires expanding the source corpus in future iterations to include a broader range of registers and genres.

5. Results

Our pipeline successfully generated 2,899 minimal pairs covering 18 phenomena—spanning agreement, non-local dependencies, and other key categories—from the 78 phenomena included in BLiMP-IT. We are actively working to expand this coverage to include all 78 phenomena. Following the methodology proposed for English in [18], early findings from employing BLiMP-IT to assess models that replicate the constraints children face while learning language show that strong performance on standard evaluation metrics doesn't translate to equally strong results on minimal pair tests, and these models fail to capture the linguistic patterns typical of children [19]. These initial findings indicate that children's language learning follows expected linguistic principles, while large language models demonstrate inconsistent behavior. Specifically, preliminary results [4] reveal that although training different language models (GPT-2, BERT, ad hoc RNN) on approximately 10 million tokens increases overall accuracy (rising from 40% to 79%), its performance on certain BLiMP-IT components actually worsens (dropping from 61% to 52%). The models' reliability in distinguishing correct from incorrect language forms decreases from 44% to 32%, falling short of human benchmarks (around 86% accuracy and 72% consistency observed in seven-year-old children). We are still in the process of testing and evaluating different models on our automatically-generated minimal pairs.

[4]Forthcoming in Proceedings of GLOW 47

Macro-phenomena	Phenomena	Micro-phenomena	Source
Agreement and Inflection	A1. D-N	A1. Num-N, num; D_def-N, num; D_indef-N, num	A1. COnVERSA
Non-local dependencies	1. wh-island_root 2. adjunct_island 3. complex NP island 4. sentential subject island 5. coordinate structure constraint_complex_object_extraction 6. Left_branch_island_echo_question 7. parasitic gap/adjunct island1 8. parasitic gap/adjunct island2 9. wh- island_embedd 10. wh-extraction_embedd 11. wh-extraction_embedd2 12. ATB_affirmative1 13. ATB_affirmative2 14. ATB_interrogative 15. RC-subject 16. RC-object 17. ATB_RC_object	1. affirmative; affirmative_dove 2. - 3. - 4. - 5. - 6.- 7. - 8. - 9. - 10. clitic_inanimate; NP_inanimate; NP_inanimate_dem 11. clitic_inanimate; NP_inanimate; NP_inanimate_dem 12. mi; ti 13. - 14. nogap_gap_clitic; gap_nogap_clitic; nogap_nogap_clitic; gap_nogap_NP_aux; nogap_gap_NP_aux; nogap_nogap_NP_aux 15. subject_nogap; subject_attraction 16. object_nogap; object_attraction 17. gap_nogap; nogap_gap; nogap_nogap	1. AcCompl-it 2. adapted from BLiMP 3. adapted from BLiMP 4. adapted from BLiMP 5. adapted from BLiMP 6. adapted from BLiMP 7. new (inspired by Lan et al., 2024) 8. new (inspired by Lan et al., 2024) 9. AcCompli-it 10. AcCompl-it 11. adapted from AcCompl-it 12. new (inspired by Lan et al., 2024) 13. new (inspired by Lan et al., 2024) 14. AcCompl-it 15. new (cf. BLiMP) 16. new (cf.BLiMP) 17. AcCompl-it

Figure 1: The linguistic phenomena (with different levels of granularity) reflected in the automatically generated minimal pairs. A detailed description of the phenomena and the acronyms, with relevant references, can be found at https://nets-lab.github.io/blimpit-generation/

6. Discussion

BLiMP-IT represents a significant step forward in the evaluation of Italian language models by providing a benchmark that combines manually curated linguistic phenomena with an innovative pipeline for automatic minimal pair generation. Through the integration of diverse resources and a structured methodology, our approach ensures both linguistic relevance and scalability. One of the strengths of our approach lies in the combination of curated content and automation. While the manual adaptation of resources such as COnVERSA and AcCompl-it guarantees that the dataset reflects core aspects of Italian grammar, the automated generation pipeline makes it possible to scale the number of minimal pairs efficiently and consistently. This dual strategy enables us to address a broader range of morphosyntactic phenomena while maintaining control over the

grammatical integrity of the examples. Moreover, by implementing a human-in-the-loop quality control process, we ensure that automatically generated sentence pairs remain grammatically accurate and linguistically natural. Linguistic experts systematically validate the outputs, which strengthens the internal consistency of the dataset and enhances its reliability for downstream evaluation tasks. This step is crucial given the complexity of Italian syntax and morphology, where subtle changes in word form or word order can significantly affect acceptability. Another key contribution of BLiMP-IT is its focus on minimal pairs as an evaluation methodology. This approach provides a fine-grained tool for testing specific grammatical contrasts, such as subject-verb agreement or clitic placement, that are often underrepresented in broader benchmarks. By isolating individual linguistic features, BLiMP-IT allows researchers to probe the syntactic sensitivity of language models in a controlled and interpretable way. The breadth of phenomena included in BLiMP-IT, spanning from local agreement patterns to long-distance dependencies, also makes it a valuable diagnostic resource. In particular, the inclusion of lesser-tested constructions such as parasitic gaps or ATB (Across-The-Board) movement contributes to a more comprehensive picture of a model's grammatical competence. This is especially important in the context of evaluating transformer-based models, which may succeed in surface-level generalizations but struggle with deeper syntactic dependencies. Furthermore, the design of BLiMP-IT allows for ongoing extension and refinement. Since the core generation pipeline is modular, it can be expanded to incorporate additional phenomena as more linguistic data becomes available. The current focus on 18 phenomena, though already substantial, represents only a subset of the 78 phenomena identified in the full benchmark framework. Ongoing work is directed toward increasing this coverage while maintaining the same level of quality control. Finally, by grounding our dataset in a linguistically annotated corpus developed for the BabyLM Challenge, we ensure that our lexical and syntactic inputs are well-attested and systematically organized. Although this corpus primarily reflects child-directed language, it still provides sufficient lexical and morphosyntactic variety to generate a diverse and representative set of sentence pairs. The detailed analysis of type-token ratios across subdomains (e.g., fairy tales, songs, subtitles) confirms that the source material supports the goals of minimal pair generation in a linguistically meaningful way.

7. Conclusions

We have presented BLiMP-IT, a novel evaluation benchmark for Italian language models that integrates curated linguistic resources with an automated pipeline for minimal pair generation. This hybrid methodology allows us to systematically and efficiently generate sentence pairs that test key morphosyntactic competencies—such as agreement, inflection, verb argument structure, and non-local dependencies—across 78 targeted phenomena. Our approach ensures scalability while maintaining high linguistic quality through expert validation. The contribution of BLiMP-IT is twofold: first, it addresses the significant gap in Italian-specific evaluation datasets for language models, and second, it proposes a generalizable, language-agnostic framework for benchmark construction. These features make BLiMP-IT a valuable tool not only for evaluating existing LMs, but also for supporting their training and fine-tuning—particularly in low-resource or developmentally plausible settings, such as those promoted by the BabyLM challenge. The automatic generation pipeline opens the door for large-scale, consistent, and reusable evaluation items, minimizing the reliance on manual crafting, which is both time-consuming and difficult to scale. This makes it feasible to evaluate a wide range of grammatical contrasts in a way that is both linguistically informed and computationally practical. Looking forward, we aim to expand coverage to all 78 phenomena, increase the lexical and syntactic diversity of the generated items, and incorporate more advanced linguistic annotations, such as semantic roles and animacy, using semi-supervised or model-assisted techniques. Additionally, we plan to develop a fully language-independent version of the pipeline, enabling researchers to create similar benchmarks for other morphologically rich languages. By combining linguistic depth with computational scalability, BLiMP-IT sets a new standard for targeted evaluation of linguistic competence in Italian language models and offers a blueprint for multilingual benchmarking in the future.

8. Limitations

As discussed in Section 4.4, several methodological challenges were encountered during the design of the automatic generation pipeline. In addition to those, our current setup faces broader limitations that affect the dataset's generalizability and scalability. Most notably, the underlying corpus was originally developed for the BabyLM Challenge and, as such, is largely composed of texts classified as 'child-directed speech'. This focus limits the diversity of the lexicon used for minimal pair creation and may not fully represent the broader spectrum of language registers. In future work, we plan to extend our dataset to incorporate a wider range of text sources, thereby enriching the lexicon and enhancing representativeness. Additionally, our current pipeline relies on manual processes for animacy annotation and the

construction of sequence tags. This dependency on manual efforts introduces potential inconsistencies and limits scalability. We aim to transition to a fully automated approach in subsequent iterations, which will improve both the reliability and efficiency of our pipeline.

References

[1] N. Chomsky, Aspects of the Theory of Syntax, 11, MIT press, 2014.

[2] C. T Schütze, The empirical base of linguistics: Grammaticality judgments and linguistic methodology, Language Science Press, 2016.

[3] T. Linzen, E. Dupoux, Y. Goldberg, Assessing the ability of lstms to learn syntax-sensitive dependencies, Transactions of the Association for Computational Linguistics 4 (2016) 521–535.

[4] R. Marvin, T. Linzen, Targeted syntactic evaluation of language models, arXiv preprint arXiv:1808.09031 (2018).

[5] E. Wilcox, R. Levy, T. Morita, R. Futrell, What do rnn language models learn about filler-gap dependencies?, arXiv preprint arXiv:1809.00042 (2018).

[6] J. Hu, J. Gauthier, P. Qian, E. Wilcox, R. P. Levy, A systematic assessment of syntactic generalization in neural language models, arXiv preprint arXiv:2005.03692 (2020).

[7] E. G. Wilcox, R. Futrell, R. P. Levy, Using computational models to test syntactic learnability, Linguistic Inquiry 55 (2022) 805–848. URL: https://api.semanticscholar.org/CorpusID:247235030.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[9] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: https://aclanthology.org/2020.acl-main.463/. doi:10.18653/v1/2020.acl-main.463.

[10] S. T. Piantadosi, Modern language models refute chomsky's approach to language, From fieldwork to linguistic theory: A tribute to Dan Everett 15 (2023) 353–414.

[11] R. Katzir, Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023), Manuscript. Tel Aviv University. url: https://lingbuzz. net/lingbuzz/007190 (2023).

[12] K. Ethayarajh, D. Jurafsky, Utility is in the eye of the user: A critique of nlp leaderboards, arXiv preprint arXiv:2009.13888 (2020).

[13] J. Coda-Forno, M. Binz, J. X. Wang, E. Schulz, Cogbench: a large language model walks into a psychology lab, arXiv preprint arXiv:2402.18225 (2024).

[14] S. R. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, S. R. Bowman, Blimp: The benchmark of linguistic minimal pairs for english (electronic resources) (2020).

[15] B. Xiang, C. Yang, Y. Li, A. Warstadt, K. Kann, Climp: A benchmark for chinese language model evaluation, arXiv preprint arXiv:2101.11131 (2021).

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, arXiv preprint arXiv:1804.07461 (2018).

[17] J. Steuer, M. Mosbach, D. Klakow, Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures, in: A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, R. Cotterell (Eds.), Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Singapore, 2023, pp. 142–157. URL: https://aclanthology.org/2023.conll-babylm.12/. doi:10.18653/v1/2023.conll-babylm.12.

[18] C. Chesi, V. Bressan, M. Barbini, A. Fusco, M. L. P. Bianchessi, S. Neri, S. Rossi, T. Sgrizzi, Different ways to forget: Linguistic gates in recurrent neural networks, in: M. Y. Hu, A. Mueller, C. Ross, A. Williams, T. Linzen, C. Zhuang, L. Choshen, R. Cotterell, A. Warstadt, E. G. Wilcox (Eds.), The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 106–117. URL: https://aclanthology.org/2024.conll-babylm.9/.

[19] A. Fusco, M. Barbini, M. L. Piccini Bianchessi, V. Bressan, S. Neri, S. Rossi, T. Sgrizzi, C. Chesi, Recurrent networks are (linguistically) better? an (ongoing) experiment on small-LM training on child-directed speech in Italian, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 382–389. URL: https://aclanthology.org/2024.clicit-1.46/.

[20] D. Trotta, R. Guarasci, E. Leonardelli, S. Tonelli, Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Lin-

guistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2929–2940. URL: https://aclanthology.org/2021.findings-emnlp.250/. doi:10.18653/v1/2021.findings-emnlp.250.

[21] D. Brunato, C. Chesi, F. Dell'Orletta, S. Montemagni, G. Venturi, R. Zamparelli, et al., Accompl-it@ evalita2020: Overview of the acceptability & complexity evaluation task for italian, in: CEUR WORKSHOP PROCEEDINGS, CEUR Workshop Proceedings (CEUR-WS. org), 2020.

[22] C. Chesi, G. Ghersi, V. Musella, D. Musola, et al., Conversa: Test di comprensione delle opposizioni morfo-sintattiche verbali attraverso la scrittura (2024).

[23] J. R. Ross, Constraints on variables in syntax. (1967).

[24] N. Lan, E. Chemla, R. Katzir, Large language models and the argument from the poverty of the stimulus, Linguistic Inquiry (2024) 1–28.

[25] L. Choshen, R. Cotterell, M. Y. Hu, T. Linzen, A. Mueller, C. Ross, A. Warstadt, E. Wilcox, A. Williams, C. Zhuang, [call for papers] the 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus, arXiv preprint arXiv:2404.06214 (2024).

[26] I. Alfano, F. Cutugno, A. De Rosa, C. Iacobini, R. Savy, M. Voghera, et al., Volip: a corpus of spoken italian and a virtuous example of reuse of linguistic resources, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 3897–3901.