Automatic Detection of Direct and Self Repetitions in Naturalistic Speech Recordings of French- and Dutch-Speaking Autistic Children

Anonymous ACL submission

Abstract

This study investigates the use of cosine similarity measures across lexical, syntactic, and semantic vectors to detect direct and self-repetitions in the spontaneous speech of autistic children. Using datasets of French and Dutch autistic children's speech, the results show that semantic and lexical similarity provide reliable cues for identifying self-repetitions, achieving high precision and recall scores. However, direct repetitions are more challenging to detect. Overall, the best models for the detection of both types of repetition are based on lexical and semantic similarities. By contrast, models based on syntactic similarity perform worse in all conditions. Further research is needed to refine models for direct repetitions and explore their cross-linguistic applicability.

1 Introduction

Autism is a neurodevelopmental condition with a wide range of symptoms that relate to social communicative impairments and repetitive behaviors (American Psychiatric Association, 2013; Schaeffer et al., 2023).

Echolalia, the repetition of previously heard speech, is often regarded as a core feature of autism due to its prevalence in the language of autistic individuals, with variations depending on language proficiency (Maes et al., 2024). However, definitions of the phenomenon vary widely, and the distinction with 'common' repetition as it occurs in neurotypical language development is not clear cut.

Traditionally, categories of echolalia differ both in their formal resemblance with the source utterance (*pure* vs. *mitigated* echolalia) and in their timing with regard to the source (*direct* vs. *delayed* echolalia, where the latter can also comprise sources from outside the conversation, such as songs). However, the definitions of these categories, and their inclusion under the phenomenon 'echolalia' differ between authors. Similarly, self-repetitions may (McFayden et al., 2022) or may not (van Santen et al., 2013) be considered as echolalia, or as a related 'non-generative' phenomenon (Luyster et al., 2022). Some researchers exclude all repetitions that display communicative intent (e.g., question for clarification) or that do not mimic the prosody of the source (Amiriparian et al., 2018; Marom et al., 2018), while others accept variations to form and function (Pascual et al., 2017; Xie et al., 2023). This lack of consensus complicates systematic analyses, particularly in large language corpora, as definitions often rely on detailed pragmatic and conversational analyses to determine whether an utterance qualifies as echolalia (Ryan et al., 2024).

In this context, some researchers have attempted to develop methods to automatically extract segments of echolalic speech. Some approaches rely on acoustic analysis to examine spectral similarities between sentences (Amiriparian et al., 2018), while others focus on transcription-based analyses to identify repetitions (Bigi et al., 2014; van Santen et al., 2013). From this perspective, Fusaroli et al. (2023) have made significant contributions by reframing the study of echolalia through the lens of alignment theory. Their methodology involves computing alignment rates across multiple linguistic levels — lexical, syntactic, and semantic — between autistic children and their caregivers to quantify the degree of 'recycling' of language material by the children. This approach offers valuable insights into the interactive dynamics of language in autism.

Building on this foundation, our study adapts and extends Fusaroli et al. (2023)'s approach with a novel aim: instead of computing a global alignment or repetition rate, we seek to detect recurring utterances by comparing alignment scores between pairs of utterances, contrasting those classified as repetitive with those classified as non-repetitive. By establishing thresholds for syntactic, lexical, and

semantic similarity on an extensively annotated gold-standard dataset (*cf* 2.1), we enable an efficient and scalable approach for detecting repetitive speech. This approach facilitates a detailed analysis of echolalia, providing insights into its linguistic features, length, and communicative functions. Furthermore, the success of each of the similarity computations for detecting repetitive pairs informs us of the linguistic levels (lexical, syntactic and/or semantic) that lead listeners to the impression of 'sameness' in a source-echolalic pair.

2 Methods

The data used for the development of the models presented are drawn from the **XXX** Study. The sample comprises naturalistic speech recordings from 15 Dutch- and 14 French-speaking children aged 2 to 6 years (mean = 57.5 months, SD = 9.6 months; 19 males, 10 females); the study itself included more Dutch-speaking children, but 15 among them were selected to ensure a comparable sample between languages. All children had a formal autism diagnosis, further confirmed through the second edition of the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al. 2012). The ADOS-2 assessment also demonstrated that all children were verbal, albeit at varying levels. Given that the children were of similar ages, the ADOS-2 modules administered (Module 1 for children with some words, Modules 2 and 3 for those capable of combining words and forming sentences) provided a relative qualitative measure of their verbal abilities.

Within our sample, 1 out of 15 Dutch-speaking and 6 out of 14 French-speaking children were assessed with the first ADOS module, while the second was administered to 5 and 4, and the third 9 and 4, respectively. These differences suggest that Dutch-speaking children in our sample are generally more verbal than their French-speaking peers. However, only their linguistic productions (isolated words, word combinations, sentences) were considered in the construction of our model, excluding pre-linguistic productions (vocalizations, babbling, etc.).

Speech recordings were collected over six hours in the children's homes using a small label recorder placed in the pocket of a project-designed T-shirt. We selected the hour during which each child spoke the most, identified using a pre-trained diarization model (Lavechin et al. 2021). From this selected hour, we make an orthographic transcription of at least 20 minutes of speech per child, adjusting the duration based on their verbal output.

2.1 Gold standard annotation

To establish a gold standard annotation for the repetition detection task, we manually coded direct and self-repetitions in 76 audio samples of 10 minutes each (760 minutes; or 12 hours and 40 minutes). A total of 360 minutes were annotated for the 14 French-speaking children and 400 minutes for the 15 Dutch-speaking children. Coding was performed in Praat (Boersma and Weenink, 2025). Direct repetitions were defined as sentences occurring within a maximum of 10 seconds of the source clause, sharing at least one content word irrespective of morphological changes. Self-repetitions were defined as verbatim repetitions of sentences or single words by the child. For more information about the coding protocol for the gold standard, see Appendix.

2.2 Model Development for Repetition Detection

Since the recordings were obtained without explicit instructions or control over background noise, we opted against using an audio-based model for detecting repetitions. Instead, we developed a model based on the orthographic speech transcriptions by the autistic children and other speakers in the recordings. This approach adapts methodologies proposed by Fusaroli et al. (2023), with modifications to accommodate languages other than English and to include additional interlocutors in the dataset.

Furthermore, in addition to direct repetitions, we also tested self-repetitions with this framework. Thus, we computed cosine similarity on syntactic, lexical, and semantic vectors of each sentence, comparing them with those from 10 seconds prior (for direct repetitions) or with other utterances in the child's speech (for self-repetitions).

2.3 Vector representation, Similarity Measures, and Performance Evaluation

For syntactic vectors, we used SpaCy models (*nl core news sm* for Dutch and *fr core news sm* for French; Honnibal and Montani (2017)) to determine part-of-speech (POS) tags, grouped into n-grams with *n=2*, as per Fusaroli's findings. Due to the large number of short utterances (< 4 words),

we opted against using larger n-grams. If an utterance contained fewer tokens than the selected *n=2*, the entire utterance was treated as a single n-gram. Similarly, we used spaCy to extract lemmas, creating a list of unique lemmas. Then for each file, a list of all unique lemmas and POS n-grams was constructed. Each utterance was then represented as a vector, where each value indicated the number of times (0, 1, 2...) each lemma or POS n-gram from the list appeared in the utterance. This ensured uniform vector structure across speakers, facilitating meaningful comparisons regardless of utterance length. Function words were included, as their proportional presence across utterances minimally affected similarity measures. For semantic vectors, we employed Sentence BERT embedding models trained on French (CamemBERT large, Martin et al.2020) and Dutch (RobBERT, Delobelle et al. 2020). These models generated fixed-length embeddings of 1024 dimensions for French and 768 dimensions for Dutch, aligning with the one-dimensional format supported by the Python SentenceTransformers library (Reimers and Gurevych 2019, 2020).

This multi-level linguistic approach integrates lexical, syntactic, and semantic representations. After constructing vector representations, cosine similarity scores were calculated using the Sentence Transformers *cos sim* function to compare pairs of utterances. After constructing vector representations, cosine similarity scores were calculated using the Sentence Transformers cos sim function to compare pairs of utterances. The autistic child's utterances were compared to (i) all those they had previously produced (self-repetition) and (ii) those of other speakers that occurred at most 10 seconds earlier.

Next, we aimed to determine which cosine similarity thresholds yielded the best results in distinguishing non-repetitive from repetitive utterance pairs. A range of 100 thresholds between -1 and 1 (the range of the cosine similarity function) with a step size of 0.02 was tested for each measure, and the resulting recall and precision values were evaluated. Our goal was to maximize recall (indicating the proportion of repetitions correctly detected) while maintaining precision (indicating the proportion of predicted 'repetitive' cases that were actually repetitive) at an acceptable level (*cf* Table 1). Finally, we evaluated the performance of the selected thresholds for each measure on the test set. Data visualization was conducted using the Python library Plotly (Inc. 2015) and Matplotlib (Hunter 2007). Generative AI tools were used to debug Python code (OpenAI 2025).

3 Results

This section presents the results for both direct and self-repetitions, comparing cosine similarities of lexical, syntactic, and semantic vectors across French and Dutch datasets.

3.1 Overall performance of the model

Figures 2 and 1 illustrate the overall performance of models based on lexical, semantic, and syntactic cosine similarities in distinguishing non-repetitive pairs from direct or self-repetitions. Receiver Operator Curves (ROC) in dashed lines plot the True Positive Rate against the False Positive Rate for the thresholds detecting self-repetitions. By contrast, full lines do so for the thresholds detecting direct repetitions. Overall, the Area Under the Curve (AUC) scores are quite satisfactory for all linguistic measures (above 73%), in both languages and phenomena. However, the ROC are higher for self-repetitions than for direct repetitions across the three measures. Secondly, AUC-scores are markedly lower for thresholds on syntactic similarity (73.2% and 76.2% for Dutch and French direct repetitions; 92.8% and 94.5% for Dutch and French self-repetitions) than for those on lexical and semantic similarity. Indeed, the latter score between 88.6% for direct repetition and 99.9% for self-repetition. Lastly, performances of the thresholds on Dutch data are generally slightly lower than those of models on French data. In sum, the best-performing models are those that detect self-repetitions based on lexical and semantic similarity, achieving an AUC score of more than 99.7% in both languages.

In the following, we will illustrate the observed differences on the basis of the distributions of the different linguistic measures in repetitive vs. non-repetitive utterance pairs in both repetitive phenomena for the two languages. Figure 3 shows the distribution for candidates of self-repetition and figure 4 that of candidates for direct repetition. The thresholds that achieved the best precision-recall combination are indicated as reference lines on the box plots.

The effectiveness of the measure in detecting direct or self-repetitions can be evaluated in multiple ways:

- Ability of the best threshold to "split the plot in two": Nearly all values for repetitive pairs should appear above the threshold, while those for non-repetitive pairs should be below it.

- Similarity in distribution between languages: if overlap at a linguistic level (lexicon, syntax, semantics) is expected to characterize direct echolalia, it should do so consistently across different languages (i.e., Dutch and French in our dataset).

- High recall, precision, and F1 score (harmonic mean of recall and precision) for the chosen threshold: See Table 1.

3.2 Performances of the model detecting self-repetitions

The box plots in Figure 3 illustrate the distribution of similarity measures for self-repetitions versus non-repetitive pairs. As expected, non-repetitive pairs predominantly exhibit low similarity values, whereas repetitive pairs show high values. The thresholds for all measures consistently exceed 0.8, effectively dividing the plots into two distinct areas with relatively few outliers on either side. Moreover, these thresholds remain highly similar across both languages. These observations suggest that self-repetitions are characterized by substantial overlap across all linguistic levels (lexical, syntactic, and semantic).

Nevertheless, differences in distribution are evident across measures. Syntactic similarity plots display greater dispersion in similarity scores, with notably more repetitive outliers in the lower range (0.0–0.6 cosine similarity) and more non-repetitive outliers above the threshold (0.879 or 0.899) compared to lexical and semantic measures. Consequently, the syntactic similarity threshold results in lower precision values, particularly for the Dutch data (French: 61.5%, Dutch: 46.5%) in contrast to precision scores between 86.5% and 87.9% for other measures (*cf.* Table 1). Additionally, cosine similarity scores for non-repetitive utterance pairs are generally more concentrated in the lower range (0–0.2) for Dutch than for French, except for semantic cosine similarity scores.

Recall scores are high for all thresholds, particularly for lexical and semantic similarity, ranging between 84.3% and 89.0%, with the highest values found in lexical and semantic cosine similarities. These results indicate that high lexical and semantic similarity serve as robust cues for distinguishing self-repetitions from non-repetitive utterance pairs by the same speaker.

3.3 Performances of the model detecting direct repetitions

According to Table 1, the best overall results for detecting direct repetitions are achieved using thresholds based on semantic and lexical cosine similarity, yielding recall rates of 73.7% and 75.2% for French and Dutch, respectively. However, the low precision values suggest a high proportion of false positives.

Furthermore, Figure 4 indicates that the conditions observed in the distribution of self-repetitions are not fully replicated for similarity measures applied to direct repetitions. While non-repetitive pairs are largely concentrated in the lower range of the plots, a significant proportion of outliers appear in the upper range, particularly for syntactic similarity. Moreover, the distribution of repetitive pairs deviates from the expected pattern, exhibiting considerable dispersion. Consequently, a substantial number of repetitive pair values fall below the thresholds and are thus not detected as repetitive. Additionally, the threshold values for direct repetitions are markedly lower than those for self-repetitions, indicating a reduced degree of linguistic overlap between utterance pairs.

Lastly, cosine similarity distributions and selected thresholds vary between languages, with consistently lower values for Dutch than for French. This difference is most pronounced in lexical similarity, where the optimal threshold is 0.293 for French and 0.232 for Dutch.

Discussion

Extending the approach proposed by Fusaroli et al. 2023, which computes cosine similarity across lexical, syntactic, and semantic vectors to detect direct and self-repetitions in children's speech, has proven effective, particularly for self-repetitions. While our model successfully detects an acceptable proportion of direct repetitions (recall around 75% or higher) using lexical and semantic similarity measures, a high number of false positives remains (*cf.* lower precision values), largely due to the presence of high outliers in the non-repetitive group. Thus, the model's predictions for direct repetitions should be interpreted with caution. This is-

Phenomenon	Similarity Type	Language	Threshold	Precision	Recall	F1 score
Self-repetition	Lexical	FR	0.879	87.9%	88.8%	88.3%
		DU	0.919	86.5%	89.1%	87.8%
	Syntactic	FR	0.899	61.5%	84.3%	71.1%
		DU	0.879	46.5%	85.0%	60.1%
	Semantic	FR	0.879	87.8%	89.0%	88.4%
		DU	0.879	86.8%	87.8%	87.3%
Direct repetition	Lexical	FR	0.293	59.3%	73.7%	65.7%
		DU	0.232	60.3%	75.2%	66.9%
	Syntactic	FR	0.232	41.2%	58.0%	48.2%
		DU	0.212	39.1%	47.9%	43.0%
	Semantic	FR	0.394	55.9%	76.1%	64.5%
		DU	0.374	52.0%	68.6%	59.2%

Table 1: Results of precision, recall, and F1 scores for the best thresholds across different phenomena, linguistic levels, and languages.

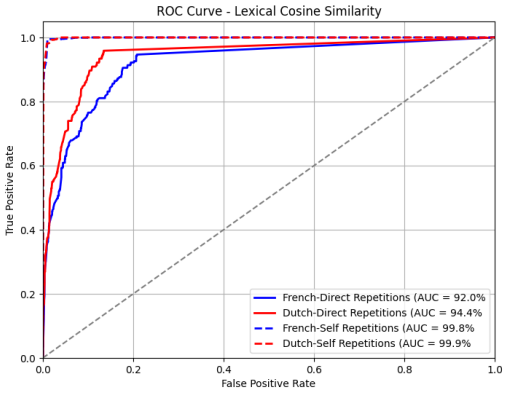
sue may stem from our annotation protocol, which classifies utterances as direct repetitions even when they share only a single content word (e.g., "Do you want a banana?" – "I like bananas"). Since this single word constitutes only a small portion of an utterance's lexical, syntactic, or semantic vector—especially in longer utterances—vector-level comparisons may not be well suited for detecting direct repetition. A simple solution aligned with our annotation protocol could involve a rule-based algorithm that checks for lemma correspondence of content words between utterances.

Moreover, the poor performance of models using syntactic similarity, as evidenced by low precision and recall values, suggests that syntactic structure is highly variable in spontaneous speech. This variability complicates detection without more sophisticated syntactic processing. In contrast, self-repetitions yield strong and consistent results for both lexical and semantic similarity in both languages, with high alignment scores for both French and Dutch data. Semantic similarity appears to be the most reliable cue for detecting both direct and self-repetitions across languages. With recall and precision scores exceeding 86

This study highlights the potential of using machine learning models based on cosine similarity to analyze spontaneous speech in naturalistic settings. Future research could extend this methodology to a broader range of languages and age groups to explore how repetition patterns vary across different linguistic and developmental contexts. A more detailed investigation into the factors influencing performance differences between languages (e.g., linguistic structure and speech patterns) could help refine the models for more accurate repetition detection.

For instance, similarity distributions and thresholds differ between Dutch and French data, with consistently higher values for French in direct repetition comparisons, whereas results for self-repetitions are highly comparable. This pattern may indicate that lexical, semantic, and syntactic overlap between speakers is influenced by language-specific interaction styles (i.e., French-speaking children in our sample may align more closely with their conversational partners than Dutch-speaking children). This could be due to the generally lower verbal output observed among French-speaking children in our dataset (*cf.* Methods section). However, it may also reflect inherent linguistic differences in the 'default' overlap between utterance pairs (see Limitations section).

Future research should evaluate the success of



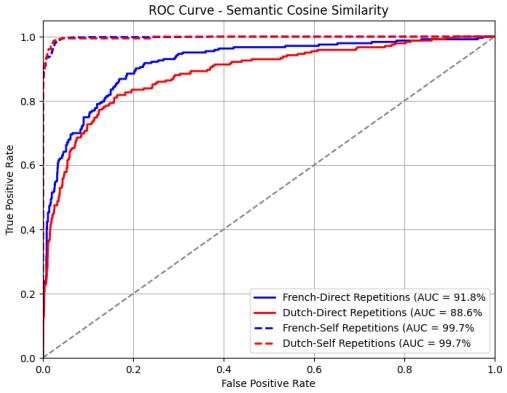


Figure 1: ROC and AUC for lexical and semantic similarity in each phenomenon (direct repetition vs. self-repetition) and language (French vs. Dutch).

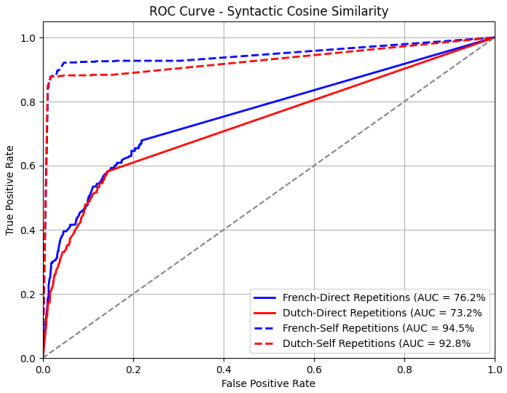


Figure 2: ROC and AUC for syntactic similarity in each phenomenon (direct repetition vs. self-repetition) and language (French vs. Dutch).

our approach in different languages and conversational contexts (e.g., structured oral conversations such as debates). Additionally, the poor performance of syntactic similarity measures suggests that alternative syntactic representation methods—such as more advanced syntactic parsing techniques or deeper contextual analysis—could enhance the detection of syntactic repetitions.

Another possible explanation for the performance differences between French and Dutch lies in variations in the technical capabilities of the NLP algorithms used for each language (spaCy and SentenceBERT models). These algorithms, trained on less extensive datasets than their English counterparts, may introduce biases. Applying our models to English data with corresponding NLP models could provide valuable insights into the impact of algorithmic differences on repetition detection. Furthermore, these models are optimized for written language, whereas our study focuses on spontaneous children's speech, which features informal grammar and vocabulary that standard NLP models are not specifically designed to handle. Future research should compare different models and embeddings to assess their impact on repetition detection.

We encourage interested researchers to test our model on their conversational data while considering its potential limitations. To facilitate this, our model is publicly available at this anonymous repository. Users can select linguistic levels for comparison (lexical, syntactic, semantic) and adjust cosine similarity thresholds. They are not restricted to the thresholds presented in this paper but may experiment with values within an acceptable range.

Finally, a key limitation of this study is the absence of a widely accepted definition of echolalia that allows for purely linguistic detection without requiring extensive conversational or psychological analysis. Our annotation protocol (*cf.* Section 2.1) attempts to address this issue by using simple linguistic criteria (e.g., comparing lemmas, POS, and dependency structures between utterances) designed with potential automation in mind. However, this approach has limitations: for instance, in the case of direct repetition, evaluating similarity at the utterance level instead of individual lemmas led to poorer model performance. Additionally, our model was trained to detect utterance pairs that would not traditionally be classified as echolalic in previous research (e.g., repeated single words used for calling someone). Thus, our models serve as an initial filtering step to identify potential echolalic utterances, which users can then refine based on their specific criteria. However, our approach fails to capture echolalic phrases that do not fit our sim-

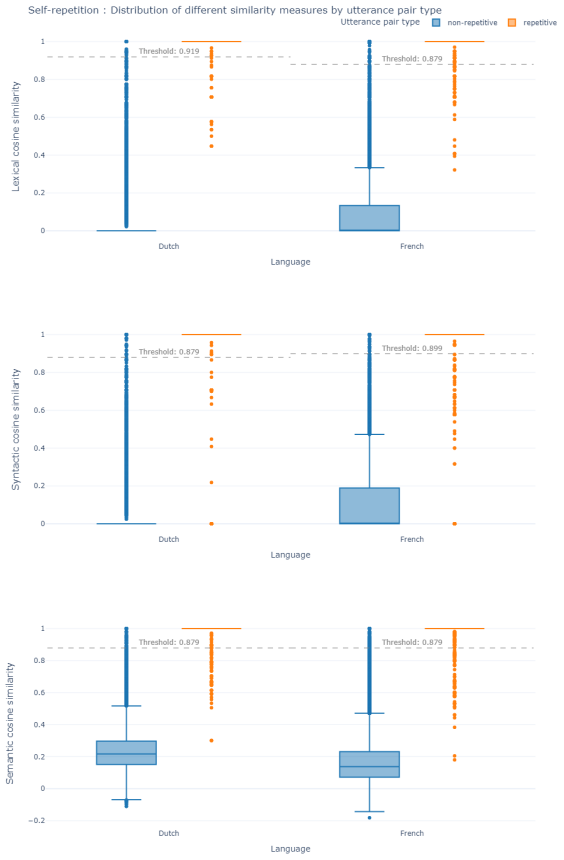


Figure 3: Distribution of lexical, syntactic and semantic cosine similarity measures in self-repetition vs. non-repetitive utterance pairs in the Dutch and French datasets.

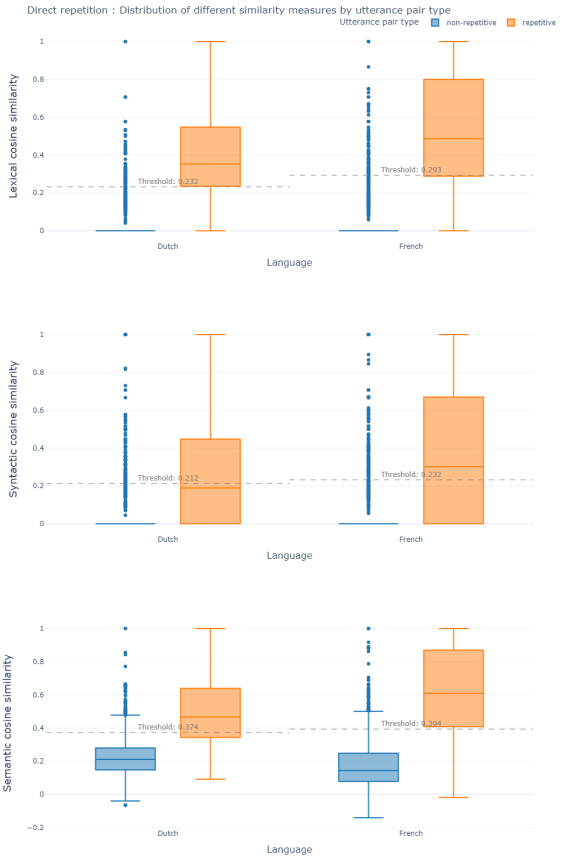


Figure 4: Distribution of lexical, syntactic and semantic cosine similarity measures in direct repetition vs. non-repetitive utterance pairs in the Dutch and French datasets.

plified definition of self-repetitions (e.g., the same word used in different syntactic structures). Establishing more precise definitions and clearer criteria for identifying echolalia would improve repetition detection accuracy in future studies.

While our study demonstrates the effectiveness of cosine similarity-based models for detecting self-repetitions, the challenges in detecting direct repetitions highlight the need for refined methods, such as lemma-based rule systems or adaptive thresholding techniques. The observed differences between French and Dutch suggest that linguistic structure and NLP model limitations influence performance, underscoring the need for further exploration of cross-linguistic generalizability. Future research should also consider testing multilingual and fine-tuned models to enhance repetition detection across languages and spontaneous speech settings.

Limitations

This study has several limitations that should be acknowledged to contextualize its findings and inform future research.

First, a significant limitation lies in the lack of a universally accepted definition of echolalia. To facilitate detection, we employed simplified linguistic criteria designed for potential automation. While effective in some cases, this approach led to the identification of certain segments that do not qualify as true echolalic instances (e.g., single-word vocatives, such as names or calls, repeated during the recording). Conversely, it also failed to capture echolalic phrases that did not align with the adopted definition, such as repetitions involving the same word used in different syntactic structures. The trade-off between simplicity and comprehensiveness highlights the need for more precise definitions of echolalia. Establishing clearer criteria would improve the reliability and validity of automated detection methods, ensuring better

alignment with the nuanced patterns of echolalic speech.

Second, technical challenges associated with pretrained NLP models must be addressed. The tools used in this study, including SBERT and spaCy, exhibited variable performance across the two analyzed languages. These models are typically optimized for formal written text and are not designed to account for the unique characteristics of spontaneous children's speech. As such, they may struggle to process features such as informal grammar, incomplete sentences, or age-specific vocabulary. Developing or fine-tuning NLP models specifically for spontaneous speech data could significantly enhance the accuracy and reliability of repetition detection in this domain. Moreover, the quality of these models varies by language, with NLP algorithms for French and Dutch generally being less robust than their English counterparts due to more limited training data. Future research could benefit from employing more advanced or domain-specific NLP models to mitigate these limitations.

Third, the transcription protocol used in this study introduces additional constraints. Specifically, a new sentence was defined when there was a pause of one second or longer in the child's speech. While necessary for standardization, this approach may have inadvertently excluded pairs of self-repetitions with different syntactic structures simply because they were followed by another sentence. This limitation underscores the need for more flexible transcription criteria that account for the temporal dynamics of naturalistic speech or for a more precise definition of the phrase unit to be considered during comparisons.

Fourth, our analysis revealed potential language-specific variability in repetition patterns and model performance. For instance, thresholds for detecting direct repetitions were consistently higher in French than in Dutch. This variability raises questions about the generalizability of the established thresholds to other languages. Additionally, the lack of validation on independent datasets limits the broader applicability of our models, particularly for detecting direct repetitions. Future studies should test these models across diverse linguistic contexts to refine their utility and generalizability.

Fifth, limitations in the syntactic representations used in this study must also be noted. For syntactic vectors, spaCy was used to extract POS tags, which were grouped into n-grams (n=2). While this approach facilitated uniform vector structures, it introduced potential biases when utterances contained fewer tokens than the selected n, resulting in less informative representations. Additionally, the inclusion of function words may have had minimal influence on similarity measures. Further exploration of alternative vectorization strategies, such as experimenting with different values of n, is warranted to address these concerns.

Despite these limitations, the methodology and findings presented in this study provide a valuable foundation for advancing the automated detection of self-repetitions and direct repetitions. Future research should aim to refine these methods and extend their application to a wider range of languages, age groups, and conversational contexts.

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5 edition. American Psychiatric Publishing, Arlington, VA.

S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrović, E. Ainger, N. Cummins, and B. Schuller. 2018. Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks. In *Proceedings of Interspeech 2018*, pages 2334–2338.

B. Bigi, R. Bertrand, and M. Guardiola. 2014. Automatic detection of other-repetition occurrences: Application to french conversational speech. In *Proceedings of Speech Prosody 2014*.

P. Boersma and D. Weenink. 2025. Praat: doing phonetics by computer [computer program]. Version 6.4.26, retrieved 8 January 2025 from http://www.praat.org/.

P. Delobelle, T. Winters, and B. Berendt. 2020. Robbert: a dutch roberta-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265. Association for Computational Linguistics.

R. Fusaroli, E. Weed, R. Rocca, D. Fein, and L. Naigles. 2023. Repeat after me? both children with and without autism commonly align their language with that of their caregivers. *Cognitive Science*, 47(11):e13369.

M. Honnibal and I. Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Plotly Technologies Inc. 2015. Collaborative data science.

M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia. 2021. An open-source voice type classifier for child-centered daylong recordings. *Preprint*, arXiv:2005.12656.

C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. Bishop. 2012. *Autism diagnostic observation schedule, second edition (ADOS-2)*. Western Psychological Services.

R. J. Luyster, E. Zane, and L. Wisman Weil. 2022. Conventions for unconventional language: Revisiting a framework for spoken language features in autism. *Autism & Developmental Language Impairments*, 7:23969415221105472.

P. Maes, C. La Valle, and H. Tager-Flusberg. 2024. Frequency and characteristics of echoes and self-repetitions in minimally verbal and verbally fluent autistic individuals. *Autism & Developmental Language Impairments*, 9:23969415241262207.

M. K. Marom, A. Gilboa, and E. Bodner. 2018. Musical features and interactional functions of echolalia in children with autism within the music therapy dyad. *Nordic Journal of Music Therapy*, 27(3):175–196.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. 2020. Camembert: A tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

T. C. McFayden, S. M. Kennison, and J. M. Bowers. 2022. Echolalia from a transdiagnostic perspective. *Autism & Developmental Language Impairments*, 7:23969415221140464.

OpenAI. 2025. Chatgpt.

E. Pascual, A. Dornelas, and T. Oakley. 2017. When "goal!"means 'soccer': Verbatim fictive speech as communicative strategy by children with autism and two control groups. *Pragmatics & Cognition*, 24(3):315–345.

B. M. Prizant. 1983. Language acquisition and communicative behavior in autism: Toward an understanding of the 'whole' of it. *The Journal of Speech and Hearing Disorders*, 48(3):296–307.

N. Reimers and I. 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.

N. Reimers and I. Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S. Ryan, J. Roberts, and W. Beamish. 2024. Echolalia in autism: A scoping review. *International Journal of Disability, Development and Education*, 71(5):831–846.

J. Schaeffer, M. Abd El-Raziq, E. Castroviejo, S. Durrleman, S. Ferré, I. Grama, P. Hendriks, M. Kissine, M. Manenti, T. Marinis, N. Meir, R. Novogrodsky, A. Perovic, F. Panzeri, S. Silleresi, N. Sukenik, A. Vicente, R. Zebib, P. Prévost, and L. Tuller. 2023. Language in autism: Domains, profiles and co-occurring conditions. *Journal of Neural Transmission*, 130(3):433–457.

L. Sterponi and J. Shankey. 2014. Rethinking echolalia: Repetition as interactional resource in the communication of a child with autism. *Journal of Child Language*, 41(2):275–304.

J. P. H. van Santen, R. W. Sproat, and A. P. Hill. 2013. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383.

F. Xie, E. Pascual, and T. Oakley. 2023. Functional echolalia in autism speech: Verbal formulae and repeated prior utterances as communicative and cognitive strategies. *Frontiers in Psychology*, 14.

Appendix

In this Appendix, we specify the annotation protocol that was used to annotate direct repetitions and self-repetitions in the speech of autistic children. Echolalia, or the repetition of previously heard speech, is a phenomenon commonly associated with the language of autistic children, but even within autism studies, definitions and categories within echolalia vary widely. We refer to the paper (Introduction section) for a summary of the debate and the most frequently mentioned categories of echolalia. The goal of our research was to develop a computational model that can detect immediate and self-repetitions in transcriptions of conversations, based on linguistic (lexical, syntactic, semantic) similarities between spoken utterances. Bearing in mind the limitations of existing NLP methods and the challenges that are inherent to detecting echolalia (e.g., deciding whether an utterance is novel or recycled from previous conversations on the basis of limited contextual information), we decided to create a model that could detect candidate utterances for echolalia, i.e., repetitions from utterances previously spoken by interlocutors or the child itself. In other words, our models were designed as a first 'filtering step' to determine possible echolalic utterances, where the user can filter

out examples that do not correspond to their definition of the phenomenon using detailed conversational and pragmatic analysis. As a consequence, our annotation protocol to develop the ground truth data is specifically designed with possible automatization in mind: it uses simple linguistic criteria (comparing lemmas, POS, and dependency structures between utterances) that may be replicated by NLP methods. We will therefore also refer to the phenomena we are describing as 'direct repetitions' and 'self-repetitions' (not echolalia) to ensure a correct interpretation of the models' results by end users.

Direct repetitions

In our effort to develop an algorithm capable of capturing the widest possible range of candidate utterances of echolalia, we adopt a broad definition of direct repetition. Specifically, we define it as *an utterance that includes the repetition of at least one content word (verb, noun, adverb, or adjective) from a prior utterance spoken by an interlocutor, provided that the onset of the preceding utterance occurs no more than 10 seconds before the onset of the utterance under consideration*. This definition identifies repetitive utterances following two main criteria: (i) the number of identical words in the source and the repetition, and (ii) the distance between the source and the repetition. For the first criterion, we decided to consider repetitions of at least one content word as examples of direct repetition. This is because the literature does not provide a clear guidance about the number or the proportion of words that should be repeated between the source utterance and the echolalic utterance. We therefore chose to set a low threshold to capture as many candidates for echolalia as possible. Moreover, this definition resembles that of (Bigi et al., 2014) for other-repetition: the authors consider a pair of word sequences pronounced by two speakers as a source-repetition pair when at least one relevant word is repeated (i.e., the probability that the word occurs in the speech of the original speaker in the dialogue is smaller than a given threshold), or when the source has a predetermined number of words that are repeated exactly. Since we think that direct repetitions may also concern words that are salient in the conversational context, and thus appear multiple times in the speech of the participants (i.e., non-relevant), we decided not to include this definition of 'relevant' words in our criteria. However, we approximate the criterion of relevance by considering only content words, and not function words (conjunctions, pronouns, prepositions, determiners, auxiliaries and interjections (huh)). An example of a repetition of only one word characterized as immediate echolalia is illustrated in (1). Here, the autistic child only repeats the noun eten ('food') from the adult's previous utterance. On the contrary, in (2), the autistic child repeats the other child's utterance word by word: this is an example of exact echolalia. For determining whether an utterance was echolalic, we did not consider any morphological changes to the words, following the approach of (Bigi et al., 2014) for other-repetition and of (Fusaroli et al., 2023) for lexical alignment. Thus, we consider an utterance as repetitive if the lemma (the unconjugated and uninflected form) of at least one content word is identical to the ones in the source utterance. This is the case in (3), where the mother produces the verb koken ('to cook/ to boil') in the infinitive form, and the autistic child uses the first-person (present) form of the same verb (ik kook: 'I cook').

The second criterion for determining whether the child's utterance was an direct repetition of another speaker's previous utterance was the temporal distance between the utterances. We decided to limit the candidates for source utterances to those starting within 10 seconds before the start of the child's utterance. In that way, we approximate the general definition of immediate echolalia as occurring within two conversational turns (Marom et al. 2018, McFayden et al. 2022; (Prizant, 1983) 1983; Sterponi and Shankey 2014; van Santen et al. 2013; Xie et al. 2023), while accounting for the fact that in our data, a source utterance may first be answered by another speaker before the autistic child produces an utterance, or that the child themselves may first produce another utterance before (partially) repeating the source utterance. This is the case in (4), where the autistic child first refers to the red car mentioned by the other child using an anaphorical pronoun (l': 'it') before repeating the other speaker's reference and adding an extra adjective to it (la (dernière) voiture rouge: 'the (last) red car') in a second utterance.

In our annotation protocol, we also take into account that one source utterance can correspond to multiple repetitive utterances. This is shown in (5), where the autistic child reproduces the word eten ('food') in two different utterances.

Conversely, we also account for the occurrence of multiple source candidates for one repeated utterance. If within a distance of 10 seconds from the

start of the utterance multiple utterances are found that the child repeats (partially), then they were all annotated as source utterances. In (6) for example, all three utterances transcribed below occur 10 seconds before the autistic child's utterance and contain the word château ('castle') that the child repeats.

The last guideline goes against the guidelines of (Bigi et al., 2014) for detecting other-repetition: their algorithm only keeps the source with the longest repetition, and then the nearest to the source. However, we aim to detect as many source-repetition pairs as possible, so that the human expert can afterwards decide which of them are echolalic and which are not.

Self-repetitions

In the second place, we annotated self-repetitions in the speech of the autistic child. We aim to identify self-repetitions because it has been hypothesized in the literature that when an autistic child produces delayed echolalia of which the source does not come from an utterance inside the conversation (e.g., utterances from movies and songs), they mostly repeat the utterance in question several times within a short time span (Marom et al. 2018; Sterponi and Shankey 2014). We furthermore hypothesize that these repetitions of the source utterance should be (almost) identical to each other: we presume that this type of delayed echolalia mirrors the source utterance as well as possible, so that it can be recognized by the conversational partners. We approximate this intuition by imposing that for verb phrases, the dependency structure of both utterances (subject, verb, objects) must be identical so that the basic lexical-semantic representation is the same; optional elements such as discourse markers and adjuncts may be added or deleted. For other types of phrases that contain only one major constituent (e.g., noun phrases), we consider that the two phrases must be exactly identical, in correspondence to our first criterion, where we do not allow words to be substituted, added or deleted inside constituents either. We thus apply a stricter definition to self-repetition than to immediate echolalia. We define self-repetitions as *the repetition by the autistic child of an utterance previously pronounced by themselves in the same conversation, containing the same verb and dependency structure (subject and objects). Alternatively, if the utterance is not a verbal phrase, but for example a noun phrase, the repetition needs to be exact, i.e., all (non-filler) words need to be identical.*

In (7) below, the autistic child repeats the subject (tu: 'you'), the verb (as foutu: 'have done'), and the direct object (qu': 'what') from a previous utterance. Thus, he repeats the entire dependency structure of the verb foutre. He does not repeat the adjunct avec ta voiture that is not commanded by the verb, nor the discourse marker mais ('but') or the vocative mec ('dude'). Following our definition, (7) is an example of a self-repetition: the verb and its dependency structure are identical for both utterances, although optional elements are not.

On the contrary, we do not consider (8) a self-repetition, because the subject (t' 'you' vs. il 'he') is different. Similarly, (9) is also not a self-repetition because the direct objects are slightly differently formulated (une voiture de police 'a police car', la police 'the police').

Multiple main verbs can be present in the source utterance and/or the candidate for repetition, for example when the utterance contains a coordinated or subordinated clause. We consider these utterance pairs a self-repetition if at least one of the dependency structures is identical. This is the case in (10), where the autistic child adds a subordinated clause after the repetition of the dependency structure c'est des pies ('those are pies').

As is stated in the definition, when the utterance does not contain a verb, all words of the source and possible repetitive utterance, both content and function words, need to be identical (not considering fillers like uhm). For example, in (11) the noun phrase une voiture ('a car') is a self-repetition of the previous utterance une voiture. Repetitions of discourse markers like yes, no, okay (12) and of vocatives like mom? (13) are also considered self-repetitions: although they most likely do not reflect delayed echolalia, it is important that the automatic detection models detect all one-word repetitions.

Unlike for direct repetitions, we do not impose a criterion for the time distance between the source and the repetition: as the identified self-repetitions may be occurrences of delayed echolalia from a source outside the conversation, they can in principle occur at any moment in the conversation. As for direct repetitions, we considered that one utterance could be repeated several times: in that case, the first occurrence was annotated as 'original' and all the other utterances as 'self-repetitions'. In this situation, an utterance indicated as 'self-repetition' was also considered the source of repetitions occurring afterwards. For example, in (14), the first utterance is considered a source for the second

and third, and the second utterance is considered a source for the third utterance.

Finally, one utterance can be implicated in both a direct repetition and a self-repetition. This is the case in (15): first, the autistic child utters il brille ('it shines'). Then, the other child repeats il brille in two different utterances. Lastly, the autistic child repeats il brille. Hence, the two utterances of the autistic child are considered a self-repetition, and the last pronounced utterance is also considered a direct repetition of the two utterances of the other child.

It is important to note that this last utterance would not be considered echolalic by most authors (e.g., van Santen et al. 2013), because the autistic child was the first to pronounce the repeated words. However, we want our automatic algorithms to detect as many potential occurrences of direct repetitions as possible: indeed, the models will only compare (i) the autistic child's utterance with other speakers' utterances 10 seconds before the start of the utterance for direct repetition, and (ii) different utterances of the autistic child for self-repetition. This entails that any utterance pair that fulfills the previously established definition of direct repetition should be considered repetitive for the training and evaluation of the models. The model for the detection of direct repetitions does not have access to the fact that the other child's utterances are a repetition of an utterance by the autistic child: if we wanted to provide the model with this information, then the second model would rely on the first model's predictions to make its own predictions, which is, of course, an unwanted situation.

1. **Adult**: en is het **eten** of is het speelgoed of wat verkoop je in je winkel?

   and is it food or is it toys or what do you sell in your shop?

   AC (Autistic Child): de jongen toch natuur **eten**

   the boy whatsoever nature food [/ eating]

2. **OC** (Other Child): **c'est ma voiture rouge**

   it's my red car

   AC: **c'est ma voiture rouge**

   it's my red car

3. **Adult**: **koken** ook?

   cooking as well?

   AC: ja in mijn winkel **kook** ik alleen maar dingen

   yes in my shop I only cook things

4. **OC**: non **la voiture rouge** !

   no the red car!

   OC: je l'ai pris [unintelligible word]

   I have taken it

   AC: j'ai pris encore **la** dernière **voiture rouge**

   I have still taken the last red car

5. **Adult**: en is het **eten** of is het speelgoed of wat verkoop je in je winkel?

   and is it food or is it toys or what do you sell in your shop?

   AC: de jongen toch natuur **eten**

   the boy whatsoever nature food [eating]

   AC: **eten**

   food [eating]

6. **OC**: **château** cha- **château** gonfab [gonflable]

   castel ca- bouncing house ['bouncing castle' in French]

   OC: au revoir **château** gonflable *bye bye bouncing castle*

   Adult: il y a pas de château gonflable [unintelligible word]

   there is no bouncing castle

   AC: c'est un **château** en bois

   it's a wooden castle

7. **AC**: mais mec **qu'est-ce que t'as foutu** avec ta voiture ?

   but dude what have you done with your car?

   AC: **qu'est-ce que t'as foutu** ?

   what have you done?

8. **AC**: t'es où Flash McQueen ?

   where are you, Lightning McQueen?

   AC: il est où Flash McQueen ?

   where is he, Lightning McQueen?

9. **AC**: c'est une voiture de police

   it's a police car

   AC: c'est la police

   it's the police

10. **AC**: **c'est des pies**
    those are magpies
    AC: **c'est des pies** parce que [. . .]
    those are magpies because [. . .]
11. **AC**: **une voiture**
    a car
    AC: **une voiture**
    a car
12. **AC**: **ja**
    yes
    AC: **ja**
    yes
13. **AC**: **maman** ?
    mom? [. . .]
    AC: **maman** ?
    mom?
14. **AC**: **c'est des pies** les oiseaux
    those are magpies, the birds
    AC: **c'est des pies**
    those are magpies
    AC: c'est des pies parce que [. . .]
    those are magpies because [. . .]
15. **AC**: et le soleil **il brille**
    and the sun it shines
    OC: **il brille**
    it shines
    AC: oui
    yes
    OC: pourquoi **il brille** ?
    why does it shine ?
    AC : bah c'est **il brille** pour faire de la lu-
    well it's it shines to make li-