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

Federica Beccaria[1,2], **Marie Kolenberg**[2], **Pierre Labendzki**[7], **BeLAS Consortium**, **Inge Zink**[2,3], **Mikhail Kissine**[1,4,5]

[1]Autism in Context: Theory and Experiment (ACTE), Center for Linguistic Research (LaDisco), ULB Neuroscience Institute, Université Libre de Bruxelles

[2]Experimental Oto-Rhino-Laryngology (ExpORL), KU Leuven

[3]Leuven Autism Research (LAuRes), KU Leuven

[4]Department of Linguistics, University College London

[5]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo

[7]University of East London

federica.beccaria@ulb.be, marie.kolenberg@kuleuven.be, labendzki@uel.ac.uk, inge.zink@kuleuven.be, m.kissine@ucl.ac.uk

Abstract

This study investigates the use of cosine similarity measures across syntactic, lexical, and semantic vector representations to detect repetitions in the spontaneous speech of autistic children. It focuses on direct repetitions (i.e., immediate verbatim repetitions of linguistic output produced by another individual) and self-repetitions (i.e., within-speaker recurrence). The performance of similarity-based methods is then compared with state-of-the-art black-box classification models based on BERT, trained on the same data. Using spontaneous speech data from French- and Dutch- speaking autistic children, the results show that lexical and semantic similarity provide reliable cues for identifying self-repetitions, achieving high precision and recall, with F1-scores exceeding 83%, comparable to those obtained by BERT-based models. In contrast, direct repetitions are more difficult to detect using similarity-based approaches, with BERT models clearly outperforming them and reaching F1-scores above 73%. Across all conditions, syntactic similarity consistently underperforms relative to lexical and semantic measures. These findings highlight the strengths and limitations of similarity-based approaches and suggest directions for future research, particularly in improving the detection of direct repetitions and assessing the cross-linguistic generalizability of these methods.

Keywords: Autism, Direct Repetitions, Self-Repetitions, Echolalia, Cosine Similarity, BERT, Repetition Detection

1. Introduction

Autism is a neurodevelopmental condition characterized by a wide range of developmental features, including differences in social communication and repetitive behavior patterns (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 variation depending on language proficiency (Maes et al., 2024). However, definitions of the phenomenon vary widely, and the distinction between echolalia and repetitions observed in neurotypical language development is not clearly delineated. Traditionally, categories of echolalia differ both in their formal resemblance to the source segment (*pure* vs. *mitigated* echolalia) and in their timing relative to the source (*direct* vs. *delayed* echolalia, where the latter may also include sources from outside the conversation, such as songs). However, the definitions of these categories and their inclusion under the phenomenon of echolalia differ between authors. Similarly, self-repetitions have been considered (McFayden et al., 2022), or explicitly excluded (van Santen et al., 2013), as instances of echolalia, or rather as a related non-generative phenomenon, broadly defined as the reuse of previously produced or perceived linguistic material (Luyster et al., 2022). Some definitions exclude all repetitions that display communicative intent (e.g., questions for clarification) or that do not mimic the prosody of the source (Amiriparian et al., 2018; Marom et al., 2018), while others accept formal and functional variation (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 a linguistic segment qualifies as echolalia (Ryan et al., 2024).

Proceedings of the 6th RaPID@MENTAl.ai Workshop @ LREC 2026, pages 146–156
12 May 2026.

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 linguistic representations of different types (syntactic, lexical, and semantic) between autistic children and their caregivers to quantify the degree of recycling language material. 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 linguistic units by comparing pairs of segments, 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, 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 similarity computation in detecting repetitive pairs informs us of the linguistic information (syntactic, lexical, and semantic) that leads listeners to perceive sameness in a source-echolalic pair. In a next step, we compare the results of these linguistically informed methods with those of state-of-the-art pretrained BERT models (Devlin et al., 2019), fine-tuned on our classification task.

2. Methods

The data used for the development of the models presented in this study were drawn from the Belgian Language in Autism Study (BeLAS). The sample comprises naturalistic speech recordings from 14 French- and 15 Dutch-speaking children aged between 2 and 6 years (mean = 55.81 months, SD = 10.66 months; 19 males, 10 females). All children had a formal autism diagnosis and were administered the Autism Diagnostic Observation Schedule - Second Edition (ADOS-2; Lord et al., 2012). Children were additionally assessed using standardized instruments to evaluate expressive and receptive language development quotients (Bayley Scales of Infant and Toddler Development, Clinical Evaluation of Language Fundamentals - Preschool, Evalo, Peabody Picture Vocabulary Test; Bayley, 2006; Semel et al., 2020; Ortho Édition, 2009; Schlichting, 2005; Dunn and Dunn, 2019) as well as non-verbal cognitive skills (Snijders-Oomen Non-verbal Intelligence Test; Tellegen and Laros, 2017). As shown in Table 1, no statistically significant differences were observed between French- and Dutch-speaking children on any of the reported measures, indicating that the two language groups were well matched. Importantly, however, participants exhibited substantial variability in both linguistic and non-verbal cognitive skills, allowing the sample to represent a broad range of developmental profiles within the autism spectrum. Speech recordings were collected over approximately six hours in the children's homes using a small lapel recorder placed in the pocket of a project-designed T-shirt. Then, the hour with the highest amount of each child's speech was selected using a pre-trained diarization model (Lavechin et al., 2021). Finally, we orthographically transcribed at least 20 minutes of child speech, with duration adjusted according to individual language 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 a total of 760 minutes of audio recordings. Each participant contributed at least 20 minutes of audio, with some providing up to 60 minutes depending on the amount of language produced. Of the total, 360 minutes were annotated for 14 French-speaking children and 400 minutes for 15 Dutch-speaking children. Coding was performed using Praat (Boersma and Weenink, 2025).

Direct repetitions were defined as linguistic units occurring within a maximum of 10 seconds of the source clause, sharing at least one content word irrespective of morphological changes. In example 1, produced by a Dutch-speaking child in our corpus, *tickle* is shared between the utterance of another speaker and the autistic child, appearing in two different morphological forms: the first in the third person singular and the second in the first person singular of the present tense.

(1) **Other Speaker:** *Kietelt* dat?
'Does that tickle?'
Autistic Child: *Kietel*
'Tickle'

Self-repetitions were defined as exact reiterations of segments from the child's own language productions. Among the three examples reported below, all produced by the same autistic child in our French-speaking sample, examples 2a and 2b

Measure	French (n = 14)	Dutch (n = 15)	t-value (df)
Age (months)	57.67 (8.45; 42.41–71.06)	55.81 (10.66; 40.01–71.12)	-0.52 (26.34)
ADOS CSS	6.14 (2.38; 2–10)	5.00 (1.60; 2–7)	-1.51 (22.58)
Non-verbal IQ	90.23 (20.82; 55–117)	92.67 (14.79; 57–115)	0.35 (21.30)
Expressive Language	77.27 (32.98; 26.80–117.53)	89.14 (19.53; 57.08–116.41)	0.88 (8.03)
Receptive Language	80.07 (27.11; 32.55–132.01)	89.19 (19.71; 56.68–116.89)	0.95 (17.39)

Table 1: Descriptive statistics of participants by language group (French vs. Dutch), including mean, standard deviation, and range. Independent-samples t-tests indicate no significant group differences

are considered self-repetitions, but 2a and 2c are not, since not all the material of 2a is repeated.

(2) a. **Autistic Child:** *c'est une voiture de police*
'It's a police car'
b. **Autistic Child:** *c'est une voiture de police*
'It's a police car'
c. **Autistic Child:** *c'est* la *police*
'It's the police'

For more information about the coding protocol for the gold standard, see this OSF repository. To assess the reliability of manual coding, 10% of all transcribed audio files were double-coded. Coders demonstrated very high agreement: for direct repetitions, agreement was 95.5% (Cohen's Kappa = 0.84), whereas for self-repetitions, agreement reached 99.4% (Kappa = 0.82), reflecting almost perfect consistency between coders.

2.2. Model Development for Repetition Detection

Since the recordings were obtained without explicit instructions or control over background noise, we opted against an audio-based approach for repetition detection. Instead, we developed a model based on orthographic transcriptions of speech produced by autistic children and other speakers, building on the methodology of Fusaroli et al. (2023) with adaptations for multiple languages and interlocutors. This framework was applied to both direct and self-repetitions.

A linguistic unit is any child's language production, regardless of syntactic complexity, ranging from single words to complete sentences. Non-linguistic vocalizations, such as babbling, were excluded. This inclusive definition allows representation of the full range of linguistic output, including children with limited verbal skills.

Following Fusaroli et al. (2023), each unit was represented using syntactic, lexical, and semantic vectors. Syntactic vectors encode sequences of Part-Of-Speech (POS) tags (e.g., nouns, verbs), lexical vectors encode sequences of lemmas (e.g., *child* for *children*; *write* for *wrote*), and semantic vectors provide a holistic representation of the unit's meaning. Together, these vectors capture the main linguistic information available from transcribed speech.

Cosine similarity was computed for each unit: direct repetitions with other speakers' units within 10 seconds, self-repetitions with the child's earlier productions. These pairs, together with gold-standard annotations, were then used to fine-tune French- and Dutch-language BERT models for repetition detection.

2.3. Cosine Similarity Models: Vector representation, Similarity Measures, and Performance Evaluation

For syntactic vectors, we used spaCy models *fr core news sm* for French and (*nl core news sm* for Dutch; Honnibal and Montani, 2017) to determine POS tags, grouped into n-grams with *n=2*, as per Fusaroli and colleagues' (2023) findings. Due to the large number of short linguistic segments (< 4 words), we opted against using larger n-grams. If a linguistic unit contained fewer tokens than the selected *n=2*, the entire one-word segment was treated as a single n-gram.

Similarly, we used spaCy's module *Lemmatizer* to create a list of unique lemmas. Then, for each file, we constructed a combined list containing all the unique lemmas and POS n-gram sequences. All linguistic units were then represented as single vectors, where each value indicated the number of times (0, 1, 2, etc.) each lemma or POS n-gram from the list appeared in the linguistic segment. This ensured uniform vector structure across speakers, facilitating meaningful comparisons regardless of the linguistic segment's length. Function words were included, as their proportional presence across the considered segments affected similarity measures minimally.

For semantic vectors, we employed BERT SentenceTransformers models trained on French (*CamemBERT* large, Martin et al.2020) and Dutch

(*RobBERT*, Delobelle et al. 2020). These models generated fixed-length embedding of 1024 dimensions for French and 768 for Dutch, aligning with the one-dimensional format supported by the Python SentenceTransformers library (Reimers and Gurevych 2019, 2020). Each dimension corresponds to a numerical feature encoding some aspect of textual meaning; the full set of dimensions jointly represents the text.

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

Next, we aimed to determine which *cosine similarity thresholds* yielded the best results in distinguishing non-repetitive from repetitive production pairs. A range of 100 thresholds between -1 and 1 (corresponding to the cosine similarity function values) with a step size of 0.02 was tested for each measure, and the resulting precision and recall values were evaluated. Our goal was to maximize recall (i.e., the proportion of repetitions correctly detected) while maintaining precision (i.e., the proportion of predicted repetitive cases that were actually repetitive) at an acceptable level (Table 2). Finally, we evaluated the performance (precision, recall, F1-score) of the selected thresholds for each measure.

2.4. BERT Models: Train-Test Split, and Performance Evaluation

In this study, we further compared the performance of our repetition-detection approach with state-of-the-art BERT models. To this end, we fine-tuned BERT models for sequence classification, using the same base models as the SentenceTransformer models employed for constructing semantic vectors (*CamemBERT* large, Martin et al.2020 for French and *RobBERT*, Delobelle et al. 2020 for Dutch). Fine-tuning and evaluation involved creating training and test sets, with a speaker-based split so that the models were evaluated on speech from children not encountered during training. For a direct comparison with BERT, the cosine similarity models were evaluated on BERT's test set here, rather than on the entire dataset as previously.

We isolated 4 French and 4 Dutch speakers (out of the total of 29) in the test set, based on a 4-means clustering of their characteristics (age, expressive and receptive language development quotients) and the number of direct and self-repetitions that they produced. This ensured that similar language profiles of speakers were found in the train and test sets, and that the repetitive phenomena of interest occurred in a similar frequency in both.

The French and Dutch BERT models were then trained on the gold-standard annotations of the remaining 21 speakers for both direct and self-repetitions. Training was conducted using 10-fold cross-validation, with adjustments for unbalanced class weights (the repetitive class being underrepresented), and parameter evaluation based on the F1-score, with 'repetitive' as the positive class. Both BERT models were evaluated on the test set using precision, recall, and F1-score.

2.5. Materials

All code for computing similarity metrics, determining best similarity thresholds, making the train-test split, training the BERT models, evaluating the models, and creating visualizations is available in this OSF repository. The repository also contains details on the characteristics of the train and test sets and the parameters used for training the BERT models.

Moreover, this GitHub contains the finished models as well as Python code that allows users to try the models on their own data.

Data visualization was conducted using the Python libraries Seaborn (Waskom 2021) 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 syntactic, lexical, and semantic vectors across the French and Dutch datasets.

An additional analysis was conducted to compare these models with BERT-based models.

3.1. Performance of the Cosine Similarity Models

Figure 1 illustrates the overall performance of models based on syntactic, lexical, and semantic cosine similarities in distinguishing non-repetitive pairs from direct or self-repetitions. Receiver Operator Curves (ROC) in full lines plot the true positive against the false positive rate for the thresholds detecting direct repetitions. By contrast, dashed lines do so for the thresholds identifying self-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 ROCs 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 French and Dutch direct repetitions; 92.8% and 94.5% for Dutch and French self-repetitions) than for those on lexical and semantic similarity. Indeed, the latter scores 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 AUC scores of more than 99.7% in both languages.

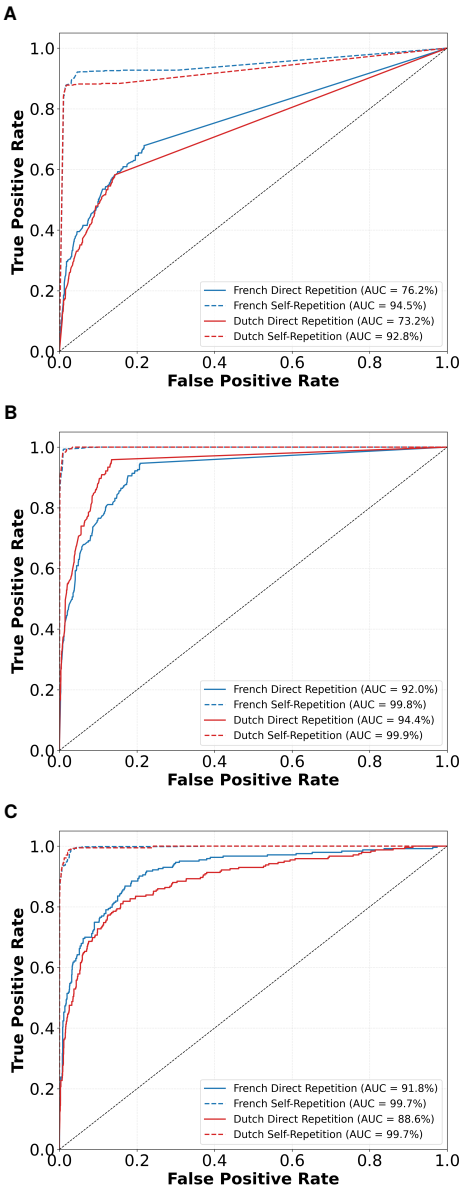

Figure 1: ROC curves and corresponding AUC values for syntactic (A), lexical (B), and semantic (C) cosine similarity models, comparing French and Dutch datasets across direct and self-repetition types

In the following, we will illustrate the observed differences in the distributions of the linguistic measures for repetitive vs. non-repetitive segment pairs in both phenomena in the two languages. Figure 2 shows the distribution for candidates for direct repetition, and Figure 3 for self-repetition. The thresholds that achieved the best precision-recall combination are indicated as reference lines on the box plots.

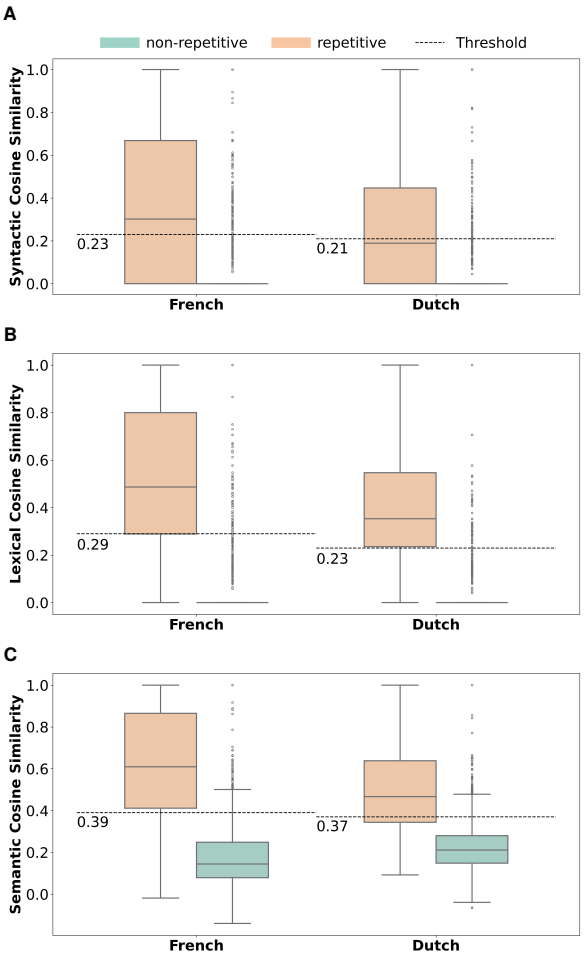

Figure 2: Distributions of syntactic (A), lexical (B), and semantic (C) cosine similarity measures for direct repetition versus non-repetitive segment pairs in the French and Dutch datasets

3.1.1. Performance of the Models Detecting Direct Repetitions

According to Table 2, the best overall results for detecting direct repetitions are achieved using thresholds based on lexical and semantic cosine similarity, yielding recall rates of 76.1% and 75.2% for French and Dutch, respectively. However, the low precision values suggest a high proportion of false positives.

Furthermore, Figure 2 shows that the distribution of the similarity values does not show the ex-

Phenomenon	Model	Language	Threshold	Precision	Recall	F1-score
Direct repetition	Syntactic CosSim	FR	0.23	41.2%	58.0%	48.2%
		DU	0.21	39.1%	47.9%	43.0%
	Lexical CosSim	FR	0.29	59.3%	73.7%	65.7%
		DU	0.23	60.3%	75.2%	66.9%
	Semantic CosSim	FR	0.39	55.9%	76.1%	64.5%
		DU	0.37	52.0%	68.6%	59.2%
Self-repetition	Syntactic CosSim	FR	0.90	61.5%	84.3%	71.1%
		DU	0.88	46.5%	85.0%	60.1%
	Lexical CosSim	FR	0.88	87.9%	88.8%	88.3%
		DU	0.92	86.5%	89.1%	87.8%
	Semantic CosSim	FR	0.88	87.8%	89.0%	88.4%
		DU	0.88	86.8%	87.8%	87.3%

Table 2: Evaluation of repetition detection across phenomena, models, and languages, reporting precision, recall, and F1-score, alongside the optimal Cosine Similarity thresholds. Results are computed over the full dataset

pected pattern (i.e., non-repetitive pairs concentrated in the lower part and repetitive pairs in the higher part of the plot). 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 exhibits 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 segment 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 0.29 for French and 0.23 for Dutch.

3.1.2. Performance of the Models 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 (syntactic, lexical, 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 (0.88 or 0.90) compared to lexical and semantic measures. Consequently, the syntactic similarity threshold results in overall 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 (Table 2). Additionally, cosine similarity scores for non-repetitive segment 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 89.1%, with the highest values 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 segment pairs by the same speaker.

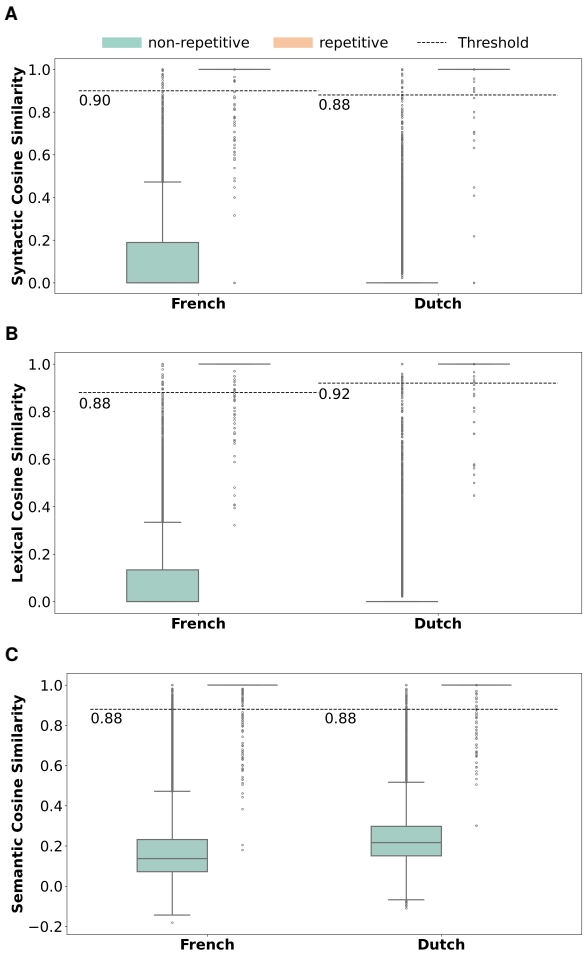

Figure 3: Distributions of syntactic (A), lexical (B), and semantic (C) cosine similarity measures for self-repetition versus non-repetitive segment pairs in the French and Dutch datasets

3.2. Comparison of the Performance of Cosine Similarity and BERT Models on Test Set

While the previous sections reported the results of cosine similarity models applied to the entire dataset, here we compare their performance with BERT models on BERT's test set (8 of 29 speakers).

Overall, BERT models achieve the highest F1-scores for direct repetition in both languages (Table 3), with the French model reaching 81.5% and the Dutch one 73.2% In contrast, for self-repetition (Table 4), lexical and semantic cosine similarity models perform comparably or slightly better than BERT in French, while BERT achieves the highest scores in Dutch (F1-score = 94.0%). Across both phenomena, syntactic cosine similarity models consistently show lower performance. These results suggest that BERT models are particularly effective for detecting direct repetition, whereas lexical and semantic similarity provide competitive performance for self-repetition, especially in French.

Model	Lang	Prec.	Rec.	F1
Syntactic CosSim	FR	45.2%	54.6%	49.4%
	DU	39.6%	40.4%	40.0%
Lexical CosSim	FR	64.7%	71.6%	68.0%
	DU	64.0%	73.7%	68.5%
Semantic CosSim	FR	64.6%	75.0%	69.4%
	DU	49.6%	68.7%	57.6%
BERT	FR	**78.6%**	**84.6%**	**81.5%**
	DU	**66.1%**	**82.1%**	**73.2%**

Table 3: Performance of models on direct repetition detection across languages, reporting precision, recall, and F1-score

Model	Lang	Prec.	Rec.	F1
Syntactic CosSim	FR	48.9%	75.2%	59.3%
	DU	46.4%	75.2%	57.4%
Lexical CosSim	FR	81.6%	**88.3%**	**84.8%**
	DU	84.9%	86.5%	85.7%
Semantic CosSim	FR	**85.7%**	80.8%	83.2%
	DU	85.7%	80.8%	83.2%
BERT	FR	78.0%	80.7%	79.3%
	DU	**92.8%**	**95.2%**	**94.0%**

Table 4: Performance of models on self-repetition detection across languages, reporting precision, recall, and F1-score

4. Discussion

Extending the approach proposed by Fusaroli et al. 2023, this study computed cosine similarity across syntactic, lexical, and semantic representations to detect repetition patterns in autistic children's speech. Overall, the approach proved effective, particularly for self-repetitions.

In the case of *direct repetitions*, models detected a substantial portion of repetitions (recall around 75% or higher) using lexical and semantic similarity measures; however, precision remained limited due to numerous false positives. These findings suggest that predictions for direct repetitions should be interpreted with caution. The limitation likely arises from the annotation protocol, which labels segments as direct repetitions even when they share only a single content word (e.g., Adult: "Do you want a banana?" Autistic Child: "I like bananas"). Because such overlap represents only a small portion of the overall vector representation (especially in longer segments), vector-based similarity measures may not adequately capture these cases. In contrast, BERT models achieved more

balanced performance, with precision above 65% and recall above 80%, suggesting that contextualized representations are better suited for detecting direct repetitions.

Differently, in detecting *self-repetitions*, lexical and semantic similarity-based models performed consistently well across languages and remained competitive with BERT models. However, BERT models showed notable variability, with an F1-score of 94.0% for Dutch compared to 79.3% for French. This discrepancy likely reflects differences in generalization from training to test data rather than data size alone, as both models were trained on substantial datasets. The detection of self-repetitions requires capturing deeper linguistic relationships (e.g., dependency structures; cf., annotation protocol), which may not have been equally well learned by the models across languages.

Across all analyses, lexical and semantic similarity emerged as the most reliable indicators of repetition, yielding high precision and recall scores. In contrast, syntactic similarity consistently showed lower performance, suggesting that syntactic structure in spontaneous speech is highly variable and difficult to capture with surface-level representations. More advanced syntactic modeling may therefore be required to improve performance.

This study highlights the potential of similarity-based approaches for analyzing spontaneous speech in naturalistic contexts. Future work should extend this framework to a broader range of languages and age groups to explore how repetition patterns vary across different linguistic and developmental contexts. A more systematic investigation of cross-linguistic differences (both linguistic and methodological) could further clarify performance variation.

For instance, similarity distributions and thresholds differed between French and Dutch, with higher values observed for French in direct repetitions, whereas self-repetition results were largely comparable. This pattern may reflect differences in interaction styles or overall verbal output, but could also be influenced by language-specific properties or limitations of the NLP tools used. Indeed, French- and Dutch-based spaCy and SentenceBERT models are trained on relatively limited datasets compared to English resources and are optimized for written language, which may reduce their effectiveness on spontaneous child speech. Future research should therefore compare alternative models and embeddings, including those trained on spoken or child-directed data.

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. For instance, lemma-based rule systems or adaptive thresholding techniques could improve detection of direct repetitions. The observed differences between French and Dutch further suggest that both linguistic structure and NLP model limitations influence performance, underscoring the need for additional cross-linguistic exploration. Future research should also evaluate multilingual or fine-tuned models to enhance repetition detection across languages and spontaneous speech contexts.

We encourage interested researchers to test our models on their conversational data while considering the potential limitations. To facilitate this, the models presented in this paper are publicly available at this GitHub repository. In the similarity-based models, users can select linguistic levels for comparison (syntactic, lexical, and 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. Users can also test the trained BERT models on their own (French or Dutch language) data.

Finally, a key limitation of this study is the absence of a widely accepted definition of echolalia that allows for fully automated detection. Our annotation protocol attempts to address this issue using simplified linguistic criteria (e.g., comparing lemmas, POS tags, and dependency structures between linguistic units). However, this simplification introduces constraints. For example, evaluating similarity at the segment level may obscure word-level repetition patterns, and some detected cases may not align with traditional definitions of echolalia. Accordingly, the proposed models should be understood as an initial filtering step rather than a definitive classification tool. Establishing clearer and more consistent definitions of echolalia will be essential for improving future detection methods.

5. Ethics Statement

I testify on behalf of all co-authors that the present article was submitted following ethical principles in publishing. All authors declare no conflict of interest and agree that this research presents an accurate account of the work performed. All data presented are accurate, and methodologies are detailed enough to permit others to replicate the study. We share all code used to produce the work, including gold standard annotation protocol, inter-rater agreement calculation, model development and evaluation, and creation of the tables and graphs in the paper. In our code repositories, we provide instructions to interested users on how to apply our code to their own datasets.

However, raw and sensitive data such as speech transcriptions and speaker identifiers cannot be shared to protect the privacy and security of the participants. Ethical approval was obtained from the Ethics Committee of Erasme Hospital on 28 March 2023 (CCB B4062023000074). Written informed parental consent and minor assent were obtained from all participants prior to enrollment in the BeLAS study.

6. 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 pre-trained NLP models must be addressed. The tools used in this study, including BERT, BERT-SentenceTransformers, 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. Further 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 linguistic unit 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 linguistic unit. 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 the linguistic segment 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.

Sixth, we explored fine-tuning BERT models for repetition detection using our gold standard annotated data. While this yielded promising results for direct repetition, performance for self-repetitions was less reliable, and precision differed significantly between languages. Future research should investigate these differences, using larger training sets and alternative (e.g., multilingual) classification models.

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

7. Acknowledgements

This work was supported by the Excellence of Science grant (FNRS and FWO) for the Belgian Language in Autism Study (BeLAS). We thank all the

families and children who participated in the study. We are grateful to our colleagues in the BeLAS consortium for their valuable contributions, and to the research assistants for their help with coding and data management.

8. 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.

Nancy Bayley. 2006. *Bayley Scales of Infant and Toddler Development*, third edition. Harcourt Assessment, San Antonio.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Lloyd M. Dunn and Douglas M. Dunn. 2019. *Échelle de vocabulaire en images Peabody, Cinquième édition (EVIP-5)*. Pearson, Paris.

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.

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

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.

Ortho Édition. 2009. *Évaluation du langage oral de l'enfant de 2 ans 3 mois à 6 ans 3 mois*. Ortho Édition, Isbergues.

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.

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.

Liesbeth Schlichting. 2005. *Peabody Picture Vocabulary Test–III–NL: Nederlandse versie. Handleiding*. Harcourt Test Publishers, Amsterdam.

Eleanor Semel, Elisabeth H. Wiig, and Wayne A. Secord. 2020. *Clinical Evaluation of Language Fundamentals Preschool–3: Dutch Version*. Pearson Assessment, Amsterdam.

Peter J. Tellegen and Jacobus A. Laros. 2017. *SON-R 2–8: Snijders-Oomen Niet-Verbale Intelligentietest*. Hogrefe, Göttingen.

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.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

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.