Object Realisation in Spoken Guadeloupean French: Evaluating NLP Models for an Under-Resourced Variety

Amalia Canes-Nápoles, Sophie Repp
University of Cologne
Albertus-Magnus-Platz, 50923 Cologne, Germany
{acanesna, sophie.repp}@uni-koeln.de

Abstract

This paper contributes to the evaluation of natural language processing models applied to colloquial speech in lesser studied varieties of a language. We report on the performance of automatic speech recognition (ASR) and universal dependency (UD) parsing models in a radio corpus of colloquial French spoken in Guadeloupe (GuaFr), which is in contact with a typologically distant language, French-based Guadeloupean Creole (GuaCr). The corpus poses specific challenges due to phonetic and syntactic specifics of GuaFr, as well as the occurrence of code switching to GuaCr. At the ASR stage, we evaluate a decoding configuration designed to privilege acoustic evidence over language-model (LM) regularisation. We show that decoding behaviour varies depending on segment duration, speaker linguistic profiles, and acoustic context. For UD parsing, we investigate utterance segmentation as the primary lever to affect model performance and compare different segmentation sources (ASR punctuation, manual chunking, UD parser tokenization) and their combination. Focussing on the expression of syntactic objects, we highlight both strengths and pitfalls of the ASR and parsing models, and propose factor-aware evaluation diagnostics for multi-condition NLP pipelines on low-resource speech data.

Keywords: spoken-language processing, ASR, NLP evaluation, object realisation

1. Introduction

Natural language processing for lesser studied varieties of a language or special situational domains of interactive speech faces multiple challenges. In the domain of automatic speech recognition (ASR), dialectal pronounciation and diverging grammar, prompted by dialectal syntax and colloquial style are challenging for ASR models mostly trained on scripted speech in shorter sequences. In the domain of dependency parsing, dialectal and colloquial syntax pose challenges for universal dependency (UD) parsing models trained mostly on written, non-spontaneous language.

Previous research has shown that despite the significant improvement of data-driven dependency parsing systems in recent years, they still achieve a considerably lower performance in parsing spoken language data in comparison to written data (Zeman et al., 2018). In challenging discourse settings, like parent-child interaction with small children, where the child's grammar substantially differs from adult grammar (Liu and Prud'hommeaux, 2023), out-of-domain (written-trained) parsers achieve only moderate accuracy on this spoken data, while in-domain training and child age both systematically affect performance. Similar out-of-domain effects have been observed for Finnish data: Introducing a UD Finnish out-of-domain-treebank spanning web, clinical notes, online discussions, tweets, and poetry, Kanerva and Ginter (2022) show that a parser trained on written data drops from $\approx$ 91 in-domain to ≈ 77.5[1] underscoring that written-trained parsers do not capture domain-specific phenomena in spoken and informal text.

Given this significant gap in parsing performance between the two modalities, spoken and written language, Dobrovoljc and Martinc (2018) investigate which speech-specific phenomena influence the poor parsing performance for speech in the Spoken Slovenian Treebank (Dobrovoljc and Nivre, 2016), and to what extent. The results show that utterance segmentation is the most prominent cause of low parsing performance, both in parsing raw and pre-segmented transcriptions (normalized transcriptions excluding disfluencies, discourse markers and fillers).

In the current study, we evaluate the performance of NLP models applied to spoken language in an under-resourced variety of French, Guadeloupean French (GuaFr), which is especially noteworthy due to its contact with a typologically distant language, French-based Guadeloupean Creole (GuaCr). The data that we submit to NLP stems from call-in radio shows broadcast in Guadeloupe, where individual speakers calling in speak GuaFr occasionally code-switching to GuaCr. The particular challenges this data type pose are the colloquial style and dialectal specifics of GuaFr phonetics and syntax as well as code-switching between GuaFr and GuaCr, i.e.,

[1] Measured in terms of parsing accuracy LAS, i.e., the percentage of words that are correctly assigned both the right head (or parent word) and the correct dependency label (or relation).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2299–2308
11–16 May 2026.

the correct classification of the code-switching sequences both in ASR and in dependency parsing.

The focus of our study regarding dialectal syntax is model performance for referential direct objects. GuaFr syntax has been reported to differ from standard Hexagonal French (HexFr) syntax in that object pronouns/clitics referring to contextually given information can be easily left out, as is illustrated below for direct objects (1) and for partitives/obliques (2) (Kriegel and Ludwig, 2018; Pustka, 2007). In (1) and (2) anaphoric null objects and their antecedent are red-coded, and the head verb is in boldface.

(1) Vous **prenez** un manioc, vous
you take.2PL a manioc you
le$_{\text{HexFr}}$/∅$_{\text{GuaFr}}$ **épluchez**...
PRO:CL-ACC peel.2PL
'You take a manioc, you peel (it)...'

(2) Il y en$_{\text{HexFr}}$/∅$_{\text{GuaFr}}$ **a** qui
EXPL LOC PRO:CL-PART COP.3SG REL
y$_{\text{HexFr}}$/∅$_{\text{GuaFr}}$ **vont**.
LOC go.3PL
'There are some (of them) who go (there).'

In addition to the occurrence of null objects, GuaFr is likely to differ from HexFr in the realization of non-given arguments that may be left implicit, depending on the particular verb lemma in this variety of French (e.g., English *I read (a book)*).

We note that these phenomena are not to be viewed as deviant from a GuaFr grammatical point of view. In general, we are not concerned with speech errors, disfluencies or repairs in this paper, whose annotation has been reported to be rather diverse, and partly conflicting in treebanks of spoken language, see Dobrovoljc (2022) for an overview. We are focussing on phonetic and syntactic specifics of colloquial GuaFr, and the challenges they pose to NLP. Apart from the realization of null and implicit arguments, this includes UD parsing for objects in general, to appreciate potential differential object realization in GuaFr vs. HexFr.

In the context of the findings regarding parsing performance for spoken language reviewed above, we treat utterance segmentation as the primary lever to affect parsing performance. As explained in Section 4, unlike Dobrovoljc and Nivre (2016) we do not post-process the transcripts; instead, we use punctuation generated by the ASR model, which we expect to be prosodically informed and therefore to yield a more natural segmentation of spontaneous speech and a better prediction of syntactic objects. This choice aligns with findings that incorporating prosodic cues boosts speech parsing: end-to-end systems parsing *from audio* outperform text-only pipelines, with graph-based models strongest on spontaneous French (Pupier et al., 2024). Our use of ASR punctuation approximates those prosodic benefits within a pipeline.

The paper is structured as follows. In Section 2 we provide information on the GuaFr spoken language corpus and give an overview of the NLP pipeline that we applied to these data. In Section 3, we report our results for the ASR recognition; Section 4 provides the dependency parsing evaluation. Section 4.3 offers the discussion and concludes.

2. Data and NLP pipeline

2.1. Data: GuaFr radio show

The data we submitted to NLP was sourced from the radio talk show *Allô j'écoute*. In the show, community members call in to share their opinions on the programme's topic, which typically is of significant relevance to the community (e.g., proposed laws or local events). There are more than 400 available programmes, each lasting 26–30 minutes. In each show, there is one host and up to 12 speakers with a balanced gender distribution, each of whom has 2–5 minutes. The host creates a friendly, conversational atmosphere to encourage open dialogues. The database used for the current evaluation contains data from 12 speakers from different installments of the show.

2.2. Overview of the NLP Pipeline

Our pipeline comprises two stages: automatic speech recognition (ASR) and universal dependency (UD) parsing. In both stages, we tuned parameters expected to influence performance. For ASR, our objective was a *literal–FR* decoding that privileges acoustic evidence over language-model (LM) regularisation. Concretely, we reduced the decoder's LM pull so that hypotheses shift from "the most plausible French" to "what was actually said." Here, "LM" denotes the decoder's language-modelling tendency to prefer high-frequency collocations given the audio and any preceding tokens. See Section 3.1 for details regarding the method for ASR.

Regarding dependency parsing, the presence of ASR in the pipeline opened up additional options for parameter manipulation. As already highlighted, the ASR model provides segmentation in terms of punctuation, and it is plausible to assume that this segmentation helps UD parsing: Spontaneous spoken language often is not well-planned, full of interruptions, omissions, truncations and dislocations, which is not frequent in written style, and these features may be captured well enough by ASR punctuation. We tested this hypothesis by providing ASR punctuation to the UD parser, or not. A second parameter that we tested as potentially influencing the predictions of the UD parser was

the application of manual chunking on the basis of prosodic information, thus identifying main clause units. The third parameter that we manipulated was the UD parser internal tokenziation. Details for all these parameters will be given in Section 4.1.

3. ASR Transcription

3.1. Method

Task. We automatically transcribed short audio segments extracted from the call-in contributions, typically ranging from 0.63 to 6.84 seconds in duration. These segments may include code-switching between GuaFr and GuaCr. We opted for the OpenAI Whisper ASR model (`large-v3`) (Radford et al., 2023) to achieve two goals. We wanted to (i) *avoid gaps* in the transcription in the context of non-standard spoken French in contact settings, and (ii) produce a transcript that stays *as close to the acoustics as possible*, i.e. a transcript that results from *literal–FR* decoding.

Strategy. To meet goal (ii), a *literal–FR* decoding, which would also feed goal (i), we deliberately weakened the decoder's language-model (LM) pull so hypotheses are driven more by acoustics than by collocational plausibility. Table 1 gives a summary of the Whispher decoding parameter settings and expected changes in literalness. First, we fixed the language model to prevent multilingual pivots and "helpful" normalisation into HexFr. Next, we reduced contextual bias: `condition_on_previous_text` was decreased (set `false`) to turn off cross-window context, because carrying over prior tokens amplifies LM expectations; we also omitted `initial_prompt` by default.

Parameter	Baseline	Literal–FR
Language	`none`	`language=fr`↗
Condition on previous text	`true`	`false`↗
Initial prompt	"text"	`null` ↗
Beam size	≥ 4	**1** ↑
Temperature	0.0 – 0.7	**0.0** ↗
Best of	≥ 3	**1** ↑
Patience	≥ 1.5	**1.0** ↗
Thresholds	enabled	`none` ↑

Table 1: Whisper decoding parameters for the *literal–FR* evaluation in comparison to the baseline: Arrows indicate *expected* change in *literal-FR*: ↑ strong increase; ↗ moderate increase; → neutral.

We then constrained the search/decoding so the model does less LM-driven "polishing." Concretely, `beam_size` was decreased to 1 (greedy) because larger beams explore and re-rank toward high-probability, standardised formulations; `temperature` was fixed at 0.0 to eliminate sampling that, together with re-ranking, can privilege LM-plausible variants; `best_of` was reduced to 1 so multiple samples are not generated and re-scored; and when beam search is enabled, `patience` was kept low ($\approx$ `1.0`) to avoid extended exploration that tends to "settle" on normative phrasing. Finally, we relaxed normalisation/filters: `logprob_threshold`, `compression_ratio_threshold`, and `no_speech_threshold` were disabled (set `none`) to minimise pruning of low-probability but acoustically supported tokens that are characteristic of our data (repetitions, clipped syllables, filled pauses).

In sum, our parameter settings in *literal-FR* push every relevant lever in the "more literal" direction: force the language model (French), *decrease* reliance on prior context and prompts, *decrease* exploration/re-ranking (greedy, deterministic, low patience, no best-of), and *disable* aggressive filtering. This favours "what was said" over "what is likely to be said," which proves advantageous for non-standard French, specifically the potential omission of object expressions, code-switching, and disfluency-rich speech.

3.2. General evaluation of decoding parameter settings

To assess the general transcription performance of the two decoding settings (baseline and *literal-FR*) as outlined in Section 3.1, we first investigated overall error rates and then examined three potential sources for error variation in the data set. Our evaluation of ASR Handling of GuaFr dialectal syntax and code-switching is presented further below, see Section 3.3.

Overall comparison. First, we considered word error rate (WER) computed against the manually segmented reference transcriptions created in Praat (Boersma, 2001). Performance differences were small. The baseline preset yielded a mean WER of 0.233 (median 0.111), while the literal setting produced 0.242 (median 0.125). Character error rate (CER) values were similarly close (0.153 vs. 0.160). The distribution of error types was also comparable across settings, with substitutions representing the largest component of the error profile (mean $\approx$ 0.94–1.02), followed by deletions ($\approx$ 0.55) and insertions ($\approx$ 0.32). Overall, these results indicate no substantial performance difference between the two decoding settings.

Error variation: Heterogeneity in the data. Treating all evaluation segments as a homoge-

neous set as in the comparison may obscure systematic variation arising from differences in the data itself. In particular, the speech material analysed here is highly heterogeneous with respect to the following three aspects: (i) segment duration, i.e., the intonation unit identified by the authors, (ii) linguistic behaviour of individual speakers, and (iii) linguistic context. To better understand the sources of variability in transcription accuracy, we examined the performance of the decoding settings across these three additional factors.

First, regarding WER changes as a function of **segment duration**, our analysis shows that short segments tend to be transcribed less accurately in all decoding settings, and that performance generally improves as segment duration increases, with the more literal settings showing better results than the baseline for longer segments (see Figure 1).

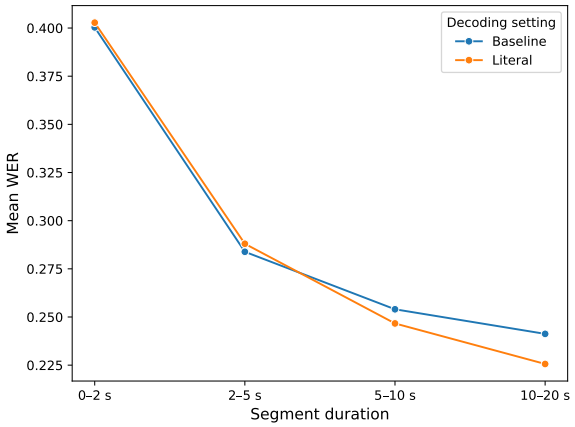

Figure 1: Mean WER by duration and decoding setting.

Second, regarding **speaker variability**, there is substantial variation in the total amount of speech the individual speakers contribute and in their linguistic profiles, particularly with respect to their use of GuaFr and GuaCr. Figure 2 shows the difference in WER between the literal and baseline settings for each speaker. Positive values indicate cases where the baseline preset performs better, whereas negative values indicate better performance for the literal setting. The results reveal substantial speaker-specific variation.

To explore whether this variation relates to speaker's linguistic profiles, we constructed a *creole profile*: the share of total speaking time labelled as 'creole' relative to the speaker's total speech duration. Figure 3 relates this creole profile to the speaker-specific WER difference between decoding settings. Although the number of speakers is at present limited, visual inspection of the regression line suggests a tendency whereby speakers with very low proportions of GuaCr tend to show

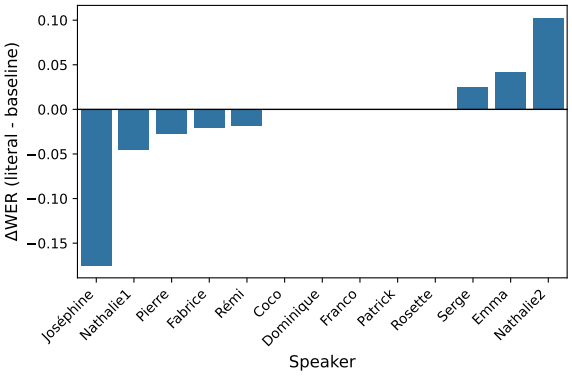

Figure 2: Inter-individual variation in the data: speaker-specific WER difference relative to the baseline, segment-based parameters.

slightly better performance under the baseline preset, whereas speakers with higher proportions of GuaCr show smaller differences between settings and occasionally better results with the literal configuration. Further studies with more speakers should confirm this exploratory trend.

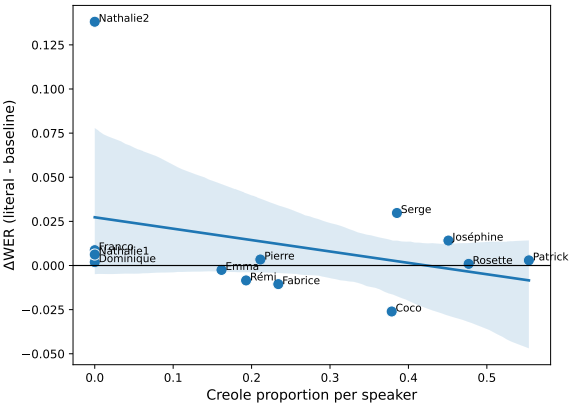

Figure 3: Relation between speaker Creole usage and decoding setting performance difference.

Finally, we examined the role of **acoustic context** during transcription (in a similar vein as done with the UD parser model, cf. Section 4). Specifically, we compared a *segment*-wise decoding condition –where each gold segment is transcribed independently– with a *joint* condition in which consecutive GuaFr segments belonging to the same speaker are concatenated and transcribed together. The resulting *joint* transcription is then projected back onto the original gold segmentation for evaluation. This comparison allowed us to test whether providing the model with a larger contextual window improves recognition accuracy, especially in light of the trend observed in Figure 1. Surprisingly, the *joint* condition resulted in substantially higher error rates. For the baseline preset, mean

WER increased from 0.233 to 0.377. A similar pattern was observed for the literal setting, where mean WER rose from 0.242 to 0.383. To assess whether this difference was systematic, we conducted a Wilcoxon signed-rank test (n = 359 per condition). The test revealed a highly significant difference between segment-wise and *joint* decoding for both settings (baseline: median ΔWER = -0.067, $p < 10^{-23}$; literal: median ΔWER = -0.056 $p < 10^{-24}$). Together with the segment-length results, this suggests that context helps recognition only within a limited range. The prosodically informed, clause-guided segmentation provided by the authors appears to offer a more appropriate inference unit than substantially longer stretches of speech corresponding to the concatenated segments.

3.3. Evaluation: ASR Handling of Spoken-Language & GuaFr Features

As mentioned in Section 1, previous studies have identified a number of features that distinguish GuaFr from HexFr (Kriegel and Ludwig, 2018; Pustka, 2007). Many of these phenomena, however, are in fact not specific to GuaFr: they have also been widely discussed in research comparing spoken and written varieties of HexFr. Typical characteristics of spoken French across varieties include, of course, repetitions and disfluencies (e.g., hesitation markers) but also differences in discourse markers, and the frequent omission of the preverbal negative particle *ne*, whose meaning is conveyed solely by the post-verbal marker *pas*. For GuaFr, which is in contact with GuaCr, additional phenomena that have been reported are the absence of copulas, clitic pronouns, certain prepositions, and complementizers.

Table 2 summarizes how the various categories are treated by the different transcription settings, indicating the extent to which they are omitted or hallucinated in the automatic transcriptions. For categories associated with spoken language (e.g., discourse markers and disfluencies), values correspond to coverage, defined as the proportion of occurrences present in the gold transcription that are preserved in the hypothesis. Higher values therefore indicate better preservation of these spoken-language features. For categories associated with potential normalization or grammatical insertion (e.g., clitics, copulas, complementizers, and prepositions), values correspond to the hallucination rate, defined as the proportion of occurrences in the hypothesis that do not correspond to an item present in the gold transcription. Lower values therefore indicate fewer inserted elements. All proportions are calculated with respect to the total number of occurrences of each category in the relevant reference set (gold transcription for coverage; hypothesis for hallucination).

Gram Class	Baseline	Literal
	Coverage ↑	
discourse marker	0.875	0.875
disfluencies	0.625	0.642
	Hallucination ↓	
ce	0.220	0.140
clitics	0.046	0.050
complementizers	0.103	0.108
copulas	0.044	0.051
neg *ne*	0.364	0.393
prepositions	0.074	0.091

Table 2: Coverage and hallucination rates for selected grammatical categories across baseline and literal settings.

We now turn to some qualitative analyses of the model's performance regarding the potential omission of object expressions, the positioning of realized object clitics, and code-switching.

Null objects. There were unexpectedly few instances of null objects in the data that would be ungrammatical in HexFr: All sentences with clitics and other short phonetically non-prominent object pronouns with a low confidence score for alignment were checked in a close phonetic analysis. For the null objects that would be ungrammatical in HexFr, Whispher ASR did not hallucinate object clitics. An example with null objects from the data is shown below.

(3) C' est le bonheur mais il
DEM be.COP DEF happiness but EXPL
faut ∅ savoir **cultiver** et
be.necessary OBJ know cultivate.INF and
on ∅ ka **chercher** toujours.
INDF OBJ PROG search.INF always

'This is happiness, but one must know how to cultivate (it) and we keep searching for (it).'

Position of object clitics. While null objects were not reconstructed by the model, the word order for object clitics was often forced to their expected distribution in HexFr. One case is illustrated in (4), where the speaker uses the clitic pronoun not pre-verbally, where it would occur in HexFr, but preceding an adverb inserted between the clitic and the verb. So, while the model does not hallucinate null objects, it does have a strong bias towards clitic position in the sentence.

(4) a. Il faut la$_{\text{GuaFr}}$ juste
EXPL be.necessary OBJ just.ADV
analyser.
analyse.INF
'You just have to analyze it.'

b. Il faut juste l'$_{\text{HexFr}}$
EXPL be.necessary just.ADV OBJ
analyser.
analyse.INF
'You just have to analyze it.'

Future work must show in how far such instances can also be found in colloquial variants of other French varieties.

Code switching. Low scores of word alignment (implemented via whisperX (Bain et al., 2023)), marked instances of single-word switches. Note that segments with longer GuaCr stretches were not part of the evaluation reported here. Two of the single-word switches stood out. For pronominal reference to non-subject first person singular, which in HexFr is realized by the clitic *me*, the GuaCr form *mwen* was occasionally used, see (5a). For pronominal reference to non-subject third person singular non-animate, which in HexFr is realized by the pronoun *ça* or the clitic *le*, the creole object pronoun *y* was occasionally used, see (5b).

(5) a. Merci déjà de
thank you already for
mwen$_{\text{GuaCr}}$(me$_{\text{HexFr}}$) **donner** la
DAT.1.SG give.INF the
parole.
word
'Thank you already for giving me the floor.'

b. Pourquoi ils **cherchent** y$_{\text{GuaCr}}$(ça$_{\text{Fr}}$)
why they search.3PL OBJ.3S
dans toutes vitesse et on pas
at full speed and they not
heureux?
happy
'Why are they searching for it so desperately and still not happy?'

Overall, the analyses reported in Sections 3.2 and 3.3 indicate that while the two decoding parameter settings yield comparable performance on average, their relative behaviour can vary under different conditions, including segment duration, speaker linguistic profiles, acoustic context. These findings next to the presence of spoken language or contact-induced linguistic features highlight the importance of considering both data heterogeneity and segment-internal linguistic variation when evaluating ASR systems of spoken language in contact settings.

4. Dependency parsing

The focus of the evaluation of UD parsing was the automatic identification and classification of direct and indirect objects because of the differential object realization in GuaFr and HexFr highlighted in Section 1. The parser we used is STANZA (Qi et al., 2020) due to its strong UD results (UD v2.5): macro-averaged across 100 treebanks, it leads most pipeline scores, and on French (UD *French-GSD*) it reports UAS/LAS of 91.38/89.05, outperforming UDPipe v1.2 (87.14/84.26) and spaCy v2.2 (67.46/60.60). These scores were computed with the CoNLL-2018 UD evaluation script. (Qi et al., 2020)

4.1. Method

Experimental factors. As laid out earlier, we manipulated three input parameters to assess the impact of segmentation on dependency parsing:

i. *Acoustically informed segmentation Whisper*: We either retained the punctuation supplied by the Whisper ASR (model `large-v3`) or removed it.

ii. *Manual clause segmentation by authors*: The input was either the caller's entire per-speaker text for the call or the same speech segmented into main-clause units.

iii. *Sentence tokenisation in Stanza*: We either applied Stanza's sentence splitter or bypassed it.

Table 3 summarizes these three parameters. Crossing the three binaries yields eight conditions.

ID	W	A	S
W-A-S	W	A	S
W-A-s	W	A	s
W-a-S	W	a	S
W-a-s	W	a	s
w-A-S	w	A	S
w-A-s	w	A	s
w-a-S	w	a	S
w-a-s	w	a	s

Table 3: Eight conditions for the evaluation of UD parsing. Legend: W/w = Whisper punctuation kept/removed; A/a = manual main-clause chunks on/off; S/s = Stanza sentence split on/off.

Chunking regimes. In the following we discuss the three segmentation input parameters.

(i) *ASR (Whisper punctuation):* Punctuation produced by Whisper during decoding is treated as a prosody-informed cue to segmentation (see *Strategy* in Section 3.1 for decoding choices that reduce

LM bias toward "most plausible French"). Whisper is trained to predict raw, unnormalised transcripts, so punctuation and casing are part of the targets. Punctuation decisions come from an audio-conditional language model: the decoder conditions on the audio representation (which implicitly captures cues like pauses and energy changes) and on preceding text tokens.

(ii) *Human (Manual segmentation):* Sentence boundaries are fixed to the externally prepared gold segments (pre-segmented and manually revised). Segmentation follows a prosodically-informed clause-based criterion, i.e., each segment contains at least one lexical verb with its dependents but can contain more verbs in the case of embedded clauses. We intentionally preserve spoken-language phenomena (discourse markers, repetitions, truncations, dislocations, disfluencies); no normalisation or removal was performed.

(iii) *Parser (Stanza tokeniser):* Stanza performs joint tokenisation and sentence splitting as a single character-level tagging task, predicting token boundaries, sentence boundaries, and multi-word-token (MWT) ends. The tokeniser/splitter is trained and evaluated on 100 UD v2.5 treebanks (Zeman et al., 2019), with development data formed by a 20% random split of training sets. In our setup, Stanza's sentence boundaries are used only when enabled; otherwise, parsing respects the chunks that are manually provided, without additional splitting.

Alignment and normalisation. To compare outputs token-by-token, we proceeded in two stages. First, to generate input without segmentation, we reconstructed one continuous text per speaker from the gold annotation (time-aligned transcripts) by concatenating tokens in gold order and preserving the original character offsets (end-exclusive). We retained all punctuation as tokens and used the expanded form of contractions (e.g., *du* → *de le*, *aux* → *à les*), mirroring the format of the gold. In the rare case of a residual contracted form (e.g., *au*), we kept the gold surface form as is, to preserve alignment. We normalised the text and preserved any gold whitespace so that each gold token remained recoverable via (`speaker, start_char, end_char`).

In the second stage, we post-processed Stanza output for alignment: To ensure 1:1 comparability at token level while keeping the gold surface text authoritative, we applied a minimal, deterministic normalisation to the Stanza output before alignment:

i. *Same-span splits (MWT):* When Stanza produced multiple tokens with identical spans (e.g., *à le* for a gold *au*), we collapsed them to a single row by concatenating their surfaces (features taken from the first non-PUNCT child).

ii. *Contiguous covers:* When a sequence of Stanza tokens exactly covered one gold span (e.g., *-t* + *-elle* for gold *-t-elle*), we collapsed the sequence into one row spanning the gold offsets and concatenated the surfaces.

iii. *Hyphen/dash harmonisation:* We unified visually identical hyphens by treating all Unicode dash punctuation equivalently, and—where the gold attaches a hyphen to the following word (e.g., *-là*)—we merged a stand-alone Stanza hyphen token into that following word if the spans were contiguous.

After the final step, Stanza's output had at most one row per gold span, enabling a strict 1:1 left-join on (`speaker, start_char, end_char`). For evaluation, we kept the gold surface forms and compared Stanza's predicted labels (e.g., *deprel, upos, lemma*) against the gold at aligned token positions.

Adaptation of UD categories. UD parsing does not assume the distinction between arguments and adjuncts that is assumed in many linguistic theories. Rather, the distinction between core arguments (subjects/objects) and other clause-internal dependents is substantially informed by the morphosyntactic characteristics of the dependent (De Marneffe et al., 2021). For instance, prepositional phrases are parsed as oblique modifiers even if the verbal predicate selects them, whereas pronouns (without a preposition) are parsed as core arguments.

In our study of object realization in GuaFr, we rely on the property of a verbal predicate to require a (core) argument irrespectively of that argument's morphosyntactic makeup, even though we are focussing on null realizations in this paper: The availability of null objects or implicit objects plausibly depends on the particular verb. Therefore, we pooled UD categories pertaining to objecthood in the sense of verbal argument selection into the following broad(er) categories: object (`obj`), indirect object (`iobj`) and prepositional object (`obl:arg`), independently of morphosyntactic make-up. `obj` did not include objects of light verb constructions; we kept the latter separate because they are not referential (`obl:lvc`). We also kept objects that were the agent of non-finite embedded clauses (aka control constructions) separate (`iobj:agent`). Finally, we used the UD categories for object clauses (finite: `ccomp`, non-finite: `xcomp`).

4.2. Evaluation

We evaluated a total of 348 tokens of object realisations by 12 speakers.

Overall evaluation metrics. Table 4 gives the macro metrics for each evaluation condition. Overall, Whisper punctuation W improved performance for dependents. Additional manual main clause segmentation A did not yield further overall improvement, although we will see further below that this general trend does not hold for all object categories. As can be seen in Table 4, in human segmentation mode A, the stanza `ssplit` toggle S/s has no effect because each manually segmented chunk contains a single gold sentence; thus *W-A-S* and *W-A-s* are identical. Punctuation removal $W \rightarrow w$ still changes $\approx$ 3–4% of predictions under A. In RAW mode a, both W/w and S/s materially change predictions ($\approx$ 10% for S/s).

However, this general trend does not hold for the identification of indirect objects (`iobj`) and for objects in light verb constructions (`obl:lvc`): `iobj` were classified most often correctly without any segmentation, i.e., also excluding Stanza tokenisation in *w-a-s*. At the same time, lack of any segmentation also produced more misclassifications of direct objects `obj` as `iobj`. We will report on the most frequent misclassifications further below.

	Precision	Recall	f1
W-A-S	0.708	0.716	0.712
W-A-s	0.708	0.716	0.712
W-a-S	0.699	0.707	0.702
W-a-s	0.633	0.672	0.651
w-A-S	0.648	0.687	0.664
w-A-s	0.651	0.693	0.668
w-a-S	0.605	0.655	0.619
w-a-s	0.589	0.686	0.627

Table 4: Macro metrics per condition.

Recall variability by object category. Figure 4 indicates which gold categories are most sensitive to the conditions. It shows the top-8 labels ranked by recall variability, the height of each bar reflecting the difference between maximum and minimum recall across conditions. As can be observed, among the parsed and classified object expressions, condition sensitivity is strongest for `obj:lvc`, `ccomp`, and `iobj` (large variability despite modest–moderate support), while `obj` shows sizeable support with only moderate variability. Expectedly, `null_obj` is consistently unrecovered (100% misclassification with 0 pp variability[2]), indicating a systematic failure to detect null objects, which are not covered in UD parsing, rather than condition effects.

[2]Throughout, pp = percentage points, i.e. absolute differences between percentages, rather than relative changes with respect to the initial value.

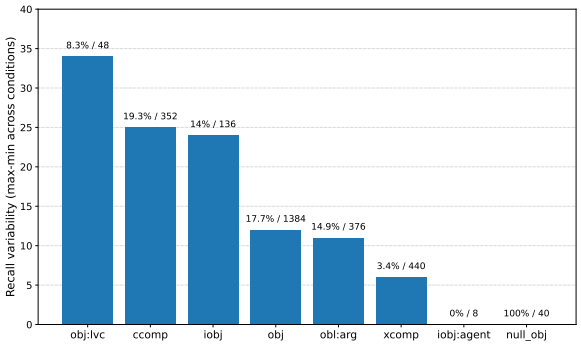

Figure 4: Recall variability (max – min) per object category across conditions in percent within category. The numbers above each bar are the overall proportion of misclassification for each category with the raw gold token count pooled across conditions (left), and raw counts referring to all predictions made for that gold category across all conditions (right).

Recall misclassification direction by object category. Despite showing the general sensitivity of object categories to the evaluation conditions, Figure 4 is not informative with respect to how each condition shifts the model's decisions or where the errors land. Figure 5 zooms in on the most volatile misclassification edges. An edge is a specific misclassification flow gold $\rightarrow$ pred(icted label) (e.g., `obj` $\rightarrow$ `iobj`). For each edge and condition we plot impact on gold = 100 × count(gold $\rightarrow$ pred)/ gold_support for that condition. Volatility is the range of this impact across conditions; we plot trend lines for the top-K edges by this volatility. The legend within Figure 5 reports the overall share of that edge among the gold label's errors (aggregated across conditions).

Figure 5 indicates that the largest swings come from off-target predictions: `ccomp` $\rightarrow$ PRED_OTHER dominates `ccomp`'s errors and is highly condition-sensitive ($\approx$ 11–12% up to $>$ 35% impact on gold, then falling and rising again), while `obj` $\rightarrow$ PRED_OTHER varies more moderately ($\approx$ 12–19%) with a late uptick. The category PRED_OTHER groups the verb instances that (i) do not have an object because they are used as participles with an attributive function (e.g., *c' est* ***noté*** 'that's noted'), (ii) occur in fixed expressions in which the object is not referential (e.g., *ça fait il y* ***a*** *quelques années* 'it's been some years ago...'), (iii) are in truncated segments (e.g., *Moi, je te* ***disais*** *...* 'I was telling you...'), (iv) occur with experiencers (e.g., *Il faut tu trouves celle qui te* ***convient***), and (v) occur in a segment that contained creole material.

Two label-conflation edges are also prominent. `iobj` $\rightarrow$ `obj` accounts for essentially all `iobj` errors: $\approx$18% early, dipping, spiking to $\approx$ 22%, then

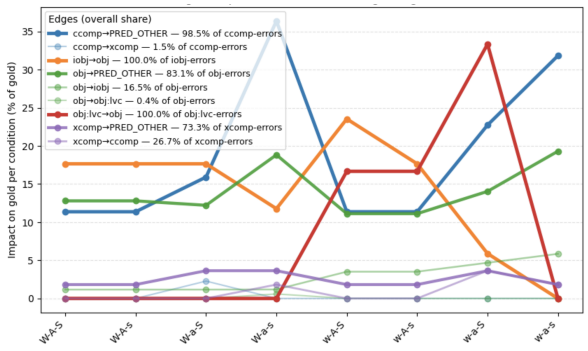

Figure 5: Recall variability per condition for the most frequent misclassification directions (edges).

collapsing near zero in the final setting, suggesting those settings largely fix this confusion. `obj:lvc` → `obj` appears in bursts, near zero in the first half, then sharp jumps (≈ 17–28%) before returning to zero, consistent with condition-triggered conflation. The remaining edges (e.g., `obj` → `iobj`, `obj` → `obj:lvc`, `xcomp` → PRED_OTHER/`ccomp`) are low-magnitude (≈ 1–4%) and comparatively stable. A main take-away is that sensitivity is edge-specific, not monotonic. The common expectation that moving from conditions *W-A-S* to *w-a-s* uniformly increases errors does not hold: some edges peak mid-grid, some decline, and others show spikes tied to particular conditions.

Concerning the conditions impact across the 2×2×2 grid, A/a is the dominant driver of volatility, W/w mostly moderates off-target drift, and S/s chiefly controls conflation edges:

- A→a **(on→off)**: Turning A off markedly amplifies off-target errors for `ccomp` and `obj` (the `ccomp`→PRED_OTHER line surges to ≈ 35 pp when A is off; obj→PRED_OTHER rises toward ≈ 19 pp). In contrast, `iobj`→`obj` declines when A is off, indicating A interacts with the parser to promote that specific conflation when on.
- W→w **(on→off)**: Switching W off generally dampens off-target drift (e.g., `ccomp`→PRED_OTHER relaxes toward ≈ 11–12 pp), but the protection disappears when A is also off: the off-target rate climbs again in the rightmost setting, showing a clear $W \times A$ interaction.
- S→s **(on→off)**: S governs label conflations. With S on, both `obj:lvc`→`obj` and `iobj`→`obj` activate (bursts to ≈ 17–28 pp and ≈ 22 pp, respectively); with S off, these conflations collapse toward zero, while off-target edges change little, so S's effect is specific to which target label errors fall into, rather than whether the model goes off-target at all.

4.3. Discussion

Stepping back from individual scores, the pattern is that segmentation drives the error behaviour in two distinct ways. First, when boundaries diverge from the parser's training regime (as with human clause chunks or when ASR punctuation is weakened), the model tends to under-commit on syntactic decisions, yielding off-target drift for `ccomp` and `obj`. Second, when tokenization or sentence splitting perturbs sentence structure, errors conflate adjacent roles, notably `iobj`→`obj` and `obj:lvc`→`obj`, revealing that segmentation is deciding where misclassifications land, not only how many there are. The one surprise, `iobj` recall peaking with no segmentation, suggests that punctuation and splitting sometimes mislead the parser for `iobj` contexts (e.g., clitic placement or truncated spans), a cue to design segmentation that preserves these micro-cues or to adapt the parser to human/ASR boundaries.

Limitations. Some labels have very small support (e.g., `iobj:agent`), making per-condition estimates unstable; moreover, macro-averages weight rare and frequent labels equally, which can under- or over-emphasize certain patterns. As the corpus grows, and especially with targeted enrichment of under-represented constructions, these estimates should stabilize; meanwhile, we report raw counts alongside percentages. The volatility metric (range) is robust and easy to read, but future work could complement it with dispersion measures or mixed-effects modelling over items.

Conclusions

We evaluated ASR and UD parsing for GuaFr under contact with GuaCr, focusing on how transcription settings, segmentation strategies, and linguistic variation shape decoding and parsing errors. At the ASR stage, decoding presets designed to privilege acoustic evidence over language-model regularisation produced comparable overall WER performance, but their behaviour varied depending on segment duration, speaker linguistic profiles, and acoustic context. At the parsing stage, sentence segmentation proved to be the main factor affecting performance, with parser-native tokenization and ASR punctuation generally providing the most reliable boundaries. Methodologically, plotting recall variability, volatile misclassification edges and their impact on gold, and separating off-target from wrong-target errors provided compact, factor-aware diagnostics. Overall, our approach offers reusable, factor-aware evaluation primitives for multi-condition NLP on low-resource, contact-variety speech.

5. Bibliographical References

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

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Kaja Dobrovoljc. 2022. Spoken language treebanks in universal dependencies: an overview. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1798–1806, Online. European Language Resources Association (ELRA).

Kaja Dobrovoljc and Matej Martinc. 2018. Er... well, it matters, right? on the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 37–46.

Kaja Dobrovoljc and Joakim Nivre. 2016. The universal dependencies treebank of spoken slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573.

Jenna Kanerva and Filip Ginter. 2022. Out-of-domain evaluation of Finnish dependency parsing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1114–1124, Marseille, France. European Language Resources Association.

Sibylle Kriegel and Ralph Ludwig. 2018. Le français en espace créolophone–guadeloupe et seychelles. *Romanistisches Jahrbuch*, 69(1):56–95.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

Adrien Pupier, Maximin Coavoux, Jérôme Goulian, and Benjamin Lecouteux. 2024. Growing trees on sounds: Assessing strategies for end-to-end dependency parsing of speech.

Elissa Pustka. 2007. Le français régional émergent en guadeloupe. *Bulletin PFC*, 7:261–271.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

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

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

6. Acknowledgments

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 281511265 – SFB 1252 Prominence in Language, University of Cologne.