An intonosyntactic treebank for spoken French: What is new with Rhapsodie?

María Paz Botero-Garcia[1] **Sylvain Kahane**[1,3] **Emmett Strickland**[1]
Bruno Guillaume[2] **Anne Lacheret-Dujour**[1]

[1]MoDyCo, Université Paris Nanterre & CNRS [2]Université de Lorraine, CNRS, Inria, LORIA [3]Institut Universitaire de France

Abstract

This paper presents a new format of the Rhapsodie Treebank, which contains both syntactic and prosodic annotations, offering a comprehensive dataset for the study of spoken French. This integrated format allow us for complex multilevel queries and open the way for the extraction of intonosyntactic studies.

1 Introduction

The Rhapsodie Treebank is the outcome of the French National Research Agency (ANR) project Rhapsodie, which began in 2008. It is the fruit of years of work by a group of French researchers who collected 3 hours and 10 minutes of spoken French audio, transcribed it, analyzed it, and developed a multi-level annotation scheme (wich involves syntax and prosody) that is reproducible and allows for the study of the syntax-prosody interface in French (Lacheret-Dujour et al., 2019b).

The main interest of this corpus, in addition to its multilevel annotation, lies in the richness of its metadata. It is composed of 30 monologues and 27 dialogues produced with different communicative goals, which may belong to public or private social contexts and can be spontaneous, semi-spontaneous, or planned, with varying degrees of interactivity.

In this paper, we propose to implement and expand upon the methodology introduced in Strickland et al. (2024) for Naija, or Nigerian Pidgin, to provide a new version of the Rhapsodie treebank where the different annotation layers (morphosyntax and prosody) are represented in a unified structure. The main benefit is that this version allows for a more in-depth study of the interaction between syntax and prosody. This is illustrated in the paper cited above on Naija.

2 Combining syntax and prosody in one single format

In 2024, the intonosyntactic treebank for the Naija Strickland et al. (2024) introduced for the first time a format in which every node annotated with syntactic information, (i. e., each token), is associated with child nodes corresponding to its constituent syllables, which are annotated with automatically extracted prosodic information. Since the Rhapsodie project had already been manually annotated at both the syntactic and prosodic levels years earlier, this development represented a valuable opportunity to adapt the original Rhapsodie corpus to the new format, while preserving the original manual annotations and incorporating new ones.

The main difference between the Rhapsodie intonosyntactic treebank and the Naija intonosyntactic treebank is the presence of micro- and macro-syntactic annotations and extra prosodic annotations at the token level and not just at the syllable level, like the token's position within an intonative period, metrical foot, rhythmic group or an intonational package, that will be explained later. The inclusion of prosodic information altered the structure of the dependency tree, as two tokens may share the same syllable (fused syllable). Their corresponding subtokens are therefore connected accordingly, as shown in Figure 1.

The corpus update involved the integration of syntactic and prosodic information into a single unified CoNLL-U format to facilitate its use. The original Rhapsodie Treebank is composed of approximately 33,000 tokens, for which the multi-level annotations were distributed in various formats.[1] These included WAV/MP3 files for the audio, TXT files for the transcription, tabular formats for micro- and macro-syntactic annotations, pitch format for acoustic analysis, and XML, TextGrid, and tabular formats for prosodic annotations. Metadata was

[1] https://rhapsodie.modyco.fr

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 111–118

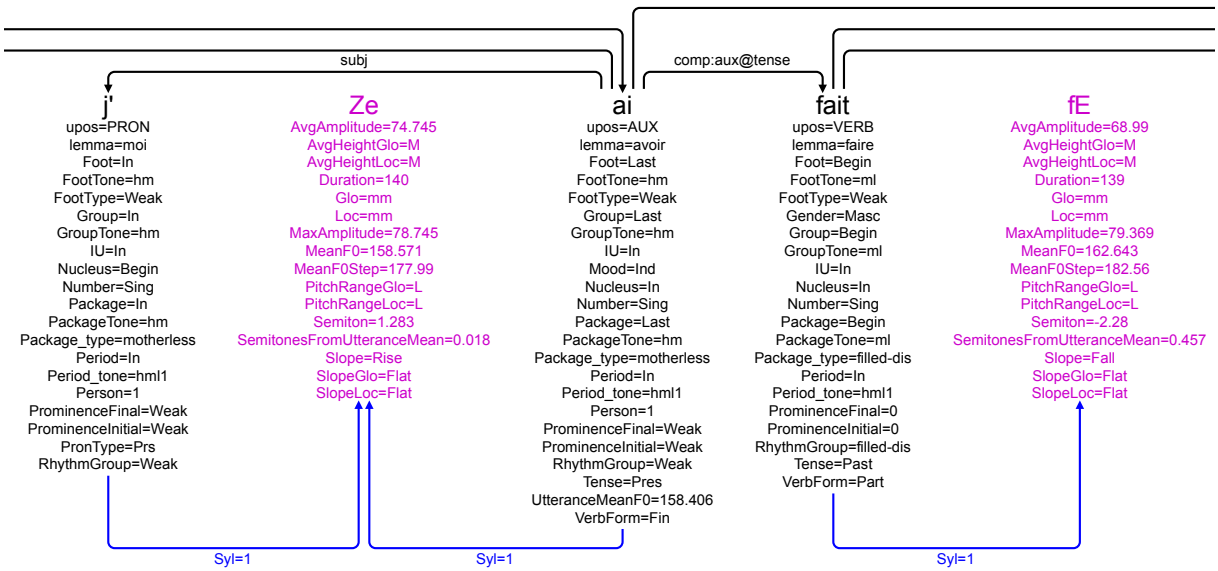

Figure 1: Example with prosodic annotations. Syllable *Ze* is fused, forming a graph structure in the sequence *et puis j'ai fait mes études au lycée, euh, de Mulhouse* ('and then I did my studies at the high school, uh, in Mulhouse'), [Rhap-D2004].

provided in both HTML and XML formats.

Since the corpus was developed over the years, with numerous reseachers involved, differences emerged in the corpus segmentations and in the alingment between each annotation. As a result, updating the corpus represented a significant challenge. The work was divided in two main steps. First, the annotations from the original Rhapsodie treebank were grouped, aligned, extracted, added and normalized. Second, automatic annotations were obtainet with the work of Strickland et al. (2023), added and normalized.

3 Integration of existing annotations

3.1 Syntax

Regarding the syntactic information provided by the original version of the corpus, in addition to a morphosyntactic analysis for word segmentation and lemmatization, the original version of Rhapsodie used its own annotation scheme, inspired by dependency syntax and the syntax of spoken corpora, in which syntactic boundaries are evaluated differently than in written corpora, relying on macrosyntax (Gerdes and Kahane, 2017).

The main difference between Rhapsodie's annotation scheme, Universal dependencies (UD) and Surface-Syntactic Universal Dependencies (SUD) is that in Rhapsodie, macrosyntax and microsyntax were annotated separately, neither of which involves prosodic criteria, although both interplay in complex ways in spoken language. Macrosyntax refers to syntactic cohesion ensured by the illocutionary act and microsyntax refers to syntactic cohesion based on government relations. The latter is encoded in a dependency tree, where a single head, which is not governed itself, projects governability onto the other lexemes (Lacheret-Dujour et al., 2019b).

The samples of the corpus are macrosyntactically segmented into groups of syntactic constituents, major syntactic units, that perform the same illocutionary act, called Illocutionary Units (IUs). Illocutionary acts include assertion, induction, interrogation, and exclamation. We can say that they perform the same illocutionary act because they can be placed under the scope of a verb that makes explicit the force of the illocutionary unit. In the first example, each IU boundary is marked with a double slash ("//").

(1) *L2 donc < moi < "ben" { je vais | {*
L2 so < me < "well" { I'm going | {
je | je } prends le mét~| je prends le
I | I } take the met~| I take the
métro } le matin "bon" jusqu'au Palais
metro } in the morning "okay" up to the Palais
Royal //+ L1 à quelle heure //
Royal //+ L1 at what time //
"excusez-moi" // [Rhap_D0001]
"excuse me" //

From a microsyntactic perspective, a Government Unit (GU) consists of all the lexemes that form the dependency graph. Even though *je prends*

le métro le matin bon jusqu'au Palais Royal and *à quelle heure* are two different IUs and two different turns of speech, they belong to the same GU (Kahane et al., 2019). The second illocutionary unit is governed by the first one. The dependency relationship between the two is categorized as *mod* in SUD.[2]. It can be queried on `https://universal.grew.fr/?corpus=SUD_French-Rhapsodie_db`

Based on this analysis, the project established the following annotations for macrosyntax. Each annotation describes the token's inclusion within a constituent or IU and its place, in BILU format (Begin, Inside, Last, Unique).

In the second example, the token *nous* 'we' is annotated as `IU=Begin` and *déja* 'first' as `IU=Last`.

(2) *nous* < *dans* *le* *quartier* <+ ***on n'a***
we < in the neighborhood <+ **we don't**
on n'a pas ***de lycée*** > *déjà* //
we don't have **any high schools** > first //
`[Rhap_D0004]`

Each IU has a **nucleus**, which is the autonomous constituent that makes clear what kind of act the speaker is performing. In this example, the token *on* 'we' is annotated as `Nucleus=Begin` and *lycée* 'school' as `Nucleus=Last`.

Other constituents inside the IU that cannot be the scope of a predicate without the nucleus, because they depend illocutionarily on it, are considered ad-nuclei. In Rhapsodie, they are classified as pre-nuclei (on the left of the nucleus), in-nuclei (within the nucleus), and post-nuclei (on the right of the nucleus). In the second example, *dans le quartier* 'in the neighborhood', *dans* is annotated as `Prenucleus=Begin` and *quartier* as `Prenucleus=Last`.

A **graft** is when the speaker does not find the good denomination and graft an IU where a proper noun was expected. In the third example, *je crois que c'est une ancienne caserne, je crois* 'I think it is an ancient barracks, I think' is used instead of *une ancienne caserne* 'an ancient barracks'. All tokens of the graft bear a feature IU_graft, with a BILU value, and the root of the graft (the first *crois*) has an additional feature Graft=Yes.

(3) *vous* *t~* *vous* *suivez* *la* *ligne du tram* *qui*
you t~ you follow the tram line which
passe *vers* *la* *&*
goes toward the &
[je crois que c'est une ancienne caserne "je crois" //
[I think it is an ancient barracks "I think" //]
] // `[Rhap_M0003]`
//

The label **AssociatedNucleus** appears when a GU shares distributional properties with nuclei, but carries a weak illocutionary force, as *je pense* 'I think' in the fourth example.

(4) *ça* < *c'est* *le* *problème* *de* *Paris*
that < that's the problem of Paris
"je pense" // `[Rhap_D0004]`
"I think" //

There are also differences between the Rhapsodie annotation scheme and the UD and SUD annotation schemes in the naming of dependency relations at the microsyntactic level, as shown in Figure 2. The updated version of the corpus was produced from a recent version annotated according to the SUD scheme which was developed on the basis of the Rhapsodie format, enriching it to allow for conversion into UD. The segmentation in IUs is preserved and the previously mentioned macrosyntactic annotations were added.

Rhapsodie	SUD	UD
sub	subj	nsubj, csubj
obj	comp:obj	obj, xcomp, ccomp
obl	comp:obl	obl:arg, iobj, xcomp, ccomp
obj + […//]	comp:obj + Reported=yes	ccomp + Reported=Yes
ad	mod	advmod, advcl, obl:mod
pred	comp:aux	aux (reversed)
	comp:pred	cop (reversed)
dep NOUN->ADJ	mod	amod
dep NOUN-> NOUN	udep	nmod
dep NOUN -> NUM	mod	nummod
dep NOUN->DET	det	det
dep ADP->NOUN	comp	case (reversed)
dep SCONJ-> VERB	comp	mark (reversed)
"…"	discourse	discourse
	parataxis	parataxis: discourse
"<…"	dislocated	dislocated
	vocative	vocative
	mod	advmod, advcl, obl:mod
"(…//)"	parataxis:parenth	parataxis:parenth
	parataxis:insert	parataxis:insert

Figure 2: Correspondence between the Rhapsodie, SUD, and UD annotation schemes.

3.2 Prosody

Regarding the prosodic annotation in Rhapsodie, it follows a data-driven approach which has been divided into three stages: prominence and disfluency

[2]This can be verified in this query, which shows the *mod* relation between illocutionary units in the new dependency-based version of the Rhapsodie treebanks of GREW-MATCH

annotation, segmentation into maximal prosodic units called Intonational Periods (IPEs), and intonational annotation relative to the intonation contour. (Lacheret-Dujour, 2019)

A syllable is considered **prominent** if it is perceptually more salient than its surrounding context. It can be annotated as `Weak`, `Strong`, or `0` if it is not prominent (Avanzi et al., 2019). If the syllable belongs to the filler *euh* 'uh' or exhibits features such as extra lengthening, infra-low register, or creaky voice, it is marked with `H`, indicating the presence of a **hesitation** (see Figure 3).

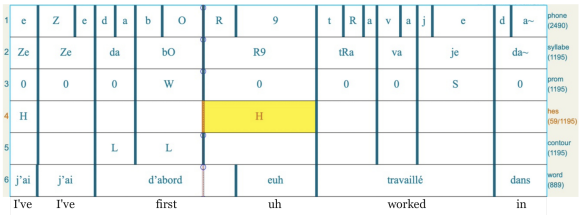

Figure 3: Original TextGrid with annotations for period, word, contour, prominence, syllable, and phonetic transcription for the sequence *j'ai j'ai d'abord euh travaillé dans* 'I've I've first uh worked in' [`Rhap_D0005`].

Segmentation into Intonative Periods (IPEs) is based on perceptual and acoustic cues. It is important to note that segmentation into IPEs does not necessarily align with that of IUs, since the IPE identification does not involve syntactic criteria.[3] It occurs when a silence of 300 milliseconds, with an absence of a filler contiguous to the pause, is detected and associated to a marked terminal contour before the pause and a melodic resetting after the pause. The detection of a speech turn is also associated to the end of a period (Lacheret-Dujour and Victorri, 2019). The token's position within an IPE is marked in BILU format too.

The period represents the root of the prosodic tree which is articulated around 3 levels of constituency from the bottom up: metrical foot, rhythmic group and intonation package as shown. Every time there is a non-disfluent but prominent syllable within an IPE, the end of a **metrical foot** (MF) is marked. In other words, a metrical foot can be composed of non-prominent syllables followed by a prominent syllable, also called the Right Head of the Foot (RHF). The prominence of the RHF determines the label of the foot as either strong or weak. A **rhythmic group** (RG) boundary is marked when a RHF (Right Head of the Foot) coincides with the final syllable of a token. When rhythmic groups occur in succession, an **intonation package** (IPA) is marked by the first group that carries a strong prominence (see Figure 4).

Sy	re	po~	sa	sHi	vRa	vE	ke	lEn	S@	va	lije
Prom	S		W	S				W	W		S
MF	re		po~ sa	sHi	vRavEkelEn				S@	valije	
RG	réponse à			suivre	avec Hélène				Chevalier		
IPA	réponse à suivre				avec Hélène Chevalier						

Figure 4: Example of a segmentation in MF, RG and IPA, where four rhytmic groups form two intonation packages in the sequence *réponse à suivre avec Hélène Chevalier* 'response to follow with HC', [Rhap-M2006]. Extracted from Lacheret-Dujour et al. (2019a).

3.3 Integration of new features

The main contribution of the treebank format introduced in Strickland et al. (2024) to ours lies in enabling quantitative studies based on acoustic analysis, both segmentally and suprasegmentally. This work used SLAM 3 (Strickland et al., 2023), the latest version of the SLAM prosodic modeling software that generates discrete labels from continuous F0 contours, which are otherwise difficult to manipulate. We also continued the Naija corpus's annotation of continuous features extracted directly from the .TextGrid and .PitchTier files of an audio, such as the duration of each syllable or `UtteranceMeanF0`, which indicates the mean F0 of each sequence in the corpus. It should be noted that our data is macrosyntactically segmented into IUs, and therefore, this annotation appears at the token that is the root of every IU.

Information related to the pitch contour is extracted from the raw pitch curves with SLAM 3, both at the global level (token) and at the local level (syllable). The pitch onset and offset of each selected segment are considered to generate a discrete label, along with its most prominent point. These discrete labels can take the values very low (L), low (l), medium (m), high (h), or very high (H). The system applies a glissando threshold formula to ensure that only pitch changes perceptible over time are taken into account (Strickland et al., 2023). It then assigns one `Glo` and one `Loc` label per syllable, corresponding respectively to `[pitch's start - pitch's end - pitch's most salient point of the contour - syllabic tier in which this prominence occurs]` like this **`[mlh1]`**. In this label, the pitch starts at a medium level (m), ends at a low level(l), with a salient high peak (h), all occurring in the first tier of the syllable (1). The

[3]This segmentation was performed using the Analor tool (Avanzi et al., 2008) and was manually verified by an expert.

full version of the annotations that were used for the Rhapsodie's new format is available in the appendix, items (27–41).

4 Use of this resource

The updated corpus is still undergoing refinement, but it is already available on GREW-MATCH.[4] For most samples in the corpus, users can visualize a dependency tree enriched with 41 new features, with full access to metadata. In cases where only part of the audio was analyzed due to overlap, these instances are marked with the "Overlap" feature, and the number of annotations is accordingly reduced. The complete list of the new features [5] added to the corpus is provided in the appendix.[6]

In GREW-MATCH, numerous queries are possible. An example[7] in the syntax-prosody domain shows that when a token is the last element of a prenucleus (`X.Prenucleus=Last`), it tends to exhibit strong final prominence (`X.ProminenceFinal=Strong`) in 65.10% of cases, as shown in Figure 5. In contrast, when the token is the first element of a prenucleus (`X.Prenucleus=Begin`), strong prominence occurs in only 23.20% of cases.[8]

Whether_1 \ X.ProminenceFinal	1291 0	880 Strong	483 Weak
2236 No	1207	589	411
449 Yes	84	291	72

Figure 5: Results of the query on the position of the IPE and final prominence.

GREW-MATCH also allows querying the newly extracted features, which contain numerical values. For example, this query [9] aims to investigate the correlation between part of speech (`X-[Syl=*]->Y; X[upos=VERB|NOUN|PRON)` and the duration of the final syllable (`Y.duration`). (see Figure 6).

We observed that the final syllables of pronouns are the shortest in 75.63% of the cases, followed by verbs at 61.96%, and nouns at 42.23%. In other words, the final syllables of nouns tend to be longer compared to those of verbs and pronouns.

X.upos \ Y.Duration[gap=200]	8228 [0, 200[	4857 [200, 400[	954 [400, 600[
5454 NOUN	2303	2553	598
4424 PRON	3346	955	123
4159 VERB	2577	1349	233

Figure 6: Results of the query on the verb's position within the IPE and its syllable's duration.

This enriched format facilitates its use in various tasks related to the syntax-prosody interface, and also opens possibilities for sociolinguistic research, even though the original version of Rhapsodie was not initially designed for this purpose.

For instance, the results of a query[10] designed to determine the percentage of IPE boundaries (`X.Period=Last`) that coincide with the end of an illocutionary unit (`X.IU=Last`), according to the social context of the sample, show that the professional social context exhibits the highest alignment between the end of an IU and the end of an IPE (68.56%), followed by the public context at 56.09%. The private context displays the lowest alignment, at 49.30%. (See Figure 7)

global.social_context \ Whether_2	1499 Yes	1261 No
1487 public	834	653
1079 private	532	547
194 professional	133	61

Figure 7: Results of the query that combines the end of an IPE, an IU and the social context.

5 Conclusion

In this paper, we discussed the process involved in updating the Rhapsodie corpus, which now forms the most feature-rich intonosyntactic treebank available (with the treebank of Naija, (Strickland et al., 2024)). It includes manual, automatic, and semi-automatic annotations, a rare achievement in current research. In the future, we believe this corpus could extend beyond the study of the syntax-prosody interface. Considering that the audio files are available, along with the information provided in the treebank, tasks such as speech modeling or classification in the prosody-syntax-sociolinguistics interface could be explored.

[4] https://universal.grew.fr/?corpus=SUD_French-Rhapsodie-prosody

[5] All information used to describe the Rahpsodie's annotations is drawn from Lacheret-Dujour et al. (2019b) and Bawden and Wang (2015).

[6] Items 1 to 26 were extracted from the original Rhapsodie version; items 27 to 41 were obtained using the tools developed for the Naija intonosyntactic treebank.

[7] universal.grew.fr/?custom=684767a3d2be8

[8] universal.grew.fr/?custom=684ab7e86ff10

[9] universal.grew.fr/?custom=68504ba9ed6b6

[10] universal.grew.fr/?custom=6847617636d58

References

Mathieu Avanzi, Guri Bordal, Anne Lacheret-Dujour, Nicolas Obin, and Julie Sauvage-Vincent. 2019. Chapter 9: The annotation of syllabic prominences and disfluencies. In Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors, *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*, pages 158–173. John Benjamins Publishing Company.

Mathieu Avanzi, Anne Lacheret-Dujour, and Bernard Victorri. 2008. Analor. a tool for semi-automatic annotation of french prosodic structure. In *ANALOR. A tool for semi-automatic annotation of French prosodic structure*, pages 119–122.

Rachel Bawden and Ilaine Wang. 2015. *Description of the Rhapsodie TreeBank's Tabular Format: Version: morpho-syntax, micro-syntax, macro-syntax, prosody*. Creation of the tabular format: Rachel Bawden, Ilaine Wang, with the collaboration of Julie Belião. Coordination: Kim Gerdes, Sylvain Kahane. Annotation Platform (Arborator): Kim Gerdes. Microsyntactic annotation: Rachel Bawden, Christophe Benzitoun, Marie-Amélie Botalla, Adèle Désoyer, Sylvain Kahane, Paola Pietrandrea. Macro-syntactic annotation: Christophe Benzitoun, Jeanne-Marie Debaisieux, José Deulofeu, Anne Dister, Florence Lefeuvre, Paola Pietrandrea, Nathalie Rossi-Gensane, Frédéric Sabio, Noalig Tanguy, Bernard Victorri. Prosody: Mathieu Avanzi, Julie Belião, Jean-Philippe Goldman, Anne Lacheret-Dujour, Philippe Martin, Nicolas Obin, Arthur Truong, Bernard Victorri.

Kim Gerdes and Sylvain Kahane. 2017. Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe. In *Actes de l'atelier « ACor4French – Les corpus annotés du français » (ACor4French 2017)*, Paris, France. ACor4French. LPP, Université Paris 3 Sorbonne Nouvelle & CNRS; Modyco, Université Paris Nanterre & CNRS.

Sylvain Kahane, Kim Gerdes, and Rachel Bawden. 2019. Chapter 4: Microsyntactic annotation. In Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors, *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*, pages 49–68. John Benjamins Publishing Company.

Anne Lacheret-Dujour. 2019. Chapter 8: Prosodic annotation of the rhapsodie corpus: Expectations and issues. In Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors, *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*, pages 147–155. John Benjamins Publishing Company.

Anne Lacheret-Dujour, Guri Bordal, and Arthur Truong. 2019a. Chapter 11: Derivation of the prosodic structure. In Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors, *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*, pages 213–231. John Benjamins Publishing Company.

Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors. 2019b. *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Anne Lacheret-Dujour and Bernard Victorri. 2019. Chapter 10: Segmentation into intonational periods. In Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea, editors, *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, volume 89 of *Studies in Corpus Linguistics*, pages 175–211. John Benjamins Publishing Company.

Emmett Strickland, Marc Evrard, and Anne Lacheret-Dujour. 2023. SLAM 3: An Updated Stylization Model for Speech Melody. In *Proceedings of the 20th International Congress of Phonetic Sciences (ICPhS2023)*, Prague, Czech Republic. Submitted on 26 Jul 2023.

Emmett Strickland, Anne Lacheret-Dujour, Sylvain Kahane, Marc Evrard, Perrine Quennehen, Bernard Caron, Francis Egbokhare, and Bruno Guillaume. 2024. New methods for exploring intonosyntax: Introducing an intonosyntactic treebank for Nigerian Pidgin. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12207–12216, Torino, Italia. ELRA and ICCL.

A Full version of the new annotations of the intonosyntactic Rhapsodie treebank

1. Layer: the token's inclusion within a layer and its position. Groups of elements that pile up in another element and have the same syntactic position are considered as lists, and the elements inside a list are considered as layers.

{il y a | il y a | il y a | il y a | il y a } *des bons établissements //=*
{‘there are | there are | there are | there are | there are } *good schools //=*
`[Rhap_D0002]`

2. Type_para: indicates the kind of relationship that links one layer in a list to another. For instance, para_disfl, para_coord, para_intens, para_dform, para_reform, para_hyper, para_negot.

je travaille à la préfecture de Paris qui { n'est pas connue |
ˆmais néanmoins existe *} "euh" //*
‘I'm working at the prefecture of Paris that { isn't well known |
ˆbut yet exists} *"uh" //’*
`[Rhap_D0001, para_coord example]`

3. Type_inherited: in an asymmetrical analysis of lists, there's a dependence with the context in

which the lists appears and and inherited dependency that is passed down to the other layers in the list, except in the case of coordination (para_coord), which is exempt from this inheritance.

4. IU: the place of a token inside an illocutionary unit, in BILOU format without the O (Begin, Inside, Last, Unique)

nous < dans le quartier <+ on n'a on n'a pas de lycée > déjà //
'we < in the neighborhood <+ we don't we don't have any high schools > first //'
`[Rhap_D0004]`

5. Nucleus: the token's inclusion within a nucleus and its place, in BILOU format. Each IU has a nucleus, which is the autonomous constituent that carries the illocutionary force of the IU.

nous < dans le quartier <+ ***on n'a on n'a pas de lycée*** *> déjà //*
'we < in the neighborhood <+ ***we don't we don't have any high schools*** *> first //'*
`[Rhap_D0004]`

6. Prenucleus: the token's inclusion within a prenucleus and its place, in BILOU format.

moi *< j'ai eu aucun problème scolaire pour mes enfants //*
'me *< I don't have any school problem for my children //'*
`[Rhap_D0002]`

7. Innucleus: the token's inclusion within an innucleus and its place, in BILOU format.

*vos journaux (****Jean-Christophe****) qui soulignent également la faiblesse de la mobilisation des électeurs >+ hier //*
*'your newspapers (****Jean-Christophe****) which also underline the poor voter turnout >+ yesterday //*
`[Rhap_D2013]`

8. Postnucleus: the token's inclusion within a postnucleus and its place, in BILOU format.

ça a duré dix ans > ***le silence autour de moi //***
'it lasted ten years > ***the silence around me//'***
`[Rhap_D2010]`

9.IU_parenthesis: the token's inclusion within a in parenthetical IU and its place, in BILOU format.

*il y a une petite rue (****+ ˆmais dont je ne sais pas le nom //****) une petite rue en & qui tourne un peu //*
*there is a small street (****+ but I don't know its name //****) a little street in & which winds a little //*
`[Rhap_M0011]`

10. IU_graft: the token's inclusion within a graft, in BILOU format. It occurs when an item appears in an unexpected position.

vous t~ vous suivez la ligne du tram qui passe vers la & [je crois que c'est une ancienne caserne ***"je crois" //] //***
'you t~ you follow the tram line which goes toward the & [I think it used to be barracks ***"I think" //] //***
`[Rhap_M0003]`

11. IU_embedded: the token's inclusion within an embedded unit that takes a governed place in another IU, in BILOU format.

*Marcel Achard écrivait (****[elle est très jolie //= elle est même belle //= elle est élégante //]****)*
*'Marcel Achard wrote (****[she is very pretty //= she is even beautiful //= she is elegant //]****)*
`[Rhap_D2001]`

12. AssociatedNucleus: the token's inclusion within an associated nucleus and its place, in BILOU format. Associated nucleui are presented as GUs that share distributional properties with nuclei but carry a weak illocutionary force.

ça < c'est le problème de Paris ***"je pense" //***
'that < that's the problem of Paris ***"I think" //***
`[Rhap_D0004]`

13. Intro_IU: The token's inclusion within an IU opener, which is an element distinct from prenuclei, is always at the beginning of an IU and is not microsyntactically dependent on another word.

et *tu arrives à la fontaine "euh" place Notre Dame //*
'and *you arrive at the fountain "erm" in Notre Dame square //*
`[Rhap_M0001]`

14. Period: the token's inclusion within an intonative period and its place, in BILOU format, without the O. Segmentation into IPEs is based on perceptual and acoustic cues. In *tu prends le boulevard euh là qui part de Nef Chavant là le boulevard qui passe à côté d'Habitat* 'you take the boulevard um there that runs from Nef Chavant the boulevard that runs past Habitat' [RhapM0001], *tu* 'you' is annotated as `Period=Begin` and *Habitat* 'Habitat' is annotated as `Period=Last`.

14. Period_tone: the contour of IPE that contains the token. The contour is considered as the first point and last point of a unit, in both points the height of the F0 in relation to the speaker average pith is labeled with five possible levels : very low (L), low (l), middle (m), high (h) and very high (H).

15. Prominence_initial: the degree to which the first syllable of a token is perceived as more salient compared to its surrounding context. This degree is annotated as Weak, Strong, or 0 if the syllable is not prominent.

16. Prominence_final: the degree to which the final syllable of a token is perceived as more salient compared to its surrounding context. This degree is annotated as Weak, Strong, or 0 if the syllable is not prominent.

17. Hesitation: particles such as 'euh' or hesi-

tant syllables, which may exhibit extra-lengthening, infra-low register, or creaky voice.

18. Foot: token's position in the metrical foot, in BILOU format without the O. Within an IPE, every time there is a non-disfluent but prominent syllable, the end of a metrical foot is marked.

19. FootTone: the contour of the last metrical foot of the token.

20. FootType: the category of the token's final metrical foot. The following annotations can be found: dis-strong, dis-weak, filled-dis, filled-pause, silent-pause, strong, tail, or weak.

21. Group: token's position inside a rhythmic group, in BILOU format without the O. A rhythmic group boundary is marked when a RHF (Right Head of the Foot) coincides with the final syllable of a token.

22. GroupTone: the contour of the group rhythmic that contains the token.

23. GroupType: the category of the rhythmic group that contains the token. It can take the same labels as the foot type annotation.

24. Package: token's position within an intonative package, in BILOU format without the O. When rhythmic groups occur in succession, a package is marked by the first group that carries a strong prominence.

25. PackageType: the category of the intonative package that contains the token. Possible annotations include: filled-dis, filled-pause, included, lone, lone-dis-strong, motherless, motherless-dis-weak, silent-pause, or tail.

26. NextBreakLength: the duration of the pause following the token.

27. AvgAmplitude: the mean amplitude of the syllable in decibels.

28. AvgHeightGlo: rough categorical average of the Glo pitch values, with possible values being L, M, and H.

29. AvgHeightLoc: rough categorical average of the Loc pitch values, with possible values being L, M, and H.

30. Duration: syllable's length in milliseconds.

31. Glo: the SLAM3 contour of a global unit, in this case, the token.

31. Loc: the SLAM3 contour of the immediate context of the target unit, in this case, the syllable.

32. MaxAmplitude: the maximum amplitude detected within the syllable in decibels.

33. MeanF0: syllable's mean F0.

34. MeanF0Step: the lowest F0 measurement which would be noticeably higher than the MeanF0, set to two semitones. This is useful for distinguishing between perceptively meaninful pitch differences in continuous data.

35. PitchRangeGlo: A categorical measurement of the pitch difference between the start and end of the `Glo` SLAM contour, with possible values being L, M, and H.

36. PitchRangeLoc: A categorical measurement of the pitch difference between the start and end of the `Loc` SLAM contour, with possible values being L, M, and H.

37. SemitonesFromUtteranceMean: number of semitones between MeanF0 and UtteranceMeanF0.

38. Slope: The slope derived from performing a linear regression

39. SlopeGlo: the slope derived from the `Glo` SLAM value, with possible values including *Rise, Fall, and Flat*.

40. SlopeLoc: the slope derived from the `Loc` SLAM value.

41. UtteranceMeanF0: the utterance's mean F0, annotated in the governing token.