

SegCor Macrosyntactic Segmentation

Version 1.1

Biagio Ursi, Nathalie Rossi-Gensane, Margot Lambert, Luisa Acosta

Introduction

The aim of this guideline is to develop a practical approach to the macrosyntactic segmentation of French oral corpora. This document relies on the theoretical principles and the analytical issues elaborated by C. Blanche-Benveniste and colleagues (1990), and the Groupe Aixois de Recherche Sur le français parlé (GARS). Subsequent syntactic descriptions of oral French have been proposed by Blanche-Benveniste (2010a, 2010b) with a simpler terminology. Through this guideline, we offer an annotation scheme for typical phenomena in spoken French.

The macrosyntactic level is a wider analytical level, which is situated beyond government and the traditional domain of syntax (i.e. the relations within the subject – verb – complement model, the relations between main clauses and subordinate ones). The notions used in this document and some segmentation choices are the result of teamwork and are grounded on previous contributions in oral corpora segmentation, namely the protocols produced within the frameworks of the Rhapsodie¹ and the ORFEO² ANR projects.


From a methodological point of view, two general criteria have been considered: substitution and paraphrase.

Segmentation and Annotation with EXMARaLDA Partitur-Editor

This guideline proposes the use of the [EXMARaLDA Partitur-Editor](#) for the segmentation and annotation of the transcripts. Clicking on *File* → *Import*, you can import various transcript files in the Partitur-Editor.


- Shared introduction (Macrosyntax + Syntax + Interactional Units) ?
- Stylesheets for the automatic generation of the tagset

You can segment the “Speaker[v]” tiers by splitting and merging the existing segments.

In order to split a segment that is too long, you click in the text where you want to split the segment. Then you either choose the split button  or the key sequence **ctrl+2**.

¹ <http://www.projet-rhapsodie.fr>

² <http://www.projet-orfeo.fr>

In order to merge several segments, you mark the segments you want to merge with the left mouse button and either choose the merge button  or the key sequence **ctrl+1**.

If you aim at an inter-annotator agreement, please be careful with the splitting and merging of segments as far as time alignment is concerned. The examples mentioned above are based on a tokenised transcript (i.e. each morpheme represents a single event). It is recommended to tokenise the transcript before annotating it, in order to limit alignment errors due to various ways of splitting segments (you can use the function “export segmented transcription” in the latest preview of the partitur-editor but you might want to check for tokenisation errors in cases of overlapping speech contributions). If you do not use a tokenised transcript, please make sure you always split **after** a space.

Annotation Tagset

<i>TAG</i>	<i>English label</i>	<i>French label</i>
MU - Macrosyntactic Units		
MU	Maximal Macrosyntactic Unit	Unité macrosyntaxique maximale
MU_dis	Discontinuous MU	UM discontinue
MU_par	Parenthetical MU	UM parenthétique
MU_col	Collaborative MU	UM collaborative
MU_ab	Abandoned MU	UM abandonnée
A-AdN - Components (ab & dis as extra information)		
N / N_dis	Nucleus / Discontinuous N	Noyau / N discontinu
PreN / PreN_ab	Pre-nucleus / Abandoned PreN	Prénoyau / Prén abandonné
PostN / PostN_ab	Post-nucleus / Abandoned PostN	Postnoyau / PostN abandonné
InN / InN_ab	In-nucleus / Abandoned InN	Innoyau / InN abandonné
Graft - Graft Phenomena		
DS	Direct Speech	Discours direct
Graft	Other cases of grafts	Autres cas de greffes
Comment - Segmentation Comments		
Series of MU	Series of MU	Série d'unités macrosyntaxiques

Segmentation problem	Segmentation problem (to be defined)	Problème de segmentation (à définir)
-----------------------------	--------------------------------------	--------------------------------------

Overview

A maximal macrosyntactic unit is an utterance constituted at least of a nucleus. The nucleus may occur with peripheral elements (prenuclei, postnuclei) or with inserted elements (innuclei), i.e. adnuclei. Nuclei, prenuclei, postnuclei and innuclei are minimal macrosyntactic units.

Minimal Macrosyntactic Units

Nuclei

A nucleus is the fundamental part of a maximal macrosyntactic unit. A maximal macrosyntactic unit (MU) contains at least a nucleus. A nucleus corresponds to a potentially isolated, autonomous utterance, which is characterised by an illocutionary force on its own (from a pragmatic perspective, a nucleus can be considered as a minimal illocutionary unit). A nucleus can namely be negated, whereas adnuclei cannot:

Ex: *j'ai mon frère il est malade*

Negation test: **je n'ai pas mon frère il est malade vs j'ai mon frère il n'est pas malade*

The nucleus is therefore *il est malade* (and *j'ai mon frère* is a prenucleus).

A maximal macrosyntactic unit can consist of a grouping, i.e. a nucleus accompanied by one or several adnuclei. Beyond macrosyntax, it is worth noting in the commentary tier certain configurations, i.e. series of nuclei that are characterised by similar syntactic constructions and/or lexicon in the same semantic field. Configurations of this kind have to figure in the commentary tier (see below for details, fig. 32).

Nuclei may have different realisations. They may be organised around pivots of different categories.

Verbal Nuclei

Verbal nuclei are organised around verbs, this kind of unit is usually called a clausal structure. In the example below (fig. 1), the intransitive verb *passe* is the pivot of the unit and is preceded by the nominal phrase *le temps*, whose syntactic function is subject.

so then time goes by

NAT[ortho] [v]	puis	après	le	temps	passe
NAT [MU]	MU				
NAT [N-AdN]	N				
NAT [Phen]					
NAT [Comment]					

Fig. 1: CLAPI_Phone-call_Copines (169-173)

Nominal Nuclei

Nominal nuclei are organised around nouns.

	2365 [05]	2366 [05]	2367 [05]	2368 [05]	2369 [05]	2370 [09:19.3]
Clai [ortho]	et	(0.3)	euh	une	autre	information
Clai [MU]	MU					
Clai [N-AdN]	N					

Fig. 2: CLAPI_Social-Meeting_MJC (2365-2370)

In the example above (fig. 2), the nominal pivot *information* is part of a minimal macrosyntactic unit, which is not coextensive with a turn of talk and projects more to come.

one hundred twenty-six
one hundred thirty

Léa[ortho][v] [v]	cent	vingt	six
Léa[MU] [v]	MU		
Léa[v] [AdN]	N		
Léa[Phen] [v]			
Léa[Comment] [v]			
Paul[v] [v]		[cent trente]	
Paul[ortho][v] [v]	cent	trente	-
Paul[MU] [v]	MU		
Paul[v] [AdN]	N		
Paul[Phen] [v]			
Paul[Comment] [v]			

Fig. 3: CLAPI_Social-Meeting_MJC (323-333)

In the example above (fig. 3), Léa and Paul produce two numeral phrases in overlap, which have been annotated as nuclei (*cent vingt-six* and *cent trente*, respectively).

In some cases, nominal pivots can be introduced by presentative constructions. In the example below, *y a* introduces the nominal phrase *une jeune apprentie*.

	no	currently	there's a young apprentice too							
spk2[ortho] [v]	non	en	ce	moment	y	a	une	jeune	apprentie	aussi
spk2 [MU]	MU									
spk2 [N-AdN]	N									
spk2 [Phen]										
spk2 [Comment]										

Fig. 4: ESLO_Interview (83-92)

Pronominal Nuclei

Pronominal nuclei are organised around pronominal pivots.

In the example below, the final element *it* represents the pronominal pivot of the minimal macrosyntactic unit, which is relevant for its microsyntactic characterization.

	that's it		
Sar[ortho][v] [v]	c'	est	ça
Sar[MU] [v]	MU		
Sar[N-Adj] [v]	N		

Fig. 5: CLAPI_Social-Meeting_MJC (2348-2350)

Adjectival Nuclei

Adjectival nuclei are organised around adjectival pivots.

	oh so funny				
	2916 [t]	2917 [t]	2918 [t]	2919 [t]	2920 [t]
ELI [ortho]	oh	trop			drôle
ELI [MU]	MU				
ELI [N-AdN]	N				

Fig. 6: CLAPI_Table-talk_Kiwi (2916-2920)

that's too lovely hm

BEA[ortho] [v]	c'	est	trop		joli		hm
BEA [MU]	MU						
BEA [N-AdN]	N						
BEA [Phen]							
BEA [Comment]							

Fig. 7: CLAPI_Table-talk_Kiwi (169-177)

In some cases, adverbial pivots can be introduced by prepositions. In the example below, the preposition *since* preceded the (modified) adjective *young*.

yeah yeah no no me since very young uhm

spk2[ortho] [v]	ouais	ouais		non		non		moi	depuis	tout	petit	hein
spk2 [MU]	MU											
spk2 [N-AdN]	N_dis1						InN	N_dis2				
spk2 [Phen]												
spk2 [Comment]												

Fig. 8: ESLO_Interview (511-523)

Adverbial Nuclei

Adverbial nuclei are organised around adverbial pivots.

exactly

MEL[ortho] [v]		exactement
MEL [MU]		MU
MEL [N-AdN]		N
MEL [Phen]		
MEL [Comment]		

Fig. 9: CLAPI_Phone-call_Copines (180-182)

absolutely

	2263 [08:53.6]
Ben [ortho]	absolument
Ben [MU]	MU
Ben [N-AdN]	N

Fig. 10: CLAPI_Social-Meeting_MJC (2263)

“Sentence Word” Nuclei

The nucleus can boil down to a sentence word (such as *ouais*, *oui*, *voilà*, *hein*, according to Tesnière 2015, see also Quirk et al. 1985 for similar issues in English). In the following excerpt (fig. 11), the little word *voilà* (“here we are”) is the pivotal element of the unit.

	er well here we are		
ELI[ortho] [v]	bon	ben	voilà
ELI [MU]	MU		
ELI [N-AdN]	N		
ELI [Phen]			
ELI [Comment]			

Fig. 11: CLAPI_Table-talk_Kiwi (57-59)

“Little words” in oral French (see Morel & Danon-Boileau 1998)

Items such as *oui*, *ouais*, *non*, *voilà*, *bon*, *ben*, *quoi*, *hein*, *enfin* are frequently used in oral French. Their segmentation can be problematic. They can act in two ways: either as nuclei or as discourse markers. The latter are attached to a larger macrosyntactic unit, i.e. are part of nuclei and of pre/postnuclei.

	well in any case			it is nice			all the same			isn't it
ELI[ortho] [v]	bon	en	tout	cas	ça	fait	plaisir	n'	empêche	quoi
ELI [MU]	MU									
ELI [N-AdN]	PreN				N			PostN		
ELI [Phen]										
ELI [Comment]										

Fig. 12: CLAPI_Table-talk_Kiwi (564-574)

In the example above (fig. 12, see also fig. 11), the first particle *bon* (“well”) is part of the prenucleus, together with the adverbial group *en tout cas* (“in any case”). Additionally, the nucleus is composed of a verbal construction (*ça fait plaisir*). A closing particle (*quoi*) is attached to the postnucleus *n’empêche* (“all the same”).

When several little words are attested in a series (e.g. *ouais ouais*, two elements in a row) and there is a pause in between, the timing of the pause represents a criterion for segmentation. If the pause is equal to or greater than 0.2 second (cf. Levinson &

Torreira 2015), the first little word constitutes a macrosyntactic unit in itself, as in the example below (fig. 13):

	yeah		yeah it's it's true					
ANA[ortho] [v]	ouais	_	ouais	c'	est	c'	est	vrai
ANA [MU]	MU		MU					
ANA [N-AdN]	N		N					
ANA [Phen]								
ANA [Comment]								

Fig. 13: CLAPI_Phone-call_CLE1 (92-99)

Similarly, in the example below (fig. 14), the initial (and only) *ouais* (“yeah”) is produced as a separate prosodic unit from *ça c’est ton job* (“that is your job”), hence two macrosyntactic units with a nucleus each.

	yeah		that it's your job				
FLO[ortho] [v]	ouais	0.287	ça	c'	est	ton	job
FLO [MU]	MU		MU				
FLO [N-AdN]	N		N				
FLO [Phen]							
FLO [Comment]							

Fig. 14: CLAPI_Preparing-meal-together_Patapizza (227-233)

Conversely, when, at the beginning of an utterance, several *ouais* (or other little words) are attested without pauses or other specific prosodic cues, they are attached to the following element.

As for the **hesitation marker** (*euh*), it is attached by default to the preceding element, except when it occurs in turn-initial position and after a long pause.

Vocalisations as Nuclei (“bruits-sons”)

Nuclei can boil down to vocalisations.

spk1[ortho] [v]	hm	hm
spk1 [MU]	MU	
spk1 [N-AdN]	N	
spk1 [Phen]		
spk1 [Comment]		

Fig. 15: ESLO_Interview (1441-1442)

AdNuclei

An adnucleus is a minimal ungoverned macrosyntactic unit depending on a nucleus, which contains a pivot on the microsyntactic level. The impossibility to apply a negative modality is due to its non-autonomous character:

- a. il est trop âgé pour le poste de directeur je pense
- b. *il est trop âgé pour le poste de directeur je ne pense pas/je pense pas

Its position can vary within a maximal macrosyntactic unit (MU): at the periphery, before (prenucleus) or after (postnucleus) the nucleus, or inside (innucleus) the nucleus.

Peripheral elements

Prenuclei

Pre-positioned complements (French *compléments antéposés*) are not considered as prenuclei, except for cases when the complement is analysed as being beyond verbal government (French *non régi*).

and in the end | the core of my talk will be a little bit on that |

spkl[ortho] [v]	et	au	fond	_	l'	essentiel	de	mon	discours	va	porter	un	petit	peu	là	dessus
spkl [MU]	MU															
spkl [N-AdN]	PreN					N										
spkl [Phen]																
spkl [Comment]																

Fig. 16: ESLO_Conference (311-327)

In the example above (fig. 16), the element at the beginning (*au fond*, “in the end”) is considered as an enunciative complement (*adverbe de commentaire énonciatif*, according to Riegel *et al.* 1994), dealing with the production of an entire utterance (and not with part of it). The connective *et* (“and”), traditionally seen as a coordinating conjunction, is the initial particle of the prenucleus. The subsequent pause is integrated to the adjunct.

Conversely, in the example below (fig. 17), the pre-positioned complement *chez mes voisins* (“at my neighbours”) is governed and therefore analysed as the first part of the nucleus (N_dis1), whose second part is introduced by the presentative *y a* (*y a peut-être dix pour cent des des cent locataires euh qui viendront euh*, “there is maybe ten percent of of the hundred tenants er who will come er”, which is annotated as N_dis2). As a governed complement, *chez mes voisins* can occur in a cleft

structure: *c'est chez mes voisins qu'y a peut-être dix pour cent des cent locataires qui viendront.*

well the at my neighbours me there is maybe ten percent of of the hundred tenants er who will

Paul[ortho] [v]	enfin	le			chez	mes	voisins	moi	y	a	peut	être	dix	pour	cent	des	des	cent	locataires	euh	qui	viendront	euh
Paul[MU] [v]	MU																						
Paul[e] [AdN]	N_dis1										InN		N_dis2										
Paul[Phen] [v]																							
Paul[Comment] [v]																							

Fig. 17: CLAPI_Social-Meeting_MJC (1207-1233)

In the example below (fig. 18), the pre-positioned complement *demain* (“tomorrow”, which is governed by the second verbal form *manger*) is part of the nucleus, and is thus analysed as the first part of a discontinuous nucleus (N_dis1). The complement can occur in a cleft structure: *c'est demain que ça te dit qu'on aille manger dehors.*

tomorrow | seeing that we have only one hour to eat er | do you fancy if we eat outside

ANA[ortho] [v]	demain	vu	que	on	a	qu'	une	heure	pour	manger	euh	ça	te	dit	que	on	aille	manger	dehors
ANA [MU]																			
ANA [N-AdN]	N_dis1		InN										N_dis2						
ANA [Phen]	DS																		
ANA [Comment]																			

Fig.18: CLAPI_Phone-call_CLE1 (28-47)

In the following example (fig. 19), the same complement (*demain*), which is produced at the end of an MU (*si tu veux on avise demain*), is part of the nucleus:

well listen if you want we decide tomorrow

FLO[ortho] [v]	ben	écoute	si	tu	veux	on	avise	demain
FLO [MU]	MU							
FLO [N-AdN]	PreN1			N				
FLO [Phen]								
FLO [Comment]								

Fig. 19: CLAPI_Phone-call_CLE1 (406-413)

When a series of prenuclei is annotated, each segment is numbered progressively: preN1, preN2, etc.

Postnuclei

Postnuclei are adjuncts to the nuclei and are right-positioned in an utterance. They are non-governed elements as in the example below (fig. 20):

oh no but I gonna I gonna tell you this several times I think

FLO[ortho] [v]	ah	non	mais	je	vais	je	vais	te	la	ressortir	plusieurs	fois	je	pense
FLO [MU]	MU													
FLO [N-AdN]	N											PostN		
FLO [Phen]														
FLO [Comment]														

Fig. 20: CLAPI_Preparing-meal-together_Patapizza (274-290)

The discourse marker *je pense* (“I think”) is annotated as a postnucleus because it can also occur inside a nucleus, as an innucleus (see below): *ah non mais je vais je vais je pense te la ressortir plusieurs fois*.

oh yeah he said

Sar[ortho][v] [v]	ah	ouais	dit	il
Sar[MU] [v]	MU			
Sar[v] [AdN]	N		PostN	
Sar[Phen] [v]	DS			
Sar[Comment] [v]				

Fig. 21: CLAPI_Social-meeting_MJC (2071-2074)

In the example above (fig. 21), the postposed reporting verb (*dit-il*, “he said”) is a postnucleus following a nucleus (*ah ouais*, “oh yeah”). The macrosyntactic unit is thus composed of a nucleus and a right-positioned element.

Inserted elements

Innuclei

We consider innuclei as macrosyntactic sub-units that occur in the middle of a larger macrosyntactic unit.

Frequently, innuclei are inserted verbal constructions (see the notion of “propositional discourse marker”, French *marqueur discursif propositionnel*, cf. Andersen 2007).

The following list (provided by the ORFEO annotation guidelines) presents the most common verbal innuclei in French. Alternative forms are in parentheses and a rough translation is proposed (see appendix for detailed examples):

mettons, disons, on va dire, si on peut dire: “let’s say”

allez: “come on”

je crois, je pense: “I think, I guess”

je sais pas: “I don’t know”

je dirais: “I would say”

je veux dire: “I mean”

tu sais (vous savez): “you know”

tu vois (vous voyez): “you see”

tu me dis (tu m'as dit, vous me dites, vous m'avez dit): “you say, you said”

il me semble: “it seems to me”

si tu veux (si vous voulez): “if you want”

comment dire (comment dirais-je): “how can I say, how can I put it”

More generally, verbal constructions of this sort can be considered as adnuclei, insofar as they occur not only in the middle of a nucleus, but also at the beginning as prenuclei or at the end as postnuclei.

Discourse markers and connectives in initial position (*ligateurs*, according to Morel & Danon-Boileau 1998, such as *donc, bon, eh bien*, etc.) are not considered as prenuclei. They are therefore not annotated as innuclei when they occur in a middle position.

the the wine in the end what is it in from an er historical point of view eh |

spk1[ortho] [v]	euh	le	le	vin	au	fond	c'	est	quoi	au	d'	un	point	de	vue	_	euh	historique
spk1 [MU]	MU																	
spk1 [N-AdN]	N_dis1				InN				N_dis2									
spk1 [Phen]																		
spk1 [Comment]																		

Fig. 22: ESLO_Conference (1365-1385)

In the excerpt above (fig. 22), the adverbial group *au fond* (“in the end”) is annotated as an innucleus and is inserted within a larger nucleus, which consists of two discontinuous nuclei (N_dis1 and N_dis 2). Disfluencies at the beginning (*le le vin*, “the the wine”) and in the middle (*au d’un point de vue*, “in from a point of view”) are part of the nucleus, as well as hesitations (*euh*) and the final particle (*hein*).

Maximal Macrosyntactic Units

A maximal macrosyntactic unit is composed at least by a nucleus, which can be extended by one or several adnuclei. For example, a prenucleus, a nucleus and a postnucleus, as shown in the excerpt below (fig. 23):

	well in any case				it is nice			all the same			isn't it
ELI[ortho] [v]	bon	en	tout	cas	ça	fait	plaisir	n'	empêche	quoi	
ELI [MU]	MU										
ELI [N-AdN]	PreN				N			PostN			

Fig. 23: CLAPI_Table-talk_Kiwi (564-574)

A macrosyntactic unit is at (the very) least constituted of a nucleus, which can be a verbal construction and even a noun phrase.

Discontinuous Macrosyntactic Unit

A discontinuous macrosyntactic unit is a unit that is momentarily suspended by another maximal macrosyntactic unit, namely a parenthetical MU. The discontinuous unit is completed after the realisation of the parenthetical unit.

Parenthetical Macrosyntactic Unit

A parenthetical unit is an insertion within a larger macrosyntactic unit. Parenthetical units can occur in series. In this case, each parenthetical unit corresponds to an utterance (i.e. a nucleus).

	I had		but now it's over				it will also be for after the summer holiday a debate on nutrition																
Paul[ortho] [v]	j'	ai	_	j'	avais	mais	maintenant	c'	est	terminé	ça	sera	aussi	pour	la	rentrée	_	un	débat	sur	la	nutrition	
Paul[MU] [v]	MU_dis1				MU_par1			MU_par2				MU_dis2											
Paul[v] [AdN]	N_dis1				N			N				N_dis2											
Paul[Phen] [v]																							
Paul[Comment] [v]																							

Fig. 24: CLAPI_Social-Meeting_MJC (1388-1411)

In this excerpt (fig. 24), the main verb (*j'avais*, "I had") is separated from the object (*un débat sur la nutrition*, "a debate on nutrition"). The two in-between parenthetical sequences are constituted of two different nuclei (*mais maintenant c'est terminé*, "but now it's over"; *ça sera aussi pour la rentrée*, "it will also be for after the summer holiday"). Within a series, each parenthetical sequence is numbered progressively: MU_par1, MU_par2, etc.

Note that segmentation relies exclusively on syntactic criteria (and not on prosodic cues). The elements within the parenthetical sequence(s) are not governed by the surrounding elements. Parenthetical sequences are verbal constructions possessing an illocutionary force (as opposed to innuclei, verbal constructions not possessing an illocutionary force: see above).

Collaborative Macrosyntactic Unit

Collaborative configurations can be the result of syntactic co-constructions, when a first speaker produces a macrosyntactic unit (abandoned or not) and a second speaker produces its syntactic completion.

	well from	oh no	you	we met		in the underground yeah					
BEA[ortho] [v]	ah	non	toi	on	s'	est	croisées	_			
BEA [MU]	MU_coll										
BEA [N-AdN]	PreN			N							
BEA [Phen]											
BEA [Comment]											
MAR[ortho] [v]								dans	le	méto	ouais
MAR [MU]								MU_coll2			
MAR [N-AdN]								N			
MAR [Phen]											
MAR [Comment]											

Fig. 25: CLAPI_Table-talk_Kiwi (637-647)

In the excerpt above (fig. 25), Béatrice produces a first MU (*ben par*, which is abandoned and syntactically incomplete), then she utters a complex MU (*ah non toi on s'est croisées*, “oh no you we met”), which is eventually completed by Marion (*dans le méto ouais*, “in the underground yeah”). The last two units represent a collaborative configuration and are labelled as MU_coll1 and MU_coll2, respectively. In this case, it is interesting to note that MU_coll1 (*ah non toi on s'est croisées*) is a potentially complete syntactic construction, which is composed by a prenucleus (*toi*, “you”, plus the initial particles *ah non* that are produced continuously) and a nucleus (*on s'est croisées* is a well-built (macro)syntactic unit). MU_coll2 can be considered as a delayed realisation of a complement (*epexegeisis*, here produced by another speaker). *Ouais* in final position is attached to the nucleus (once again, the lack of a pause in between and the continuous prosody are segmentation criteria).

it's like you know being on come dine with me you know
 come dine with me

ELI[ortho] [v]	on	se	croirait	tu	sais	dans	un	dîner	presque	parfait	tu	sais
ELI [MU]	MU_col1											
ELI [N-AdN]	N_dis1			InN			N_dis2			PostN		
ELI [Phen]												
ELI [Comment]												
BEA [v]	[dans un dîner _]							pres[que parfait _]				[]
BEA[ortho] [v]							dîner	presque	parfait	-		
BEA [MU]	MU_col2											

Fig. 26: CLAPI_Table-talk_Kiwi (525-541)

Abandoned Macrosyntactic Unit

An abandoned macrosyntactic unit is an unfinished construction from a syntactic point of view (fig. 27).

oh well anyway from the moment er | it's funny |

ELI[ortho] [v]	oh	ben	de	toute	façon	à	partir	du	moment	où	euh	-	c'	est	rigolo
ELI [MU]	MU_ab														
ELI [N-AdN]	N_ab														
ELI [Phen]															
ELI [Comment]															

Fig. 27: CLAPI_Table-talk_Kiwi (393-413)

Here, Elise produces a first utterance “well anyway from the moment er” (*ben de toute façon à partir du moment où euh*), which is unfinished. This first utterance is made up of one prenucleus and an abandoned nucleus (as *à partir du moment où* is considered as governed: see below) where the governing element is nonetheless absent. After a long pause, the subsequent unit consists only of a nucleus (*c’est rigolo*, “it’s funny”).

Graft phenomena

A graft (Deulofeu 2010) is a verbal construction produced, in the place of another category, in a governed position.

i) Direct Speech

What constitutes a direct speech unit could normally act as a macrosyntactic unit, in a non-reporting speech context. Yet, direct speech and the introductory sequence, which consists of a *verbum dicendi*, form a macrosyntactic unit (only) together. In other words, direct speech is analysed as part of the MU. It figures as a graft (in particular, a direct speech sequence, DS in abbreviated form) in a third tier (called Phen). The resulting configuration illustrates a specific phenomenon (for example called “clause complex” by M.A.K. Halliday 1994).

look I wanted to ask you
tomorrow seeing that we have only one hour to eat er do you fancy if we eat outside

ANA[ortho] [v]	dis	je	voulais	te	demander	_	demain	vu	que	on	a	qu'	une	heure	pour	manger	euh	ça	te	dit	que	on	aille	manger	dehors
ANA [MU]	MU																								
ANA [N-AdN]	PreN	N					N_dis1	InN											N_dis2						
ANA [Phen]							DS																		
ANA [Comment]																									

Fig. 28: CLAPI_Phone-Call_Cle1 (22-47)

Here (fig. 28), the *verbum dicendi demander* (“to ask”) introduces a direct speech sequence (*demain vu que on a qu'une heure pour manger euh ça te dit que on aille manger dehors*), which consists of a discontinuous nucleus in two parts (see above for a detailed analysis of this excerpt) and an innucleus. Note that the *verbum dicendi* is part of the first nucleus *je voulais te demander* (“I wanted to ask you”), which is preceded by the verbal particle *dis* (translated by “look”, literally: “tell”).

Direct speech sequences can be quite complex as well as the preceding structures, which contain *verba dicendi*.

make a living my father had told us
guys as you wish but if you are not careful you will end up in a factory

spk2[ortho] [v]	gagner	sa	vie	mon	père	nous	avait	dit	les	gars	comme	vous	voulez	_	mais	_	si	vous	faites	pas	attention	vous	vous	retrouvez	à	l'	usine	
spk2 [MU]																												
spk2 [N-AdN]	PreN				N					PreN																		
spk2 [Phen]										DS																		
spk2 [Comment]	floating segment																											
spk2 [Syntax]	P		D(D(F))	D(F)	D(A)	A	D(P)	P		D(P)	P	R	A	P			DM		D(P)prop			A	D(P)a	P		R	D(D(P)h)	D(P)h
spk2 [Syntax]																		R	A	P	negP	D(F)						

Fig. 29: ESLO_Interview (1789-1820)

In the example above (fig. 29), the first nucleus *mon père nous avait dit* (“my father had told us”) is preceded by a hanging topic, i.e. a prenucleus, which is an infinitive phrase (*gagner sa vie*, “make a living”). The direct speech sequence is then made up of the noun phrase *les gars* (guys”), which represents a prenucleus and projects the nucleus *comme vous voulez* (“as you wish”). Another nucleus, introduced by the connective *mais* (“but”), is present at the end of the embedded sequence (note that there is no prenucleus as this kind of *if*-construction is considered as governed according to Sabio 2013): *mais si vous faites pas attention vous vous retrouvez à l’usine* (“but if you are not careful you will end up in a factory”). Again, the whole configuration represents a maximal macrosyntactic unit.

ii) Other cases

In cases other than direct speech, a governed syntactic element may not be realised through the expected component and a verbal construction occurs in the place of noun phrases, prepositional phrases, etc., thus representing a macrosyntactic insertion at the interface with microsyntax (Deulofeu 2010).

	because it is a software					that costs I don't know how much					they they will not lend them for training							
ELI[ortho]	parce	que	c'	est	un	logiciel	qui	vaut	je	sais	pas	combien	ils	vont	pas	leur	prêter	pour
ELI[MU]	MU																	
ELI[N-Adj]	N																	
ELI[Phen]											Graft							

Fig. 30: CLAPI_Table-talk_Kiwi (1928-1954)

In the example above (fig. 30), Elise produces a macrosyntactic unit and a graft is annotated in the phenomena tier. More specifically, the cleft structure *c’est un logiciel qui vaut je ne sais pas combien* (“it is a software that costs I don’t know how much”) contains a graft sequence, which occupies a syntactic slot that would have been reserved to the price of the software, normally expressed through a noun phrase (e.g. *c’est un logiciel qui vaut dix mille euros*, “it is a software that costs ten thousand euros”).

Series of Nuclei within a Macrosyntactic Unit

Syntactic configurations with two introductory elements related to each other are annotated within the same macrosyntactic unit. This is the case for elements like *soit... soit...* (“either... or...”, see fig. 31) or *plus... plus...* (such as in the example: *plus il mange plus il grossit*, “the more he eats the more he puts on weight”). These macrosyntactic configurations have been studied as “conjoined structures” (in French *constructions siamoises*, see Blanche-Benveniste *et al.* 1990, Savelli 1993).

er in fact it’s either one thousand euros or there is another solution

Clai[ortho][v] [v]	euh	en	fait	c'	est	soit	mille	euros	soit	y	a	aussi	une	autre	solution
Clai[MU] [v]	MU														
Clai[v] [AdN]	N1								N2						
Clai[Phen] [v]															
Clai[Comment] [v]															

Fig.31: CLAPI_Social-meeting_MJC (2234-2248)

In this case, though the former *soit* does not occur in the initial position of the first nucleus, it links this first nucleus to the second one, where another *soit* occurs. This is the criterion for the grouping of several nuclei within the same macrosyntactic unit. Every single nucleus is numbered in a progressive way (N1, N2... Nn).

BEYOND MACROSYNTAX

Series of Macro syntactic Units

When there is a series of nuclei, i.e. verbal or nominal constructions that are not governed, every single nucleus is considered as a macrosyntactic unit, as is the case in the example below (fig. 32):

I have verbal interaction on Mon- on Monday afternoon on Tuesday i have no classes
on Wednesday I have er what do I have I have syntax

ELI[ortho] [V]	j'	ai	interaction	verbale	le		lun-	le	lundi	après	midi	.	mardi	j'	ai	pas	cours	,	le	mercredi	j'	ai	euh	,	qu'	est	ce	que	j'	ai	j'	ai	syntaxe	bon
ELI [MU]	MU											MU	MU_dis1					MU_par	MU_dis2															
ELI [N-AdjN]	N											N	N_dis1					N	N_dis2															
ELI [Phen]																																		
ELI [Comment]	Series of MU																						interruption of the series					Continuation of the series						

Fig. 32: CLAPI_Table-talk_Kiwi (928-961)

Here, there are several nuclei, which correspond to several macrosyntactic units: a series of MUs. The nuclei are characterised by a similar syntactic construction and lexical items belonging to the same semantic field, [day of the week]_{Time Adjunct} + *j'ai* + [X]_{Direct Object}, with the exception of one abandoned nucleus (without a direct object) and the nucleus *qu'est ce que j'ai* (English “what do I have”).

Commentary tier

A supplementary tier is introduced in order to take into account cases beyond macrosyntax and cases for segmentation discussion.

References

- Andersen H. L. (2007). Marqueurs discursifs propositionnels, *Langue française* 154(2), 13-28.
- Blanche-Benveniste C. *et al.* (1990). *Le français parlé. Études grammaticales*, Paris: CNRS éditions.
- Blanche-Benveniste C. (2010a). *Approches de la langue parlée en français* (second edition). Paris: Ophrys.
- Blanche-Benveniste C. (2010b). *Le français : usages de la langue parlée* (with the collaboration of Philippe Martin). Louvain/Paris: Peeters.
- Deulofeu J. (2010). La greffe d'un énoncé sur une construction : une combinaison originale de parataxe et de rection. In M.-J. Béguelin & G. Corminboeuf (eds), *La Parataxe*. Berne: Peter Lang, 175-208.
- Guimier C. (1996). *Les adverbes du français*. Paris/Gap: Ophrys.
- Halliday M. A. K. (1994). *An Introduction to Functional Grammar*. London: Edward Arnold.
- Levinson S. C. & Torreira F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6: 731.
- Morel M.-A. & Danon-Boileau L. (1998). *Grammaire de l'intonation. L'exemple du français*. Paris/Gap: Ophrys.
- Quirk R., Greenbaum S., Leech G., Svartvik J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Riegel M., Pellat J.-C. & Rioul R. (1994). *Grammaire méthodique du français*. Paris: Presses Universitaires de France.
- Sabio F. (2013). Les séquences en SI dans les corpus oraux. In J.-M. Debaisieux (ed.) *Analyses linguistiques sur corpus. Subordination et insubordination en français*. Paris: Lavoisier, 317-362.
- Savelli M.-J. (1993). *Contribution à l'analyse macro-syntaxique : les constructions « siamoises » de type « plus V1 plus V2 »*. Thèse soutenue le 13 septembre 1993 à l'Université de Provence, sous la direction de C. Blanche-Benveniste.
- Tesnière L. (2015). *Elements of Structural Syntax* (translated by T. Osborne & S. Kahane), Amsterdam/Philadelphia: John Benjamins [original French edition: Tesnière L. (1966). *Éléments de syntaxe structurale* (édition revue et corrigée). Paris : Klincksieck].

APPENDIX

List of innuclei

Here below, we present a list of the most common discourse markers that can occur as innuclei. The following examples have been provided by the ORFEO annotation guidelines. Innuclei are in red:

mettons

demain à partir (**mettons**) en fait de treize heures s'il vous plaît

disons | dites | dis

on pourrait hum (**disons**) attendre le le conducteur

allez

on va lui dire (**allez**) d'ici euh une quinzaine de minutes

je crois

c'était (**je crois**) sur la ligne cinq

je pense

elle a été verbalisée (**je pense**) dans un bus

je sais pas

il a eu en fait euh (**je sais pas**) apparemment une indemnité forfaitaire

je dirais

essayez (**je dirais**) peut-être en début d'après-midi ça sera mieux

je veux dire

c'est tellement plus simple (**je veux dire**) de bon ben de prendre le client

tu sais | vous savez

une paire de lunettes un peu (**vous savez**) un peu comme rectangulaires

tu vois | vous voyez

c'était à peu près au niveau (**vous voyez**) de Saint Germain l'Auxerrois

tu me dis | tu m'as dit | vous me dites | vous m'avez dit

et vous êtes descendue (**vous me dites**) au niveau du centre commercial

on va dire

vous allez utiliser la le passe (**on va dire**) de manière régulière ou occasionnelle

il me semble

je suis rentrée dans le bus (**il me semble**) sept cent trente-cinq

si tu veux | si vous voulez

j'ai un petit problème euh concernant euh (**si vous voulez**) un ticket que j'ai introduit

si on peut dire

mercredi c'était férié ben ce matin le trafic (**si on peut dire**) normal

comment dire | comment dirais-je

il y a (**comment dire**) une marque euh quelque chose

arrête | arrêtez

arrête c'est pas vrai

attends | attendez

à vrai dire

bien entendu

écoute | écoutez

n'empêche

regarde | regardez

remarque | remarquez

tiens

tu parles