

ANNE LACHERET-DUJOUR, PAOLA PIETRANDREA

Rhapsodie: Un treebank prosodico-sintattico per il francese parlato

Il corpus di francese parlato *Rhapsodie* presenta produzioni di 89 soggetti, uomini e donne, per una grande varietà di generi discorsivi ($\pm$ spontaneo, $\pm$ pianificato, dialoghi faccia a faccia, interviste, trasmissioni radiofoniche e televisive). Il corpus è composto da 57 estratti brevi (lunghi in media 5 minuti ciascuno) per un totale di 3 ore di parlato trascritto ortograficamente (33000 parole) e fonologicamente, allineato al suono al livello di fonema, sillaba, parola, turno, pausa e sovrapposizione. L'annotazione sintattica e quella prosodica sono organizzate ciascuna in più livelli. In sintassi, oltre ad un'annotazione delle dipendenze sono annotate le liste e la struttura macro-sintattica. In prosodia sono annotate le unità prosodiche maggiori (periodi intonativi), le prominenze, le disfluenze, le unità prosodiche interne al periodo e i contorni intonativi.

Parole chiave: interfaccia prosodia-sintassi, prominenze, disfluenze, liste, francese.

1. Introduzione

Il treebank *Rhapsodie* è un corpus di francese parlato annotato sintatticamente e prosodicamente con l'obiettivo di modellizzare il ruolo svolto dall'intonazione e la sintassi nella segmentazione del discorso in unità elementari e condurre studi funzionali sulla struttura prosodico-sintattica del francese in diversi generi discorsivi. Per quanto riguarda la sintassi, combinando il lavoro iniziato con il corpus C-Oral-Rom (Cresti & Moneglia 2005) con la proposta di analisi elaborata a Aix en Provence da Blanche-Benveniste *et al.* (1990), si propongono un'annotazione morfosintattica, microsintattica e macrosintattica. Per quanto riguarda la prosodia, *Rhapsodie* costituisce il primo esempio di corpus annotato contemporaneamente al livello delle frasi prosodiche derivate dalle prominenze accentuali e al livello dei profili intonativi degli enunciati che lo compongono. Per quanto riguarda l'interfaccia prosodia-sintassi, l'allineamento temporale delle unità prosodiche e sintattiche permette di esplorare simultaneamente que-

sti due livelli di annotazione, studiare il loro allineamento, che è variabile nel francese parlato, e proporre un'interpretazione funzionale di queste variazioni (Lacheret-Dujour *et al.* 2019).

2. *Struttura del corpus e metadati*

Il corpus *Rhapsodie* è stato costituito per verificare l'ipotesi di una correlazione tra le caratteristiche prosodico-sintattiche di un discorso e le caratteristiche della situazione in cui è prodotto, cioè il suo genere¹ (v. tabella 1).

Tabella 1 - *Variabili linguistiche, situazione e genere del discorso: una relazione causale da modellizzare nella lingua parlata*

Caratteristiche della situazione

→ Caratteristiche prosodico-sintattiche

→ Genere discorsivo

Il corpus *Rhapsodie* è organizzato in 56 brevi testi (di 5 minuti in media ciascuno) rappresentativi di vari generi discorsivi, prodotti da un numero di parlanti sufficientemente ampio da evitare idiosincrasie individuali. I dati sono stati tratti da corpora già esistenti (Durand *et al.* 2009; Eshkol-Taravella *et al.* 2011; Branca-Rosoff *et al.* 2012), e completati da corpora multimediali raccolti specificamente per *Rhapsodie*.

Per strutturare i metadati sono selezionate sei caratteristiche situazionali maggiori. Innanzitutto, si distinguono monologhi, prodotti da un solo parlante, e dialoghi, prodotti da almeno due parlanti, in situazioni più o meno interattive. Si distinguono poi discorsi privati (faccia a faccia) e discorsi pubblici (conferenze, trasmissioni televisive o radiofoniche). Gli argomenti e i compiti (racconti personali, lezioni, interviste, sermoni, allocuzioni...) affrontati nei vari estratti sono molto vari e per quanto riguarda la parola pubblica, gli estratti rappresentano esempi di parlato su canali diversi: interviste politiche, dibattiti, trasmissioni di divulgazione scientifica, ecc... Infine, ogni testo è classificato secondo cinque variabili: (i) livello di pianificazione;

¹ Cfr. la nozione di «registro» di Biber & Conrad (2009), secondo i quali un registro è caratterizzato da una serie di tratti lessicali e grammaticali ricorrenti tipici di una varietà linguistica e associati a determinate funzioni comunicative generali.

(ii) grado di interattività; (iii) compito svolto dai parlanti; (iv) canale di comunicazione; (v) tipo di sequenza discorsiva maggiormente rappresentato nel campione. Per codificare e interrogare i metadati di *Rhapsodie* è stato utilizzato il formato IMDI-CMDI, sviluppato al Max Planck di Nimega (CMDI, <http://www.clarin.eu/cmdi>, Broeder *et al.* 2012). La tabella 2 presenta sinteticamente i metadati che possono essere interrogati nel corpus.

Tabella 2 - *Variabili situazionali nel corpus Rhapsodie*

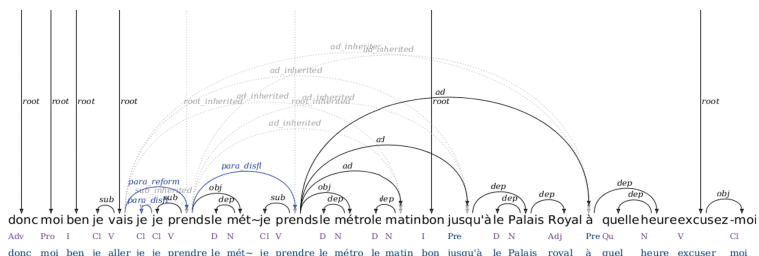
Tipo di discorso	Privato/Pubblico	Monologo
		Dialogo
<Pianificazione>		Spontaneo, semi-spontaneo, pianificato
<Interattività>		Interattivo, semi-interattivo, non interattivo
<Canale di comunicazione>		Faccia a faccia vs. Conferenza, trasmissione radiofonica, trasmissione televisiva
<Tipo di sequenza>		Argomentativa, descrittiva, procedurale, oratoria

3. L'annotazione sintattica

L'annotazione sintattica del corpus *Rhapsodie* consiste in un'annotazione morfosintattica, un'annotazione microsintattica e un'annotazione macrosintattica.

L'annotazione morfosintattica è piuttosto classica: si identificano le parole, cioè token dotati di una valenza teorica; si assegnano dei lemmi a questi token; e per ogni lemma si precisa la parte del discorso.

L'annotazione microsintattica permette di precisare la struttura di dipendenza dei dati, ma anche le strutture di lista (figura 1). Per lista intendiamo quel dispositivo sintattico per il quale due o più parole occupano la stessa posizione in un albero di dipendenza (Blanche-Benveniste 1990; Kahane & Pietrandrea 2012; Masini *et al.* 2018). È importante notare che molti fenomeni tipici del parlato come le riformulazioni, correzioni, richieste di chiarimento, di conferma, che vengono solitamente considerate come interruzioni della coesione sintattica, sono in realtà per la gran parte realizzate attraverso liste sintattiche, cioè enunciando parole che si trovano in una posizione sintattica già aperta nel discorso (1): questa caratteristica permette di assicurare nel parlato un certo livello di coesione sintattica.

Figura 1 - L'annotazione morfosintattica e microsintattica in *Rhapsodie*

- (1) *alors < { j'a~ | j'avais } beaucoup | j'avais beaucoup } trop peur de m'installer (comme ça) seule { d~ | dans } la brousse //* [Rhap-D2004, Rhapsodie]

'allora < { io a~ | io avevo } davvero | io avevo davvero } trop-pa paura di andare a vivere (così) sola { ne~ | nella } foresta //

L'annotazione macrosintattica del corpus Rhapsodie combina il modello di annotazione macrosintattica proposto a Firenze (Cresti & Moneglia 2005) con il modello di Aix en Provence sviluppato da Blanche Benveniste *et al.* (1990). Gli estratti sono innanzitutto segmentati in Unità Illocutorie (UI). All'interno di ogni UI è identificato un nucleo, cioè una sequenza portatrice della forza illocutoria dell'intera UI ed eventualmente altre unità accessorie prive di forza illocutoria che possono precedere, interrompere o seguire il nucleo (dette rispettivamente pre-nuclei, in-nuclei, post-nuclei). Sono inoltre annotati i marcatori del discorso. In (2) e (4), i pre-nuclei sono delimitati dal simbolo <, in (3) il post-nucleo è preceduto dal simbolo >, in (4) il marcatore del discorso è indicato tra virgolette ". Le UI sono delimitate dal simbolo //.

- (2) *alors < là < la psychiatrie < c'est autre chose //* [Rhap-D0006, CFPP2000]
allora < da questo punto di vista < la psichiatria < quella è un'altra cosa //
- (3) *ça a duré dix ans > le silence autour de moi //* [Rhap-D2001, Mertens corpus]
è durato dieci anni > il silenzio intorno a me
- (4) *donc < moi < "ben" { je vais | { je|je } prends le mét~ | je prends } le métro le matin "bon" jusqu'au Palais Royal //*
quindi < io < "insomma" { vado | { io|io } prendo la met~ | prendo } la metro di mattina "insomma" fino a Palais Royal //

4. L'annotazione prosodica

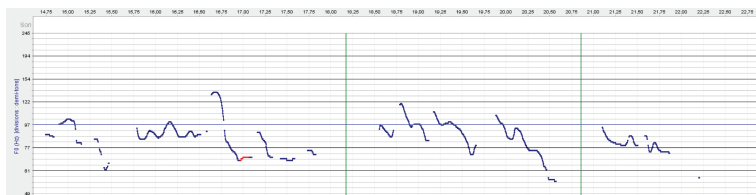
L'annotazione prosodica è stata condotta su due piani: sono stati identificati le frasi melodiche e i confini tra i segmenti, e sono stati poi identificati i profili interni alle unità. L'annotazione prosodica *Rhapsodie* ha tre specificità: (1) è un'annotazione bottom-up percettiva ('t Hart *et al.* 1990; Wightman 2002). E questo ha per conseguenza che le unità generate dal modello non sono necessariamente corrispondenti alle unità funzionali. Così il periodo intonativo, identificato da un confine terminale, non si allinea necessariamente alla clausola o all'atto linguistico. (2) Sono annotate tutte le disfluenze e sovrapposizioni. (3) Il dominio di proiezione dei contorni intonativi non è fissato a priori, ma definito dal ricercatore in funzione del suo obiettivo d'osservazione.

Ogni enunciato (o turno di parola per gli estratti dialogici) è segmentato automaticamente con il software Analor (Lacheret & Victorri 2002) in una successione di periodi intonativi (PI) sulla base di criteri essenzialmente acustici, (tavola 3, figura 2).

Tavola 3 - Criteri di segmentazione in periodi intonativi

Identificazione di una pausa di almeno 300 ms
Identificazione di un movimento di F0 ampio: il tratto $[\pm \text{ampio}]$ è fissato in funzione dell'intervallo melodico, misurato in semitoni, tra l'ultimo estremo di F0 prima della pausa e la media di F0 su tutta la porzione che precede la pausa.
Identificazione di una reinizializzazione melodica dopo la pausa.

Figura 2 - Segmentazione in 3 PI dell'enunciato *je pense aux nombreuses victimes de la tempête p_i et à toute leur famille endeuillé e_{p_i} dont nous partageons la peine p_i* 'penso alle numerose vittime della tempesta e a tutte le loro famiglie in lutto, di cui condividiamo il dolore' [D2004, Corpus *Rhapsodie*, tratto da un discorso presidenziale di Nicolas Sarkozy]



La *prominenza*, cioè la proprietà di una sillaba di distaccarsi percettivamente dal suo contesto fonetico (Terken 1991; Wagner *et al.* 2015), è un indice cruciale che il ricevente usa per processare la prosodia on-line. L'annotazione manuale è stata effettuata con PRAAT (Boersma & Weenink 2010) da cinque annotatori non esperti e rivista da tre esperti. Sono stati distinti tre gradi di *prominenza*: le sillabe non prominenti (annotate con '0'), le sillabe debolmente prominenti (annotate con 'W' per *weak*), e le sillabe fortemente prominenti (annotate con 'S' per *strong*).

Una sillaba può essere al tempo stesso prominente e disfluente nel qual caso è annotata con (H'), e le marche di disfluenza possono essere variabili e combinabili: allungamento, pausa d'esitazione, ripetizioni, eccetera (Tabella 4).

Tabella 4 - *Etichettatura delle prominenze nell'enunciato c'était assez assez terrible et les les maisons ont brûlé 'era davvero terribile e le case sono bruciate' [Rhap-D0003, Corpus PFC]; con dall'alto in basso: il livello delle sillabe, il livello delle prominenze, il livello delle disfluenze*

se	tE	a	se	a	se	te	Ri	ble
0	0	W	S	0	W	0	S	0
		H	H					
le	le	mE	zo~	o~	bRy	le		
H								

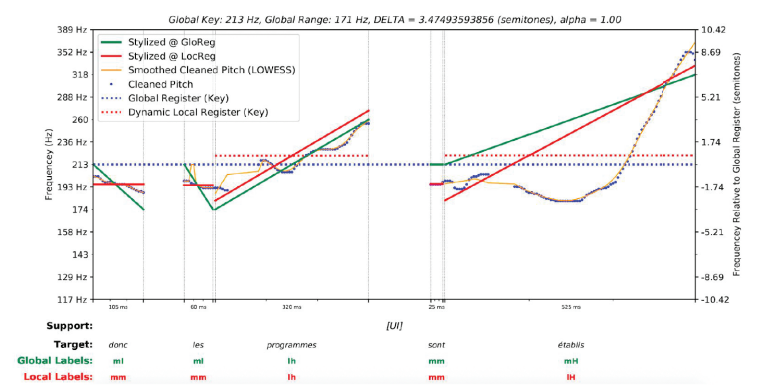
A partire dalle prominenze identificate (Prom), vengono applicate 3 regole che permettono di derivare la struttura prosodica di base (figura 3): (1) ogni sillaba prominente di un periodo marca il confine destro e la testa di un piede metrico (PM); (2) la testa di un piede metrico che si allinea con il confine destro di una parola ortografica marca la fine di un gruppo ritmico (GR); (3) la prima testa di un PM associata a una *prominenza* ('S') indica la frontiera destra di un gruppo intonativo (GI).

Figura 3 - *Generazione dell'albero prosodico del sintagma
le sherpa du président 'lo sherpa del presidente'*

Syl	l@	SER	pa	dy	pRe	zi	da~
Prom			W		W		S
PM		l@SERpa			dypRe		zida~
GR		le sherpa				du président	
GI			le sherpa du président				

Il tool SLAM+ è stato sviluppato per generare i contorni intonativi stilizzati di qualunque unità del corpus Rhapsodie (sintattica o prosodica). Il tool codifica tre informazioni principali (Yu Chen *et al.* 2019): i livelli intonativi attraversati da un contorno (da sopra-acuto a infra-grave), la direzione del contorno (ascendente o discendente), la sua forma (concava o convessa). Un insieme di etichette tonali permette di dare conto dei punti chiave d’un contorno (figura 4).

Figura 4 - *Generazione dei contorni delle parole e etichette tonali associati,
estratto di [Rhap-M1001]*



4. Conclusioni

Il corpus Rhapsodie è scaricabile liberamente secondo i termini della licenza Creative Commons Attribution – Non commercial – Share Alike 3.0 France sul sito <https://rhapsodie.modyco.fr/>. Le registrazioni (wav/mp3), le analisi acustiche (F0 grezze, pulite manualmente e F0 stilizzate automaticamente, formato pitch), la trascrizione ortografica (txt), l’annotazione macrosintattica (txt e formato tabulare),

l'annotazione microsintattica (formato tabulare), l'annotazione e la segmentazione prosodica (textgrid, formato tabulare), e i metadati (xml, html) sono accessibili sia sotto forma di archivi sia campione per campione. I metadati possono essere interrogati *on line*. Le prime analisi funzionali del corpus (strutture micro e macrosintattiche, tratti prosodici, unità intonosintattiche maggiori) sono presentati in Lacheret *et al.* 2019.

Riferimenti bibliografici

- Biber, Douglas & Conrad, Susan. 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.
- Blanche-Benveniste, Claire & Bilger, Mireille & Rouget, Christine & Van den Eynde, Karel. 1990. *Le français parlé – études grammaticales*. Paris: Editions du Centre National de la Recherche Scientifique.
- Boersma, Paul & Weenink, David. 2010. *Praat: doing phonetics by computer (Version 5.3)*. www.praat.org.
- Branca-Rosoff, Sonia & Fleury, Serge & Lefevre, Florence & Pires, Matthew. 2012. *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*. <<http://cfpp200.univ-paris3.fr>>
- Broeder, Daan & van Uytvanck, Dieter & Gavrilidou, Maria & Trippel, Thorsten & Windhouwer, Menzo. 2012. Standardizing a component metadata infrastructure. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1387-1390.
- Cresti, Emanuela & Moneglia, Massimo (eds). 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.
- Durand, Jacques & Laks, Bernard & Lyche, Chantal. 2009. Le projet PFC (phonologie du français contemporain) : une source de données primaires structurées. In Durand, Jacques & Laks, Bernard & Lyche, Chantal (a cura di), *Phonologie, variation et accents du français*, 19-61. Paris: Hermès.
- Eshkol-Taravella, Iris & Baude, Oliver & Maurel, Denis & Hriba, Linda & Dugua, Céline & Tellier, Isabelle. 2011. Un grand corpus oral «disponible»: le corpus d'Orléans 1968-2012. *Ressources linguistiques libres TAL* 52(3): 17-46.
- Kahane, Sylvain & Pietrandrea, Paola. 2012. Types d'entassement en français. In *Actes du Congrès Mondial de Linguistique Française (CMLF 2012)*, 1809-1828.

- Lacheret, Anne & Victorri, Bernard. 2002. La période intonative comme unité d'analyse pour l'étude du français parlé: modélisation prosodique et enjeux linguistiques. *Verbum*, 24/1-2, 55-73.
- Lacheret-Dujour, Anne & Kahane, Sylvain & Pietrandrea, Paola. 2019. *Rhapsodie, a prosodic and syntactic treebank of spoken French*. Amsterdam: John Benjamins.
- Masini Francesca & Mauri Caterina & Pietrandrea Paola. 2018. List constructions: Towards a unified account. *Italian Journal of Linguistics*, 30(1), 49-94.
- Terken, J. 1991. Fundamental Frequency and Perceived Prominence. *Journal of the Acoustical Society of America*, 89, 1768-1776.
- † Hart, J., Collier, R. & Cohen, A. 1990. Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody. Cambridge: Cambridge University Press.
- Wagner, Petra & Origlia, Antonio & Avesani, Cinzia & Christodoulides, George & Cutugno, Francesco & D'Imperio, Mariapaola & Escudero, David & Gili Fivela, Barbara & Lacheret, Anne & Ludusan, Bogdan & Moniz, Helena & Ní Chasaide, Ailbhe & Niebuhr, Oliver & Rousier-vercruyssen, Lucie & Simon, Anne-Catherine & Simko, Juraj & Tesser, Fabio & Vainio, Martti. 2015. Different parts of the same elephant : a roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 0202.1-5.
- Wightman, Colin W. 2002. Tobi Or Not Tobi?. In *Proceedings of Speech Prosody*, Aix-en-Provence, France, <http://sprosig.isle.illinois.edu/sp2002/>.
- Yu Chen, Liu & Lacheret-Dujour, Anne & Obin, Nicolas. 2019. Automatic Modelling and Labelling of Speech Prosody: What's new with SLAM+. In Calhoun, Sasha & Escudero, Paola & Tabain, Marija & Warren, Paul (a cura di), *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019. <https://assta.org/proceedings/ICPhS2019/>.