

Designing and implementing a multilayer annotation system for (dis)fluency features in learner and native corpora

Amandine Dumont

Université catholique de Louvain
Centre for English Corpus Linguistics
amandine.dumont@uclouvain.be

1 Introduction

The notions of fluency and disfluency have mainly been approached from two angles. The holistic approach defines fluency as the "smooth, rapid, effortless use of language" (Crystal 1987: 421); in this perspective, fluency is considered as a global phenomenon of language production (e.g. Chambers 1997; Lennon 1990). By contrast, the componential approach (Fox Tree 1995; Gut 2009; Hieke 1981, 1985) sees fluency as an isolatable dimension of language proficiency resulting from the conjunction of a series of quantifiable and qualifiable phenomena, such as filled pauses, discourse markers, false starts or restarts. Many componential studies are, however, limited to the research of one of those features without considering its interaction with other fluency features (Chafe 1980; Pallaud et al. 2013; Raupach et al. 1980), and few have considered variation in speech performances. The researcher's choice for one or the other of those two approaches obviously results in a different set of fluency features and in diverging methodologies.

Against this backdrop, Götz' (2013a) study is remarkable for combining both perspectives into what she calls an "integrated" approach: she examined a comprehensive set of fluency variables with the aim of delineating the overall fluency profiles of nonnative speakers of German. Following this new line of thinking, this paper aims to present an innovative multilayer annotation system for (dis)fluency features in learner and native speech. This system allows for the investigation of a large number of (dis)fluency features (either independently or in interaction) with a view to drawing the bigger picture of (dis)fluency behavior of nonnative and native speakers. The system has been created within a wider research project on fluency across languages and modalities¹ for which a more general framework of (dis)fluency annotation has been developed (Crible et al. 2014).

2 Design

Several theoretical and practical principles have underpinned the design of the (dis)fluency annotation system.

1. The main hypothesis of the research project out of which this annotation system has arisen is that fluency and disfluency are the two sides of the same coin. In other words, the same feature can be used as a means to enhance fluency at one point, and as a marker of disfluency at another, and it is in the recurrence and combination of those features that fluency or disfluency can be established. Consequently, the tagging system makes no a priori decision as to which elements should be considered as fluent or disfluent: all occurrences of a feature are tagged in the same way.
2. The integrated approach to (dis)fluency, i.e. (dis)fluency seen as a variety of features contributing to a holistic phenomenon, constitutes the second cornerstone of the system. For this purpose, the protocol offers a tagging system for a dozen distinct (dis)fluency features (see Table 1). It allows for the annotation of (dis)fluency features involving one (e.g. a pause) or several words (typically a repetition), and, conversely, for the annotation of words included in more than one (dis)fluency feature (e.g. a vowel lengthening within a repetition). The annotation system remains essentially linear, and (dis)fluency features are annotated at the level of the word. In addition to this componential dimension, the system makes it possible to draw a holistic picture of each individual speaker's fluency behavior.
3. Thirdly, the system is designed for and on the basis of spoken data: contrarily to some other annotation systems which have been developed with standard written text in mind (see Rehbein et al. 2012), this protocol is solely based on concepts of spoken language such as filled pause, self-correction, aborted utterance and the use of "written" concepts such as "sentence" is avoided. Preliminary versions of the annotation scheme were iteratively tested on a corpus sample and amended accordingly to reach the final version.
4. Last but not least, the (dis)fluency annotation system is aimed to be applied to large corpora, to different speaking tasks, and to both learner and native data. This implies that

¹ Université catholique de Louvain & Université de Namur (ARC grant 12/17-044).

the system must not only be grounded on well-defined (dis)fluency categories, but it also has to be flexible, straightforward, applicable to different data types and reasonably quick to implement.

3 Implementation

This multilayered (dis)fluency annotation system has been implemented within the EXMARaLDA tool (Schmidt et al. 2009) to the time-aligned version of the French component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI, Gilquin et al. 2010). This large database contains recordings and transcripts of interviews of advanced learners of English from 11 mother tongue backgrounds (50 interviews [c. 20 min. each] per L1, each interview consisting of three speaking tasks)². Although the released version of LINDSEI transcriptions contains inline annotations of several features of spoken language (including (dis)fluency phenomena), these are insufficient for (dis)fluency analyses answering the principles outlined above. Table 1 illustrates the added value such annotation system can provide to spoken corpora.

LINDSEI (inline annotations)	(DIS)FLUENCY ANNOTATION SYSTEM	
	(Dis)fluency feature	Examples (FR009, FR010 & FR011 ³)
Empty pause (perceptive transcription)	Unfilled pause (in ms; 3 sub- categories)	<i>I'd been (0.720) planning to to go</i>
Filled pause	Filled pause	<i>something to do with er politics</i>
Truncated word	Truncated word (3 sub- categories)	<i>wh when I was a little girl</i>
Foreign word	Foreign word	<i>politics or (0.820) relations internationales</i>
Lengthening	Lengthening	<i>in a (0.280) well in a real (0.750) town</i>
/	False start	<i>I don't think it's well for me er I wouldn't do it</i>
/	Repetition	<i>it's just for (0.490) for me</i>

² A comparable corpus of interviews of native speakers of English, the *Louvain Corpus of Native English Conversation* (LOCNEC, De Cock 2004) is currently being time-aligned and annotated for (dis)fluency features so as to provide a proper native benchmark.

³ Each interview in LINDSEI is identified by a specific code: "FR" corresponds to the interviewees' mother-tongue (here

		<i>it's it's meant for students I think</i>
/	Restart (5 sub- categories)	<i>last week (0.190) er last year he plays with the (0.530) he plays volleyball with the Lux</i>
/	Connector	<i>I like skiing too but it's a bit too far to go you have erm (0.570) facilities and you can get into contact with people</i>
/	Discourse marker	<i>well I (0.200) wanted to to go there I like eh making er swimming for (0.360) non stop you know</i>
/	Editing term	<i>tennis table (0.330) ta table tennis sorry</i>

Table 1. LINDSEI mark-up vs. (dis)fluency annotation system

Each feature has a corresponding tag in the form of one or two letters, e.g. FP for filled pause and T for truncated word. For repetitions, a numbering system is used to show the number of repetitions and the number of repeated words. A set of symbols is also integrated to indicate the onset ("<") and offset (">") of each feature as well as multiple tagging on one item, if any ("+"; e.g. the word "enfin" in *it was erm enfin we hadn't* [FR005], which is tagged both as a discourse marker and as a foreign word). Those tags are spread into successive layers of annotation, corresponding to different levels of precision in the characterization of (dis)fluency features, from the more generic to the more in-depth. The following examples⁴ illustrate the annotation system.

1. FR017

er	(1.590)	you	we	have	to
<FP>	<UP>		<RS>		<L>
	<P>		<SP>		

study	erm	mathematics	er	to	start
	<FP>		<FP>		

2. FR027

I	(0.220)	well	(0.690)	I	won	't
<R0	<UP>	<DM>	<UP>	R1>		
	<N>+<S>	<N>	<N>+<P>			

French), and the three-figure number (001 to 050) refers to the 50 learners.

⁴ 1st tier: FP: filled pause; UP: unfilled pause; RS: restart; L: lengthening; Rn: repetition; DM: discourse marker; T: truncated word. 2nd tier: S: short UP; P: long UP; SP: propositional substitution; N: nesting.

say	I	(0.120)	only	listen	to	classical	music
		<UP>					
		<S>					

3. FR011

a	girl	sitting	in	a	(2.030)	yes
			<R0	R0	<UP>	<DM>
					<N>+<P>	<N>

in	a	(0.230)	o	on	a	chair
R1	R1>	<UP>	<RS+<T	RS+T>	RS>	
		<S>	<SP	SP>		

In order to highlight the potential of this type of annotation, I adopt a contrastive approach to test the hypothesis that learners and native speakers differ in their quantitative use of both (dis)fluency features and (dis)fluency patterns/clusters (cf. Aijmer 1997). On the basis of the annotated LINDSEI-FR and its native speaker counterpart LOCNEC, and of the features outlined above, (dis)fluency profiles of 30 learners as compared to 30 native speakers are presented in an integrated approach to (dis)fluency. Preliminary data reveal that advanced French learners of English are a much less homogeneous group than native speakers as the extent of inter-speaker variation is far greater (about twice as large), particularly in the use of filled and unfilled pauses. The paper then shifts to the qualitative use of those two types of pauses and examines how native and nonnative speakers compare with respect to (dis)fluency clusters around those pauses. This study offers an L1/L2 counterpart to Degand & Gilquin's (2013) recent research on the environment of pauses in L1 English and French and contributes to the recent line of studies into variation between learners (from the same or different mother tongue, or with different language proficiency levels) on the one hand, and between native and nonnative speakers on the other (e.g. Gilquin & Granger 2011; Gilquin & De Cock 2011; Götz 2013 a and b).

References

- Aijmer, Karin. 1997. "I Think" - an English Modal Particle". In *Modality in Germanic Languages. Historical and Comparative Perspectives*, eds. Toril Swan & Olaf J. Westvik, 1-47. Berlin: Mouton de Gruyter.
- Chafe, Wallace. 1980. "Some Reasons for Hesitating". In *Temporal Variables in Speech*, eds. Raupach, Manfred et al, 168-80. Den Haag: Mouton de Gruyter.
- Chambers, Francine. 1997. "What Do We Mean by Fluency?". *System* 25, n° 4: 535-44.
- Crible, Ludivine, Dumont, Amandine, Grosman, Iulia, & Notarrigo Ingrid. (2014). "Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et

multimodaux, natifs et non natifs". Unpublished internal report. Université catholique de Louvain: Louvain-la-Neuve.

- Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.
- De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, New Series 2, 225-246.
- Degand, Liesbeth, & Gaëtanelle Gilquin. 2013. "The clustering of 'fluencemes' in French and English". 7th International Contrastive Linguistics Conference (ICLC 7) - 3rd conference on Using Corpora in Contrastive and Translation Studies (UCCTS 3) (Ghent, 11/07/2013 - 13/07/2013).
- Fox Tree, Jean E. 1995. "The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech". *Journal of Memory and Language* 34: 709-38.
- Gilquin, Gaëtanelle, & Sylvie De Cock. 2011. "Errors and Disfluencies in Spoken Corpora: Setting the Scene". *International Journal of Corpus Linguistics* 16, n° 2: 141-72.
- Gilquin, Gaëtanelle, Sylvie De Cock, & Sylviane Granger, eds. 2010. *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses Universitaires de Louvain. Louvain-la-Neuve.
- Gilquin, Gaëtanelle, & Sylviane Granger. 2011. "The Use of Discourse Markers in Corpora of Native and Learner Speech: From Aggregate to Individual Data". *Corpus Linguistics conference (Birmingham, 20/07/2011 - 22/07/2011)*.
- Götz, Sandra. 2013a. *Fluency in Native and Nonnative English Speech*. Studies in Corpus Linguistics (SCL) 53. Amsterdam & Philadelphia: John Benjamins Publishing Company.
- . 2013b. "How Fluent Are Advanced German Learners of English (perceived to be)? Corpus Findings vs. Native-Speaker Perception". In *Studies in Variation, Contacts and Change in English*, eds. Magnus Huber & Joybrato Mukherjee, Vol. 13. Giessen: University of Giessen.
- Gut, Ulrike. 2009. *Non-Native Speech: A Corpus-Based Analysis of Phonological and Phonetic Properties of L2 English and German*, eds. Thomas Kohnen & Joybrato Mukherjee. English Corpus Linguistics 9. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.
- Hieke, Adolf E. 1981. "Audio-Lectal Practice and Fluency Acquisition". *Foreign Language Annals* 14, n° 3: 189-94.
- Hieke, Adolf E. 1985. "A Componential Approach to Oral Fluency Evaluation". *The Modern Language Journal* 69 (2): 135-42.
- Lennon, Paul. (1990). "Investigating Fluency in EFL: A Quantitative Approach*". *Language Learning* 40, n° 3:

387-417.

- Pallaud, Bertille, Stéphane Rauzy, & Philippe Blache. 2013. "Auto-interruptions et disfluences en français parlé dans quatre corpus du CID". *TIPA: Travaux interdisciplinaires sur la parole et le langage*, n° 29.
- Raupach, Manfred, Hans-Wilhelm Dechert, & Frieda Goldman-Eisler, eds. 1980. *Temporal Variables in Speech*. Janua Linguarum 86. The Hague: Mouton.
- Rehbein, Ines, Sören Schalowski, & Heike Wiese. 2012. "Annotating Spoken Language". In *Best Practices for Speech Corpora in Linguistic Research*, 29. Istanbul, Turkey.
- Schmidt, Thomas, & Kai Wörner. 2009. "EXMARaLDA - Creating, Analyzing and Sharing Spoken Language Corpora for Pragmatics Research". *Pragmatics - Quarterly Publication of the International Pragmatics Association* 19 (4): 565. www.exmaralda.org.