Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 89–92.

Repeats in spontaneous spoken French: the influence of the complexity of phrases

Sandrine Henry

Équipe DELIC, Université de Provence, Aix-en-Provence, France

Abstract

We here present the results of a descriptive study we conducted on 383 disfluent repeats from a corpus of spontaneous spoken French. We analyze noun phrases under construction and study whether there is a co-relation between the frequency of the repeats and the complexity feature of the phrases. We then focus on complex noun phrases in order to locate precisely the repeats. We also analyze how repeats affect structures such as [Preposition + Determiner + Noun] and what the constraints upon such structures are.

1. Introduction

Disfluency phenomena (such as repeats, word fragments, self-repairs, etc.) can be found in all spontaneous oral productions. Indeed, oral speech, as opposed to finalized writing which is a deferred production, is produced online and therefore has the specificity of retaining the traces of its elaboration. It is by fits and starts and later syntactic and/or lexical readjustments that oral spontaneous speech is elaborated. It is never delivered in a smooth fashion which could be compared to edited writing, that is to say a revised, corrected and perfect form! We here study repeats such as "malheureusement + c'est comme toujours on est obligé **de de** continuer **le : ++ le** voyage".

In the past years, researches on spoken French have thrived, in fields such as syntactic studies [3, 4], prosody [5], psycholinguistics [7, 11], computational aspects or human-computer dialogue [2]. Thus, a certain number of regular features have been identified in repeats: on the morpho-syntactic level, repeats mostly involve function words (9 repeats out of 10) which, most of the time, are monosyllables [9] as 41.5% are determiners, 26% pronouns and 13% prepositions [10]. These function words partake of the structuring of language and shape content words into syntactic units. Like Blanche-Benveniste [3], we have been able to check that repeats are subjected to syntactic constraints: they mainly appear at the beginning of phrases and their structure remains stable, that is to say the simple syntactic frame – without any lexical content – appears first, and the lexical filling comes second.

We have also established in collaboration with Campione & Véronis [9] that repeats present a variable degree of "disruptivity" according to the number of other disfluencies (silent and/or filled pauses) that are combined with them at the Interruption Point: when there is only one disfluency at the IP, it is the lengthening; when they are two disfluencies, the most frequent case is a combination of a lengthening and a silent pause.

We here focus on the realization of the repeat in NPs. We first give a distribution of the repeats according to the type of chunk involved. We will then analyze how the repeats are distributed according to the complexity feature of the phrase. Is it possible to co-relate the presence of a repeat with the complexity feature? We will also determine where the repeat is locating in the complex phrase: do repeats tend to appear more often in the head zone or in the expansion zone? Does the presence of a preposition in the noun chunk have an influence on the location of the repeat? Finally, among the prepositional noun chunks that have a [Preposition + (predeterminer) + Determiner + (modifier) + Noun + (modifier)] pattern, we will give an account of the most frequent types of repeats.

Our work is based on a corpus of spontaneous[1] speech of 17,000 words. It consists of Campione's corpus [5] (54 min, 8,500 words) to which we added, whilst preserving the original sampling of the corpus, 10 other extracts in order to obtain a corpus of 1h 47min. Most of the recordings are from the *CRFP* (French Reference Corpus)[2] and our corpus is composed of 20 speakers: 10 men and 10 women. The average length of the extracts is about 5 minutes, and the passages we selected are monologues where the speaker talks about his job, or evokes past events, etc. The speaker there answers questions from an investigator who does not intervene in the selected pieces.

2. Repeats and types of chunks

We have labelled disfluent repeats and found 383 occurrences in our corpus.

If, traditionally, the phrase, "a sequence of words composing a syntactic unit" [13], is considered to be the intermediate unit between the word and the sentence, we nevertheless remark that, in some cases, this unit can again be broken up into smaller units that are not words but chunks [1].

There are 4 types of chunks: noun (NC), verb (VC), adverb (AdvC) and adjective (AdjC) chunks. The distribution of repeats according to the type of chunk is as follows:

Table 1: Distribution of the repeats according to the type of chunks.

	NCs	VCs	AdvCs	AdjCs	Other	Σ
Repeats	263	91	8	5	16	383
%	68.7	23.8	2.1	1.3	4.2	100

We note that noun chunks obviously prevail (more than 2 repeats out of 3). This is linked to the strong involvement of determiners and pronouns in repeats. We have not found any example where the repeat would be on a content word only, such cases do exist, but they are not disfluent repeats.

Less than a fourth of the repeats occur in verb chunks and only a little over 3% in adverb and adjective chunks.

If we take a closer look at these results, we remark that approximately 1 repeat out of 3 (30.5%, 117/383) takes place in a chunk introduced by a preposition. When a chunk is introduced by a preposition, in 82% of cases it is a noun chunk, in 17% a verb chunk, and in only 1% of cases an adjective chunk.

Noun chunks, whether introduced or not by a preposition, are a privileged observation zone for repeats, and that explains why we propose a detailed study of these chunks.

[1] "Not based on a written piece, not learned by heart", Candéa [6].
[2] *Corpus de Référence du Français Parlé*, for more information, *cf.* [8].

3. Repeats in noun phrases

3.1. Definition of the complexity feature and results

We here adopt Clark & Wasow's definition of the complexity feature [7]: "NPs, however, range in complexity. *The mangy dog*, for example, is slightly more complex than *the dog* because of the added modifier, but *the dog down the street* and *the dog my neighbor owns* are much more complex because of the prepositional and clausal modifiers *after* the head noun. To simplify complexity, we divided NPs into *simple NPs*, which don't have anything after the head noun, and *complex NPs*, which do" (p.211).

Repeats can occur in phrases that present no expansion of the head noun. The noun phrase (NP) is then called "simple":

ex. 1: il faut être extrêmement vigilant car [**la : + la** loi]$_{NP}$ est euh ++ est précise là-dessus

When the NP is expanded, it is called "complex":

ex. 2: elle sort [**deux : deux** boudins qui étaient pleins de sciure]$_{NP}$

Among the 223 occurrences of repeats in NPs, 57% (126/223) take place in complex NPs, and 43% (97/223) in simple NPs. If these results indicate that repeats occur more frequently in complex NPs, only relative frequencies would however allow us to conclude that the complexity feature does influence the distribution of the repeats. We do not have such figures at the moment.

We have therefore restricted our analysis to "test words" which are the more frequent determiners in French, that is to say *le*, *la*, *les*, *un*, *une*, *des*. For each of them, we have identified, according to the complexity feature of the phrase (simple *vs* complex), the number of repeats *vs* the absence of repeats. When we divided the number of occurrences of repeats by the total number of occurrences (repeated + non repeated ones) for each determiner, we obtained the following relative frequencies:

Table 2: Relative frequencies of repeats of determiners *le, la, les, un, une, des* according to the complexity feature of the phrase.

	le	*la*	*les*	*un*	*une*	*des*
Simple NPs	5.4%	4.2%	5.2%	4.0%	2.2%	7.3%
Complex NPs	14.6%	7.4%	8.8%	6.9%	6.5%	11.8%

First at all, we can remark that for each of the six test words we retained for our study the relative frequencies of repeats are systematically higher in complex NPs than in simple NPs. The repeat rate is on average of 4.7% in simple NPs and reaches 9.3% in complex NPs.

In order to measure accurately the influence of the complexity feature on the presence of repeats, we have calculated the ratios of the relative frequencies of repeats in test words:

Table 3: Ratios of frequencies of repeats for each test word.

	le	*la*	*les*	*un*	*une*	*des*
Ratios	2.7	1.8	1.7	1.7	2.9	1.6

The table above shows that, according to the determiner involved in the repeat, the ratios change drastically, from 1.6 to 2.9. *Le* and *une* have the most important ratios (respectively 2.7 and 2.9); these elements are approximately 3 times more likely to be repeated in a complex NP than in a simple NP. The other determiners (*la*, *les*, *un*, *des*) have smaller ratios, between 1.6 and 1.8. The mean of the ratios amounts to 2.1 and it allows us to affirm that determiners are on average twice more likely to be repeated in complex NPs than in simple NPs.

Clark & Wasow [7] also observe a correlation between the complexity of the phrase and the frequency of the repeat for the determiners *the* and *a* located at the beginning of a noun phrase. So far, we have not taken into account the location of the determiner in the phrase and our results apply to determiners in the head chunk (containing the head noun) as well to those in the expansion chunk (containing the expansion of this head).

Why does the complexity of the phrase influence the occurrence of repeats? The most obvious and logical explanation is that the speaker has a lot more to plan in a complex phrase. Actually, he not only has to manage the structuring of the head chunk but also plan ahead the syntactic arrangement of the expansion chunk. The local constraint of a lexical search – the repeat as delaying tactics from the speaker – would be a lighter burden than the global structuring of the phrase. If it proves true, we should logically observe more repeats in the head zone than in the expansion zone. In order to check this hypothesis, we have located all the repeats in the phrases using our 6 test words.

3.2. Locating the repeat in complex noun phrases

In our corpus, determiners can be repeated:

- in the zone which contains the head noun:

  ex. 3: et en fait [(**les : les** personnes)$_{HEAD\ CHUNK}$ (d'un certain âge)$_{EXPANSION\ CHUNK}$]$_{NP}$ aiment toujours danser

- or in the zone which contains the expansion of the noun:

  ex. 4: il y a euh [(des fiches)$_{HEAD\ CHUNK}$ (sur **la la** faune)$_{EXPANSION\ CHUNK}$]$_{NP}$

As we expected, the determiners we selected as test words are mainly repeated in head chunks (83% of cases). The head chunk is therefore a privileged site for repeats of determiners and this proves that planning the whole of the noun phrase is a major constraint on the presence of repeats.

Moreover, the analysis of the complex noun phrases containing at least one repeat allows us to bring to light the following patterns:

- repeats in the head chunk:

  ex. 5: on a par exemple [(**ces + ces** fameux oeufs)$_{HEAD\ CHUNK}$ (en meurette)$_{EXPANSION\ CHUNK}$]$_{NP}$ je sais pas si vous savez

- repeats in the expansion chunk:

  ex. 6: alors là c'est c'est [(des soirées dansantes)$_{HEAD\ CHUNK}$ (**qui : qui** sont ouvertes à tout le monde)$_{EXPANSION\ CHUNK}$]$_{NP}$

- repeats in both the head chunk and the expansion chunk:

  ex. 7: j'ai eu très sincèrement l'impression que [(**ce : + ce** jour)$_{HEAD\ CHUNK}$ (**de : de** mon mariage)$_{EXPANSION\ CHUNK}$]$_{NP}$ a été le plus beau jour de ma vie

The following figure presents the distribution of complex noun phrases according to the location of the repeat:

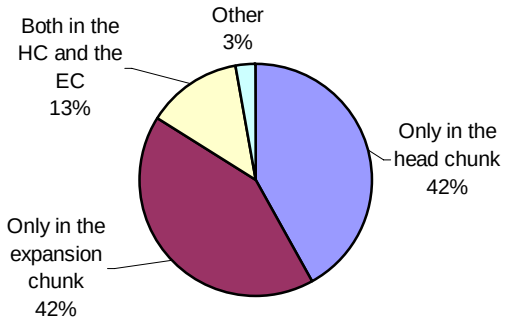

Figure 1: Distribution of complex NPs according to the location of the repeat.

Among the 112 noun phrases in which repeats occur, we notice that 13% of complex noun phrases contain repeats both in the head chunk and the expansion chunk. Repeating at the beginning of the phrase is therefore not enough to compensate for the difficulties the speaker encounters when planning the whole of the phrase. The constraint of a lexical search must not be neglected.

In addition to that, we can see that in 42% of cases repeats affect only the head chunk. The same proportion is to be found in the expansion chunk. This result seems to run counter to our previous conclusions on repeats of determiners. One would actually expect complex noun phrases to contain more repeats in the head than in the expansion zone, but it is not so. Why then?

A possible explanation would be as follows: contrary to the head chunk which, of course, contains the lexical head of the phrase and thus nearly systematically begins with a determiner (of any kind), the beginning of the expansion chunk can be composed of various elements, such as a preposition or a relative pronoun. The following figure shows the significant presence of prepositions in expansion chunks:

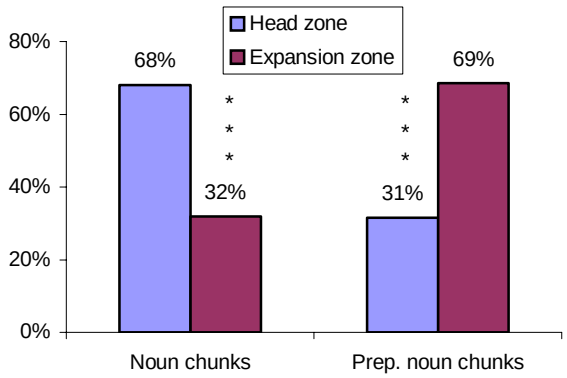

Figure 2: Comparison of the location of repeats according to the type of chunk.

When the complex noun phrase contains a determiner in the expansion chunk, the repeat tends to occur on the first element of the chunk rather than on the determiner. We are going to check this hypothesis.

4. Repeats and prepositional noun chunks

4.1. Patterns of repeats in prepositional noun chunks

Among the 96 repeats which affect prepositional noun chunks, we have kept only those which follow a [Preposition + (predeterminer) + Determiner + (modifier) + Noun + (modifier)]$_{PNC}$ pattern. We have found 59 occurrences. The study of the data allowed us to see six possible configurations, depending on whether the speaker goes back to the preposition or not:

- back to the preposition:

• the preposition only (47.5%):
ex. 8: on parle des journaux **dans dans** la plaine euh + euh bourguignonne

• the preposition and the determiner (23.7%):
ex. 9: là + **à ce : ++ beh à ce** lycée + j'ai eu des élèves absolument remarquables

• the whole of the prepositional noun chunk (1.7%):
ex. 10: et ensuite ++ euh on applique on euh l'émail **sur la pièce sur la pièce** en terre qui est déjà cuite

- later in the chunk:

• the determiner only (23.7%):
ex. 11: nous on les met dans **les : les** machines

• the determiner and the noun (1.7%):
ex. 12: aussi que je voulais dire + euh à propos des différentes terres + et de : justement que de **l'idée euh l'idée** reçue que se font les gens de la poterie

• the determiner and the “quantifier” of the noun (1.7%):
ex. 13: nous nous entendions bien avec **les deux : + les deux** Anglaises

The distribution of the repeats is as follows:

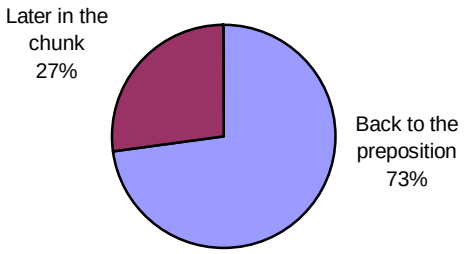

Figure 3: Frequency of the two major types of patterns of the repeats in prepositional noun chunks.

The figure above clearly indicates that, when the speaker begins a prepositional noun chunk, he tends to repeat the first element of the chunk (73%), that is to say the preposition or the unit composed of the preposition and the determiner. Cases when repeats appear later in the chunk are only a minority (27%).

Besides, in prepositional phrase expressions like *en dehors de*, *vis-à-vis de*, *dans la mesure de*, etc., the whole expression is rarely repeated:

ex. 14: et j'avoue qu'**avant de : avant de** me marier

We note that the repeat, most of the time, only affects the end of the fixed expression with *de* and not the whole of the expression[3]:

ex. 15: on essaie d'aider dans la mesure **de : + de** nos moyens actuels

ex. 16: et en fait moi je me suis aperçue au cours **de ces : de ces** soirées que beh il y a beaucoup de personnes d'un certain âge qui participent aux soirées dansantes

These expressions can not be broken down, and it is not possible to establish a relationship between the "head" and what follows[4]: ?*on essaie d'aider dans notre mesure*; that is the reason why we did not study these cases differently from the cases when the whole fixed expression is repeated.

4.2. Prepositional noun chunks as expansion chunks

When the prepositional noun chunk is the expansion chunk, the tendency to repeat the preposition only increases: 65% of cases *vs* 47,5%. Our hypothesis is thus confirmed: when complex noun phrases contain a determiner in the expansion chunk, the repeat tends to occur on the first element of the chunk rather than on the determiner.

5. Discussion

This study has permitted us to show that many different syntactic constraints bear an influence on repeats. We have been able to establish a co-relation between the complexity of the phrase and the frequency of the phenomenon. Indeed, the determiners we selected as test words are on average twice more likely to be repeated in complex noun phrases than in simple noun phrases. Furthermore, these elements are repeated mainly in head chunks (83% of cases). Our results are compatible with Clark & Wasow's findings on the English language. The "weight" of the phrase would be an additional constraint the speaker has to manage when he structures his speech.

As regards all the complex NPs in our corpus, we have counted as many repeats in the head zone as in the expansion zone. This result can be explained by the fact that the preposition – the juncture between the two chunks – is more often repeated than the determiner following it.

6. References

[1] Abney, Steven. 1991. Parsing By Chunks. In Robert Berwick, Steven Abney and Carol Tenny (Eds.), *Principle-Based Parsing.* Dordrecht: Kluwer Academic Publishers.

[2] Adda-Decker, Martine, *et al.* 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *Proceedings of DISS'03*, 5-8 September 2003, Göteborg University, Sweden, pp. 67-70.

[3] Blanche-Benveniste, Claire. 2003. La naissance des syntagmes dans les hésitations et répétitions du parler. In J.L. Araoui (Ed.), *Le sens et la mesure. Hommages à Benoît de Cornulier*. Paris: Editions Honoré Champion, pp. 40-55.

[4] Blanche-Benveniste, Claire. 1990. *Le français parlé. Études grammaticales.* Paris: CNRS Éditions.

[5] Campione, Estelle. 2001. *Étiquetage semi-automatique de la prosodie dans les corpus oraux: algorithmes et méthodologie.* Thèse d'état, Université de Provence, Aix-en-Provence, France.

[6] Candéa, Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'« hésitation » en français oral spontané.* Thèse d'état, Université Paris III, Paris, France.

[7] Clark, Herbert H. & Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. Cognitive *Psychology 37*, pp. 201-242.

[8] Équipe DELIC. 2004. Présentation du Corpus de Référence du Français Parlé. *Recherches Sur le Français Parlé 18*, Publications de l'Université de Provence, pp. 11-42.

[9] Henry, Sandrine, Estelle Campione & Jean Véronis. 2004. Répétitions et pauses (silencieuses et remplies) en français spontané. *Actes des XXV^èmes Journées d'Études sur la Parole*, 19-22 Avril 2004, Fès, Maroc, pp. 261-264.

[10] Henry, Sandrine. 2002. Étude des répétitions en français parlé spontané pour les technologies de la parole. *Actes de la 6^ème RECITAL*, 24-27 Juin 2002, Nancy, France, tome 1, pp. 467-476.

[11] Lickley, Robin. 1994. *Detecting disfluency in spontaneous speech.* Ph.D. thesis, University of Edinburgh.

[12] Martinie, Bruno. 1999. *Étude syntaxique des énoncés réparés en français parlé.* Thèse d'état, Université Paris X-Nanterre, Paris, France.

[13] Riegel, Martin, Jean-Christophe Pellat & René Rioul. 1999. *Grammaire méthodique du français.* Paris: Presses Universitaires de France (5^ème édition, 1^ère édition: 1994).

[3] We observe the same behaviour in adverbial expressions and complex determiners: "on fait beaucoup de : colonies : *beaucoup **de : + de*** choses comme ça", "on a quand même *un laps **de : de*** repos moi je suis du matin ma collègue est du soir". For more information on the behaviour of these determiners, see Claire Blanche-Benvensite's work, [4], pp. 109-111.

[4] *Cf.* Martinie [12], p.99.