Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, 5–8 September 2003, Göteborg University, Sweden.
Robert Eklund (ed.), Gothenburg Papers in Theoretical Linguistics 90, ISSN 0349–1021, pp. 77–80.

Word fragments and repeats in spontaneous spoken French

Sandrine Henry† & Berthille Pallaud‡

† Équipe DELIC, Université de Provence, Aix-en-Provence, France
‡ CNRS, UMR 6057 Langage et Parole, Université de Provence, Aix-en-Provence, France

Abstract

This paper presents the results of a study conducted on the interaction of two disfluencies: repeats and word fragments. It is based on 150 repeated word fragments (e.g., "on le **re- re-**revendique encore une fois") extracted from a one-million-word corpus of spoken French. Word fragments such as: "notre métier **spé-** spécifique", are, like repeats (e.g., "vous avez évalué **le le** montant des dégâts"), very frequent events in spoken language: on average, there is 1 word fragment every 50 seconds,[1] 1 repeat every 17 seconds. Speakers and listeners alike are generally unaware of these phenomena as if they were not part of the communication process. They seldom trigger a metalinguistic reaction from the speaker and are even more rarely acknowledged by the listener. These phenomena have sometimes been interpreted as 'errors' in the communication process, like slips of the tongue [6]. Word fragments and repeats encompass different categories of phenomena, and this enables us to define them as an heterogeneous group ruled by different types of constraints and mechanisms.[2] This analysis rests on the following criteria: structural aspects of the repeat, types of word fragments, morphological and syntactic aspects. Analyses of these repeated of identical word fragments from two different angles – that of the repeats and then that of the word fragments – confirm the relevance of the distinction between these two types of disfluencies.

1. Introduction

Disfluencies have often been considered as traces of the elaboration that encumber the oral utterance and have therefore long been ignored by the linguists. We contend that these performance phenomena are to be taken seriously into account for they reflect the production processes at work and can thus shed light on the planning of constituents.

1.1. Repeats

Actually, previous studies [2, 3] have shown that these repeats, in French as in English, mostly concern function words. We have found that function words are five times more likely to be repeated than lexical words. And, among repeated words, 91.3% are function words whereas only 8.7% are lexical words. We have also classified the repeated function words according to word classes: 41.5% are determiners, 35.5% pronouns and 13.0% prepositions. As for lexical words, most of them (52%) are adverbs, then adjectives (25.0%) and verbs (11.0%).

Repeats tend to appear at major syntactic boundaries, as in the following example:

(1) "**le le** terrain commençait à glisser beaucoup"

In (1), the determiner *le* is both at the left edge of the noun phrase *le terrain* and at the left edge of the clause *le terrain commençait à glisser beaucoup*. As repeats chiefly affect function words, it seems quite logical that repeats should occur at the beginning of phrases. However, a recent study [3] of *le*, a word that shows a multiple class membership (i.e. belongs to more than one word class: 'le' can be a determiner or a pronoun), has shown that only 1.33% of *le* as accusative pronoun are repeated *vs* 5.64% of *le* as determiner. It means that syntactic constraints – not to the morphological status of the repeated element – are responible for this tendency of repeats to appear at the beginning of phrases.

The structure of a repeat can be defined as follows:

"le {R_0} le {R_1} terrain commençait à glisser beaucoup"

repeat = R_0 ('repeatable element') + R_1 ('repeated element(s)') [2]

If we consider the larger description of disfluency phenomena provided by Shriberg [9], our term 'repeatable' (R_0) corresponds to reparandum (RM) and 'repeated' (R_1) to repair (RR). Shriberg has described an intermediate region in between the two, called interregnum (IM), that can remain empty (consecutive repeats, as can be seen in Figure 1) or can contain other disfluencies (for instance a filled pause), or editing terms [4], or again parenthetical clauses.

Interruption Point			
R_0	↓	R_1	
le		**le**	terrain commençait à glisser beaucoup
↔	↔	↔	
RM	IM	RR	

Figure 1: Structure of a repeat.

We have retained three structural criteria as regards repeats:

- The number of elements composing the 'repeatable'.
- The presence or absence of material in the interregnum.
- The span of the repeat, i.e. the number of 'repeated elements'. We have thus divided them into two categories: simple repeats (e.g. *le le*) or multiple repeats (e.g. *le le le*). Simple repeats are by far the most frequent (93.8%): in only 5.3% of cases words are repeated twice; and the rate becomes very low for three times or more (0.9%).

We have focused on the latter criterion, for we thought the first two were irrelevant to our point: first of all, repeated word fragments are rarely composed of more than one element,[3] and secondly, our aim was to focus on the interaction of two disfluencies, and we thus could not refer to other types of combinations (such as with silent or filled pauses). That would demand another study.

[1] With an average rate of 200 words/minute.
[2] Recent neurophysiological studies [5] on detection of repeats and false starts – i.e., syntactic interruptions – in utterances have shown that detection of these two types of disfluencies goes along with different event-related potentials (ERPs).
[3] 17 occurrences for repeated word fragments composed of two elements and only 1 composed of three elements in our one-million-word corpus.

1.2. Word fragments

Three basic observations were made in previous studies [7, 8]:

- Contrary to repeats, word fragments mostly affect lexical words (70%), as in the rather typical example that follows: "c'est vrai que c'est pas **b- beau** d'associer les deux choses". If we examine the distribution of word fragments according to the type of constituent, we note that more than half of them (50.2%) are in the Object position, 35.1% in the Verb Position and 12.7% in the Subject position.
- When the speaker produces a word fragment, he momentarily suspends his speech. What is at stake is to find out if the element that allows him to resume speaking belongs to the same syntactic locus as the word fragment.
- Word fragments are either listing phenomena (corresponding to a lexical search on a given syntactic locus) or elements that trigger a syntactic breach (in cases where the context following the word fragment does not belong to the same syntactic unit). There are three categories of word fragments:
  - **completed word fragments:** the word fragment is completed on the same syntactic locus:
    "c'est vrai que c'est pas **b- beau** d'associer les deux choses"
  - **modified word fragments:** the fragment is not completed but replaced by another word belonging to the same syntactic unit:
    "on va + attaquer l'autre **b- morceau** l'autre moitié du dos"
  - **word fragments left incomplete:** the word fragment does not initiate a listing phenomenon [1] and what follows belongs to another syntactic locus. It corresponds to what Levelt [4] calls a 'fresh start':
    "alors je vais euh faire un petite **diver-** on va diverger là pour expliquer ça euh au début"

Out of 948 word fragments, 59.6% were completed word fragments, 21.9% incomplete and 18.5% modified.

1.3. Repeated word fragments

Repeated word fragments form a minor sub-category among 'stumbling' events:
"il vaut mieux être **ho- ho-** honnête vis-à-vis des gens"
"mais **no- no- no-** notre base politique veut le que ouais que la que le peuple ait souvent son + son mot à dire"

2. Method

2.1. Corpus

Our corpus is composed of 1,000,382 words,[1] it corresponds to 283 situations of spontaneous speech and involves 794 different speakers.

2.2. Extracting the data

First of all we selected the repeats, word fragments and repeated word fragments using a program (script in *Perl* language, application under *Linux*). We then proceeded to a manual sifting of the data.

Repeated word fragments are only a minor phenomenon among word fragments and repeats. On a total of **6 094 word fragments**, we found only 150 repeated word fragments, that is to say **2.4%**. On a total of **16 135** (repeated word fragments included), only **0.93%** are repeated word fragments.

The frequency of repeated word fragments is low: 1.5 every 10,000 words that is – with an average rate of 200 words/minute – one repeated word fragment every 33 minutes.

3. Results

3.1. Structural aspects of repeats

Validating our hypothesis of an interaction between the two phenomena, we first notice that, much in the same way as repeated words [3], simple repeats of word fragments are the most common occurrences (84.7%), way before double (14.0%) and triple (1.3%). The likelihood of having a repeat grows smaller as the number of repeated elements rises, and this applies to the two phenomena.

3.2. Repeated word fragments and categories of word fragments

The type of word fragment is another criterion in our analysis, what is at stake here is to find out whether repeats of word fragments abide – or not – by the rules of word fragments.

Table 1: Numbers and percentages of repeated word fragments according to the type of word fragment.

	Repeated word fragments	Word fragments
Completed	123 (82.0%)	565 (59.6%)
Modified	7 (4.7%)	175(18.5%)
Incomplete	20 (13.3%)	208 (21.9%)
Σ	150 (100.0%)	948 (100.0%)

As with previous results on word fragments, we note important differences in the frequencies of the three categories of repeated word fragments (completed, revised or incomplete). Among them, completed word fragments are by far the most frequent. Repeated word fragments are rarely incomplete or modified.[2] This difference in distribution between word fragments alone and repeated word fragments shows that the repeat phenomenon affects word fragments and confirms our hypothesis of the word fragment phenomenon as a 'stumbling' event which is the site of a lexical search. The repeat sustains this search for, in most cases, the truncated word is completed. This interdependence of repeats and word fragments suggests that these repeated word fragments work like filled pauses, and from a wider perspective, like a filler: the speaker suspends his speech – there is thus a 'stagnation' on the syntagmatic axis – and then he goes on.

3.3. Repeated word fragments and word classes

An analysis of repeats of word fragments according to their morphological status has also been conducted. The results appear in tables 2, 3 and 4.

Table 2: Distribution of repeated word fragments according to the morphological status of the repeatable element.

	Repeated word fragments	Repeated words	Word fragments
Function words	100 (66.7%)	14 594 (91.3%)	151 (30.1%)
Lexical words	46 (30.7%)	1 391 (8.7%)	350 (69.9%)
Misc.	4 (2.6%)	–	–
Σ	150 (100%)	15985 (100%)	501 (100%)

Function words account for most cases of repeated word fragments. That is also the case for repeats, with an even

[1] It is composed for the main part of *Corpaix*, (a numerical corpus that took shape thanks to the work of the GARS team in the past 25 years, currently DELIC). All the transcriptions conform to the conventions established by the GARS.

[2] Repeats of incomplete word fragments compared to completed + modified: chi-square = 5.83 ; d.d.l. = 1 ; $p < .02$.
Repeats of modified word fragments compared to completed + incomplete: chi-square = 17.81 ; d.d.l. = 1 ; $p < .001$.

higher rate (90%). The two phenomena tend to follow more or less the same trends, but there are considerable differences in the distribution of lexical words: they amount to only 8.7% of repeats, *vs.* 30.7% of repeated word fragments. Lexical words are thus more sensitive to an interaction between the repeat and the word fragment, and the corollary of this is that function words – massively present in repeats – are much less involved when the word fragment phenomenon is added to the repeat. The repeat phenomenon inverts the distribution of function and lexical words for word fragments.

Table 3: Repeated word fragments: a distribution according to function word class.

	Repeated word fragments	Repeated words
Pronouns	61 (61%)	5 181 (35.5%)
Determiners	22 (22%)	6 057 (41.5%)
Prepositions and complex prepositions	9 (9%)	1 897 (13.0%)
Conjunctions (subordinators and coordinators)	5 (5%)	1 021 (7.0%)
Auxiliaries	3 (3%)	146 (1.0%)
Misc.	–	292 (2.0%)
Σ	100 (100%)	14 594 (100%)

Two grammatical categories are prevalent in both types of repeats: pronouns and determiners. The table above nevertheless shows a significant[1] difference in the distribution of these grammatical categories according to the type of repeat involved: in the case of word fragments, determiners outweigh pronouns (61% *vs* 22%), whereas the latter prevail in the case of repeats (41.5% *vs* 35.5%).

Table 4: Distribution of repeated lexical word fragments according to lexical word class.

	Repeated word fragments	Repeated words
Verbs	26 (56.5%)	153 (11.0%)
Adverbs	11 (23.9%)	723 (52.0%)
Nouns	6 (13.1%)	167 (12.0%)
Adjectives	3 (6.5%)	348 (25.0%)
Σ	46 (100%)	1 391 (100%)

There are few repeats of lexical word fragments: actually, out of the 150 word fragments in our corpus, 46 of them are lexical words. Hence, the following figures are but trends. These repeats chiefly involve verbs (56.5%), whereas repeats of lexical words are mostly adverbs (52.0%). Moreover, if we set aside the noun class which shows identical proportions whatever the type of repeat, we notice that adjectives, which were one fourth of the repeated lexical words, only represent a very small part of the repeated lexical word fragments. There is thus a strong co-relation between the type of repeat and the lexical words.

3.4. Syntactic analysis of repeated word fragments

The distribution of repeated word fragments according to the type of word fragment on the one hand and to syntactic constituents (Subject, Verb, Object) on the other enables us to examine if these two variables are interacting.

Table 5: Distribution of repeated word fragments according to the type of word fragment and to syntactic constituents.

	Completed	Modified and Incomplete	Σ
Subject	65 (52.8%)	6 (22.2%)	71 (47.3%)
Verb	17 (13.8%)	9 (33.3%)	26 (17.3%)
Object	32 (26.0%)	8 (29.6%)	40 (26.7%)
Misc.	9 (7.3%)	4 (14.8%)	13 (8.7%)
Σ	123 (100%)	27 (100%)	150 (100%)

Concerning completed repeated word fragments, no interaction was observed between the type of word fragment and the syntactic constituents. Actually, completed word fragments in a Subject position remain prevalent (52.8%), followed by word fragments in Object position (26%) and Verb position (13.8%). This distribution is equivalent to the one observed in all repeated word fragments (chi-square = 0.93 ; d.d.l. = 2 ; N.S.).

As concerns the category of modified and incomplete word fragments, the distribution according to the type of constituent seems to occur at random, all types of constituents appear more or less equally in repeats. However, the scarcity of this type of data does not allow us to form any definite judgement in that case.

Table 6: Distribution of word fragments according to the type of word fragment and to syntactic constituents.

	Completed	Modified and Incomplete	Σ
Subject	92 (15.4%)	29 (8.2%)	121(12.7%)
Verb	188 (31.3%)	147 (41.5%)	335 (35.1%)
Object	306 (51.0%)	173 (48.9%)	479 (50.2%)
Misc.	14 (2.3%)	5 (1.4%)	19 (2.0%)
Σ	600 (100%)	354 (100%)	954 (100%)

This interaction does not show for word fragments [7]: their distribution depends on the syntactic locus, not on their category. Half the word fragments are in Object position, 35% on the Verb position and only 13% in the Subject position.

Therefore, on the syntactical level, there seems to be an interaction between repeats and word fragments only as regards modified and incomplete word fragments.

Table 7: Distribution of repeated word fragments according to the type of word class and the syntactic constituent.

	Repeats of lexical word fragments	Repeats of function word fragments	Σ
Subject	3 (6.5%)	68 (68.0%)	71 (48.7%)
Verb	21 (45.7%)	5 (5.0%)	26 (17.8%)
Object	16 (34.8%)	23 (23.0%)	39 (26.7%)
Misc.	6 (13.0%)	4 (4.0%)	10 (6.8%)
Σ	46 (100%)	100 (100%)	146 (100%)

The table above proves the state of dependency between the morphological status of repeated word fragments and their syntactic situation.

In the case of repeated lexical word fragments, repeats mainly occur on the Verb (45.7%) and Object (34.8%) positions. When it comes down to function word fragments, the distribution is completely different: 5% for the Verb position and 23% for the Object. In the Subject position, repeats of function word fragments and repeats of lexical word fragments follow completely opposite trends: 68% for the first, 6.5% for the last.

[1] Chi-square = 24.86 ; d.d.l. = 1 ; $p < .001$.

4. Discussion

This study brings to light the co-relation that exists between repeats and word fragments.

As far as the span of the repeat is concerned, the same trends emerge for repeats of words and repeats of word fragments.
If we consider the type of word fragment involved in repeat, the prevalence of completed word fragments proves that the repeat phenomenon fuels the lexical search the word fragment expresses.

On the morphological level, repeated word fragments abide by the constraints of repeat phenomena, not by those of word fragments, for repeats of word fragments involve function words for the main part. According to the class of the function words, there is a significant difference between repeats of word fragments and repeats.

On the syntactic level, the distribution of repeated word fragments turns out to depend on the constituent (Subject, Object or Verb). Word fragments in Subject position represent only 12.7% of word fragments, whereas they account for half of the repeated word fragments. The repeat further also accentuates the trend towards completion in Subject positions (52.8% for repeated word fragments *vs* 15.4% for fragments words). On the contrary, the distribution of repeated word fragments seems to occur at random when the word fragment is modified or incomplete, which is not the case when the word fragment is not repeated. Repeats of lexical word fragments especially occur in Verb (45.7%) or Object (34.8%) positions, whereas repeats of function words mostly appear in Subject position (68%).

5. Acknowledgements

Thanks to Jean Véronis, Sandrine Henry's dissertation tutor for his comments and reviews, and to Daniel Hirst. We are both indebted to Sophie Girardin for the translation into English.

6. References

[1] Blanche-Benveniste, Claire. 1987. Syntaxe, choix de lexique et lieux de bafouillage. *DRLAV*, vol. 36–37, pp. 123–157.

[2] Candéa, Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané. Étude sur un corpus de récits en classe de français.* Thèse d'État, Université Paris III (Sorbonne Nouvelle).

[3] Henry, Sandrine. 2002. Étude des répétitions en français parlé spontané pour les technologies de la parole. *Actes de la 6ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'02)*, 24–27 June 2002, Nancy, France : ATALA, pp. 467–476.

[4] Levelt, Willem J. M. 1989. *Speaking. From Intention to Articulation*. Cambridge, Massachusetts: MIT Press.

[5] McAllister, Jan, Cato-Symonds, Susan & Blake Johnson. 2001. Listerners' ERP Responses to False Starts and Repetitions in Spontaneous Speech. *Proc. DISS'01*, 29–31 August 2001, Edinburgh, Scotland, pp. 65–68.

[6] Pallaud, Berthille. 2001. Les lapsus: des pierres dans le champ linguistisque. In M. Arrivé & C. Normand (eds.), *Linguistique et psychanalyse*. Colloque de Cerisy-la-Salle, 1–8 September 1998, IN Presse, pp. 47–66.

[7] Pallaud, Berthille. 2002. Les amorces de mots comme faits autonymiques en langage oral. *RSFP*, vol. 17, pp. 79–102.

[8] Pallaud, Berthille. 2003. Achoppements dans les énoncés de français oral et sujets syntaxiques. In J.M. Merle (ed.), *Le Sujet*, Paris: Éditions Ophrys, Faits de Langue, pp. 91–04.

[9] Shriberg, Elizabeth. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of Berkeley, California.