Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 145-149.

The re-adjustment of word-fragments in spontaneous spoken French

Berthille Pallaud

Parole et Langage, Université de Provence, Aix-en-Provence, France

Abstract

A study of word-fragments in spoken French has been undertaken for a few years on the basis of non directive talks corpora recorded and transcribed according to GARS'conventions (DELIC currently). These disfluencies are often analyzed within the framework of disfluent repetitions. The observations made on these two types of disfluencies led us to distinguish them. The aim of our study is to describe on the one hand insertions which take place in relation to the word interruptions and their re-adjustment, and on the other hand, to specify the types and localizations of retracing which follow these interruptions. Two kinds of incidental clauses were observed at the time of the readjustments which follow these disturbances. Some, (the more numerous) are syntactically linked to the fragment or with its retracing, others are not. Moreover, the word-fragments which will be modified are the only one to be dependent on the type of localization. For the others, this localization does not make it possible to predict the category of interruption (complemented or unfinished). Our results on word-fragments, confirm however that in contemporary French, the retracing at the head of the nominal or verbal group which contains the disfluency remains the simplest example (at the same time the most frequent, [5]. Nevertheless, a third of the retracing either does not go back to the beginning of the Group, or exceeds it.

1. Introduction

If the fluidity of an oral statement is measured by the rhythmic regularity in its production, it is clear that the statement is not fluent but disfluent [14, 12]. All speakers produce oral speech with a certain variability in the flow, pauses, whether they be silent or not [6], and in the lengthening of linguistic elements, some of which were studied at the point where the statement was interrupted, either in the middle of a word, such as in the case of a word-fragment, or at the boundary of words, such as syntagm interruptions which were or were not followed by word repetitions [15]. These interruptions are also characteristic of oral disfluencies. This study, using the grid paradigm [2], in fact has highlighted the progression of the statement by successive syntactic steps. These "halts" in the production of the text were due not only to silent or filled pauses, known as hesitation pauses [6], but also to word repetitions and fragments.

These last two types of stumbling, that is to say disfluencies, have rarely been distinguished (2,5). So, we thought it was important that they be characterized separately, and the big difference in frequency suggested that there were two kinds of phenomena. The results already obtained confirmed this hypothesis; in fact, they affected neither the same syntactic places nor the same grammatical categories [10, 11].

The aim of this study on word-fragments was to specify their syntactic position and the place in the statement where the subject returned to when it happened. Shriberg & Stockle [14] showed that by repeating words (where there is always a return to the statement) the speaker tended to go back to the beginning of the word with which he had problems formulating. This question regarding word-fragments, will only be considered from the position (of the word and the return) and morpho-syntactic and syntactic aspects.

2. Corpus and methodology

The study on spoken French throughout France, headed by Claire Blanche-Benveniste, was carried out in 1998 and 1999 on 20 corpora, the results of which were collected, with the exception of one, by the GARS. The entire survey was followed and digitized - the sound and transcription - by the DELIC team. In accordance with GARS, when transcribing conventions, which enjoin an orthographical transcription of the oral statements, the word- fragments are noted using a hyphen attached to the fragment of the word, which is then detected automatically.

All but two recordings were done in private, using non-directive interviews and only two speakers. Only two corpora were recorded in public, which consisted of improvised talks in front of a group of 40 people. All the speakers were adult.

These corpora were not labelled. The analysis was carried out exclusively on the statements of the interviewed speaker. The 441 extracted pieces of data from these corpora were then labelled (not automatically) and entered on a spreadsheet (using Contextes software by J Véronis and Excel). Our study was based on a medium flow of 200 words/min, with the total duration of the group lasting 7 hours 51 min. The average length of time of the GARS corpora was 3,080 words, thus an average duration of 16 min with the extreme values of these durations being 1,307 and 4,931 words. Hence, we found in this corpus subset, the average frequency of the apparition of word-fragments in a statement was 1/57sec - from 1/23 sec. to 1/8 min.

3. Results

3.1. Types of word-fragments

Three types of word-fragments were distinguished [10] in relation to the syntactic place occupied by what followed the word-fragment. When the following element occupied the same syntactic place, this insistence could result in completing the word-fragment (completed fragments) or replacing it (modified fragment). If the element, which followed the fragment, belonged to another syntactic place, the fragment of the word was left unfinished (unfinished fragments). A little more than half of the word-fragments were completed; modified and unfinished fragments constituted the remainder equally:

Ex. 1 completed fragment: EDF, 20,5 qui ont la possibilité de **r-** de **remonter** euh autour d'un axe

Ex. 2 modified fragment: EDF, 19,5 5 en revenant à nos **ber- Bassins** Versants Intermédiaires qu'on a qu'on par-qu'on *a* parlés précédemment euh

Ex. 3 unfinished fragment: EDF, 15,2 2 toutes les usines nous appellent nous **di-** et nous donnent par forme de fax ou d'e-mail euh

The completed and modified fragments were the only cases where the speaker resumed or “repaired” the flow of his/her speech by going back to the interrupted part of the statement. The unfinished fragments were those where the speaker simply continued the statement without any reparation or retracing.

The proportions in these three categories showed that these involuntary truncations appeared a lot more frequently as hesitation markers when developing the text than as the sign of an error to be corrected. Our observations confirmed the studies of Schegloff et al. [13] concerning auto-corrections (completed fragments in our study) and Cappeau’s findings on syntagm fragments [4].

Contrary to the disfluent repetitions [11], these hesitations were mainly related to the nominal or verbal lexicon (70%) and much less to the functional word category. In the same way, the hesitation was more frequently a repetition than a word-fragment before the verb. The functional-words were significantly less frequently modified and more often left unfinished.

When the fragment was repeated, it was, in 82% of the cases, completed (instead of 59%) and barely modified or left unfinished. This last, very different result (kh2 = 29.85; d.d.l.= 2; p<.001) of what was observed on the simple fragment justified an in-depth study of the retraces and readjustments after the interruption of the word as the repetition of the fragment seemed to allow the speaker to complete more frequently the started word, and less frequently to modify it or leave it unfinished.

3.2. Analyses of re-adjustments after the interruption of words

Two types of phenomena could be defined: firstly, linguistic **insertions** of elements, and secondly, **retracing** of the statement (whether it be by syntactic insistence or by the continuation of the statement on the syntagmatic axis). The term "retracing" was only employed if an insistence occurred in the same syntactic place (thus the case of the completed or modified word-fragments).

3.2.1. Insertions

The structure of the disfluencies [5, 14) which differentiated the three phases in stumbling revealed two possible spaces of insertion following the truncation: the space which preceded the retracing of the statement and that which began with the retracing of the statement.

The three phases of this word-fragment structure are as follows:

* **the reparandum** (RM): indicates the word-fragment The interruption point (IP): establishes the final boundary of *the reparandum*: (it is identified by the hyphen in our transcriptions).

* The **interregnum** (IM, *the hyatus* in [5]): indicates the moment between the final boundary of *the reparandum* and the initial boundary of *repair*.

* The **repair** or **reparans** (RR): represents the repaired, repeated or modified part of the *reparandum*;

Thus, it is a 3-phase structure: **interruption, latency and the retracing of the statement:**

C24Bnanc, 24,2 ça **dép-** (Reparandum RM, IP)
ouais [Space IM]
ça ***euh*** **dépend (**Reparans, space RR**)**
+ moi au début j’avais demandé à aller en au Népal

Two kinds of insertions were distinguished:

- Firstly, interpolated clauses which were inserted into the statement without being syntactically linked (pauses voiced or not, enunciator and parenthetic clauses).
- Secondly, interpolated clauses syntactically linked to the interrupted statement and its retracing: the repeated word-fragments (IM space only), repetitions, added, removed or replaced items (RR space).

3.2.1.1 The first space ([IM space] "Interregnum") was only filled in 13 % of the cases.

Enunciator insertions (voiced pauses or not, *enfin*, *ben, bon,* etc), parenthetic items and repetitions of the word-fragment were observed:

Ex 4 EDF, 14,4 c’est qu’au premier choc euh **prétro-** [*euh*] **pétrolier** euh

Ex 5 C6bBesan 2,8, et j’ai **regard -** [*et j’ai en fait choisi ça par hasard*] j’ai **regardé** ce qu’il fallait avoir

Ex 6 CäBelfo, 3,2, euh et **j’ai-** [*j’ai*] **j’aimais** pas quoi

The moment, which followed the interruption of word, was thus a potential space for enunciator or parenthetic insertions. However, in our corpora, it was only “used" by the speaker in one out of 10 cases. Besides, if we had taken out repetitions of word-fragments, which was rare, we would never have observed sequences of successive truncations as one can find in the case of stammering [21, 1, 7]

3.2.1.2. The second space ([RR space], "Reparans") began with the retracing of the statement which re-established the continuity in its development where it had initially been interrupted. We found the same insertions as described in IM space, although they were fewer:

(parenthetic clause) EDF, 18,5 pour euh mettre les **ni-** [les *comment* les **niveaux**] euh d’eau corrects
(silent pause) EDF, 13,2 mais la co-génération restera **u-[une + une] solution**
(repetition) C24bNan, 2,1 c’était un peu euh un peu **com- comment** ***comment*** vivre au quotidien

In this second space (RR), it was assumed that it could be possible to find the same insertions In fact, they accounted for only 3%. We found in particular, insertions which would modify the interrupted statement by this truncation, which was not therefore, the moment of correction. Thanks to these "back modifications", this space comprised of, in total, two times more insertions than IM space, that is to say, 22% to 13%. There were three types of modulating statement insertions: term additions, suppressions, and replacements.

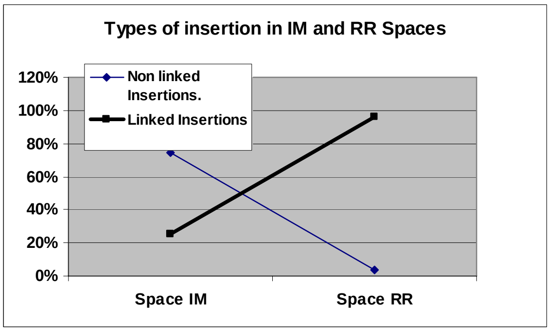

Figure 1 Insertion types linked or not syntactically to the interrupted statement in IM and RR spaces

Ex 7 Interpolated clauses which modulate the statement
*Retracted item C24bNan, 20, 1 il y **en** av- il y avait deux filles de dix-huit ans
*Added item C7cBorde, 12, 11 le judoka commence à être pas mal et euh s'il a s'il s'en- **continue à** s'entraîner
*Replaced item EDF 2,41 qui est un peu **plus** mo-un peu **moins** modulable

In total, a third of the word-fragments were followed by various insertions, in IM and RR spaces, and this in fact occurred mainly in the completed word-fragment category. As our observations on the repeated fragments suggested, these insertions, encompassing especially the completed fragments, seemed to play a role in the lexical research by facilitating it. These two spaces (Fig.1) were not composed of the same types of insertion. The space located just after the truncation seemed to be where the non-syntactically linked interpolated clauses were located, whereas they were practically absent when the retracing started. However, in the RR space, the majority of the interpolated clauses adjusted the retracing of the statement.

3.2.2. Retracing

Retracing after interruption only occurred where fragments were completed or modified, these being the only word-fragments followed by an insistence on the same syntactic place (i.e. a retracing). The unfinished fragments were precisely identified because what followed the fragment did not belong to the same syntactic place. The re-adjustment in this case was not a retrace but a continuation of the statement. The proportions of these three word-fragment categories validated the conclusions of Levelt [8] on the phenomena of an interruption in statements where the speakers retrace more than they continue their statement after an interruption: 78% to 22%. We were interested in the position of the interrupted word and the retracing which followed.

It was not possible to describe the exact place of retracing after an interruption without specifying beforehand, **the place of this interruption,** which might or might not have taken place at the beginning of the nominal or verbal group: [1]

At the beginning: C24aNanc 1, 3 euh la mygale **s' – s'arrime** avec ses ses crochets sur sa sa proie

Not at the beginning: C7dBord 5, 5 et et après j'ai **vou-voulu** changer

Clark and Wasow [5] stated that retracing after the interruption of a constituent, whether it be a group or a syntagm, more frequently occurred at the beginning of this constituent. If this was the same for the word-fragment and the latter was already at the beginning of the group, then the place where retracing would begin would not have the same signification as it would if the fragment was not already there. Indeed, if the word-fragment was not at the beginning of group but the retracing went back to this point, this would reveal a linguistic constraint which would be impossible to show as the fragment would already be at the beginning of the group.

3.2.2.1. Localization of the word-fragment

Out of the 436 word interruptions where it was possible to determine if they took place at the beginning of the group or not, more than two thirds of them did not (72%): in fact, there were no differences between the nominal and verbal groups. Hence, a large majority of these disfluencies (in French the lexicon is seldom found at the beginning of a group, but is preceded by a determinant or a pronoun) occurred later in the nominal or verbal group.

3.2.2.2. Localization of the retracing which followed a word-fragment which was not at the beginning of a group.

The mass result on the development of the word-fragment was that when this fragment was completed, it was always (without any exception in our corpora) through at least a minimal retracing of the fragment. There was never a simple completion of the fragment by the missing fragment (of the type: **un li vre**). This was another "qualitative" feature which made it possible to distinguish a disfluent statement from a statement produced by a person who stammers [16].

In addition, we noted that the localization of the disturbance did not make it possible to predict if there would be a retracing or a simple continuation of the statement, and in fact, wherever the fragment was localized, we found a similar proportion (one out of five) of unfinished fragments (without retracing). It was the same for the fragments which would be completed. Only those fragments which would be modified were significantly more numerous as the disturbance was not located at the beginning of the group (kh2 = 10,46; p<.01; d.d.l.=2):

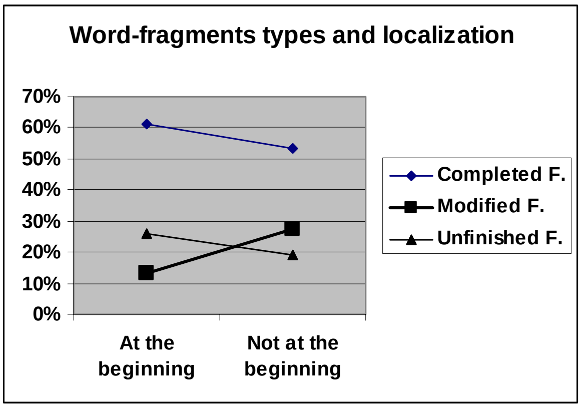

Figure 2 Localization of the word-fragments (whether or not at the beginning of a group in relation to the word-fragment types)

Moreover, the types of observed retracing could be regrouped into three categories:
* "**Minimal** " retracing which did not go back to the beginning of the group: CorpusEDF, 12,6 la note est est salée hein la note **énergé- énergétique**
* "**Beginning of the Group** "retracing which went back to the beginning of the nominal or verbal group: C5cBelfo, 2,2 alors ce monsieur s'est approché et il a dit ***vous* li- *vous* lisez** l'Est Républicain donc vous êtes de l'Est

[1] We refer here to the hierarchical concepts employed by Blanche-Benveniste [2}

* "**More** " retracing which started well before the group: Corpus EDF, 8,38 mais ***l'eau* est de vingt mè-l'entrant est de vingt metres cubes**

This type of disfluency did not make any difference to the more general results shown by [5]: retracing generally took place at the beginning of the nominal or verbal group (71%) although 29% did not actually conform to this model. Some (19%) did not retrace back to the beginning of the group and others (10%), on the contrary, exceeded this limit. This could however, have been a characteristic of our speaker. It would obviously be necessary to check these results on a much larger corpora comprising of more speakers.

3.2.2.3. Localization of the retracing which followed a word-fragment located at the beginning of the group.

In our corpora, only 28% of the fragments were found at the beginning of the group. The completed and modified fragments, followed by retracing, in fact accounted for 92% and this in turn was followed by a readjustment located at the place of the word-fragment; that is to say at the beginning of the group. As there was no opposition, retracing took place before the statement (retracing More), although 6% exceeded this limit (*[2*) and retraced back to earlier elements (*[1*):

*C6cBesan, 6, euh dans la mise en pratique *[1* je *vois* pas *[2* **c** - *[1* je *verrais* pas *[2* **comment** faire correspondre les choses

*Tropr102, 3,1, *[1* je crois que*[2* **C** + *[1* je crois que*[2* **c'est** la famille des N. mais j'en suis pas sûr

4. Conclusion

Repetitions and word-fragments described in the "standard" spoken French corpora, showed that these phenomena were very frequent and took an active part in the development of the statement through successive insistence on the syntactic places. Half of the involuntary truncations in the statements could be qualified as hesitations. If the function of the fragments, which would be followed by a modification, appeared clearly to be an error, that of unfinished word-fragments was less clear as the speaker did not confirm nor annul his production by retracing, and only this would reveal if it was an error or a hesitation.

Repetitions and interruptions of words are two distinct phenomena which do not affect the same syntactic places. We repeat more functional words than lexical words while we stop in the middle of a lexical item rather than a functional word. As regards the stammerers, Zellner [16, p482] noted that on the contrary, at the time of a hesitation, and not of a stammering, "whatever the rate of disfluencies, the monosyllabic functional words – which enables the speech to be structured – tend to be subjected to more accidents than the other words". However, we never observed, as was seen in 40% of the disfluencies produced by stammerers, "syllables gradually produced in several utterances" in our corpora, nor did we observe the completion of word-fragments without retracing to the beginning of the word. From time to time, although rarely, (once every 33 minutes), a word-fragment was repeated before the word was actually completed (this happened in 80% of the cases).

We studied the re-adjustments concerning where the speaker continued after having stopped in the middle of a word from two points of view: firstly, the types of insertions (30% of the cases) and secondly, the retracing of the statement (75% of the cases). Two types of insertions were distinguished: those which were non-syntactically linked to the statement and those which were. The first were observed, above all, immediately after the interruption, and sometimes, although rarely, when retracing started. Syntactically linked insertions were used when retracing started. These modulations consisted primarily of word additions or replacements while retracted elements were infrequent.

In the large majority of cases, word interruptions occurring in the lexicon, did not appear at the beginning of the nominal or verbal group. Retracing could occur on the interrupted word, go back to the beginning of the two groups or even exceed this limit. If the most numerous cases were those (as noted by Clark and Wasow, [5]) where the retracing of the statement went back to the beginning, a third of the retracing did not obey this schema. Nearly 20% was in fact "minimal" and did not go back to the beginning of the group whereas the others, on the contrary, exceeded the nominal and verbal group. The number of speakers was insufficient, but it seemed that the retracing (6 to 10 %) which exceeded the limits of the group where the interruption took place, did not depend on the localization of the latter.

5. References

[1] Bensalah, Y. 1997. *Pour une linguistique du bégaiement*. Paris, L'Harmattan

[2] Blanche-Benveniste C. 1997. *Approches de la langue parlée en français.* Paris, Edition Ophrys.

[3] Candéa M. 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).

[4] Cappeau P. 1998. Quelques mots sur quelques bribes liées au genre. In: Bilger M, Van den Eynde & Gadet F, (Eds) *Analyse linguitique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste.* Peeters. Leuven, Paris, pp. 301-311.

[5] Clark et Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology, 37,* pp 201-242

[6] Duez D. 2001. Signification des hésitations dans la production et la perception de la parole spontanée. *Revue Parole, 17-18-19,* pp 113-138.

[7] Van Hout A. 2002. *Les bégaiements. Histoire, psychologie, évaluation, variétés, traitements.* Paris, Masson, 2ème édition. (1997)

[8] Levelt W.J.M. 1989. *Speaking. From intention to articulation.* Cambridge, MIT Press.

[9] Pallaud B. 2002 a. Les amorces de mots comme faits autonymiques en langage oral. *Recherches Sur le Français Parlé, 17,* pp. 79-102.

[10] Pallaud B. 2003 a. Achoppements dans les énoncés de français oral et sujets syntaxiques. In Merle J.M. (Ed.). *Le Sujet*. Paris : Éditions Ophrys, Faits de Langue, pp. 91-104.

[11] Pallaud B. & Henry S. 2004. Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. In *Le poids des mots. Actes des 7èmes Journées Internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve, 10-12 mars 2004. Louvain, PUL, vol 2, pp.848-858.

[12] Pasdeloup V. 1992. A Prosodic Model for French Text-to-Speech Synthesis. A Psycholinguistic Approach. In BAILLY. Gérard; BENOIT, Christian; SAWALLIS,

Thomas R. (eds). *Talking Machines. Theories. Models, And Designs*, pp. 335-348

[13] Schegloff E., Jefferson H. and Sachs H.. 1977. The preference for self-correction in the organization of repair in conversation.. *Language, 53, 2*, pp. 351-382.

[14] Shriberg E.& Stolcke A..1998. How Far Do Speakers Back Up In Repairs? A Quantitative Model. *Proc. Intl. Conf. on Spoken Language Processing*. vol. 5, pp. 2183-2186, Sydney, Australia.

[15] Shriberg E. (1999). Phonetic Consequences of Speech Disfluency. Symposium on The Phonetics of Spontaneous Speech (S. Greenberg and P. Keating, organizers). *Proc. International Congress of Phonetic Sciences*. vol. 1, pp. 619-622, San Francisco.

[16] Zellner, B. (1992). Le bé- bégayage et euh … l'hésitation en français spontané. Actes des 19 èmes Journées d'Études sur la Parole, J.E.P. Bruxelles, pp. 481-487.