

La notation subjective de la pause constitue-t-elle un bon indice pour le découpage de textes oraux ?

Anne Dister¹

Université catholique de Louvain

Abstract

L'analyse syntaxique et morphosyntaxique automatisée se fait traditionnellement en isolant des portions de texte constituées par des phrases. Or, les études sur l'oral ont abandonné la notion de phrase et les données transcrites ne sont pas ponctuées. Dans cet article, nous tenterons de voir si les pauses notées subjectivement par les transcrip-teurs peuvent servir de base à un découpage du texte en unités plus petites que le tour de parole.

Keywords : morphosyntactic annotation, transcription, pause, subjective pause

Résumé

L'analyse syntaxique et morphosyntaxique automatisée se fait traditionnellement en isolant des portions de texte constituées par des phrases. Or, les études sur l'oral ont abandonné la notion de phrase et les données transcrites ne sont pas ponctuées. Dans cet article, nous tenterons de voir si les pauses notées subjectivement par les transcrip-teurs peuvent servir de base à un découpage du texte en unités plus petites que le tour de parole.

Mots-clés : annotation morphosyntaxique, transcription, pause, pause subjective

1. Introduction

Le présent article s'inscrit dans le cadre d'un travail plus large sur l'annotation morphosyntaxique de corpus de parole transcrite.

Les textes sur lesquels nous travaillons sont issus de la banque de données textuelles orales VALIBEL (<http://www.uclouvain.be/valibel>). Les transcriptions de l'oral se distinguent des données écrites standard en ce sens qu'elles notent une série de phénomènes propres à l'oral : amorces de morphèmes, répétitions, autocorrections, *euh*, etc. qui posent des problèmes aux analyseurs construits pour l'écrit. De plus, se pose la question du découpage préalable du texte avant son analyse automatique. En effet, traditionnellement, les analyseurs travaillent dans le cadre de la phrase, unité isolée grâce aux marques de la typographie. Or, pour des raisons théoriques largement développées dans Blanche-Benveniste et Jeanjean (1986), les transcriptions de l'oral

¹ Cental et Centre de recherche Valibel, anne.dister@uclouvain.be

ne sont pas ponctuées. Le seul découpage préalable est celui qui est effectué en fonction des prises de parole des locuteurs.

Dans cet article, nous envisageons les marques de la pause silencieuse présentes dans les transcriptions comme une possible base à un découpage du texte en unités plus petites, unités que nous ne considérerons pas comme des unités théoriques mais comme des segments pratiques qui peuvent être soumis à une analyse morphosyntaxique.

2. La pause silencieuse

La pause silencieuse participe pleinement à l'organisation temporelle des énoncés. Elle a déjà été maintes fois étudiées (*cf.* Zellner 1998 ou Duez 1997 pour un état de l'art), principalement parce qu'elle constituait déjà un objet de recherche quand les études sur la parole ne se concentraient que sur ce que l'on nomme communément la « parole de laboratoire » (principalement des phrases lues, isolées, dans des études de phonétique acoustique).

Nous appelons *pause silencieuse* une interruption perçue du continuum du signal sonore dans une prise de parole. Dans la littérature sur le sujet, la pause silencieuse est également appelée *pause non sonore* ou encore *pause vide*, par opposition à la pause dite pleine, en français le *eah*.

2.1. Les fonctions de la pause

La pause silencieuse, reprise parfois parmi les phénomènes de disfluency quand elle est perçue comme une marque d'hésitation, occupe des fonctions très différentes, tant dans la planification de l'énoncé par le locuteur que dans la compréhension de celui-ci par l'auditeur. La pause silencieuse permet au locuteur de reprendre son souffle, elle lui laisse le temps nécessaire pour planifier la suite de son énoncé, elle peut aussi avoir pour fonction de donner à l'interlocuteur le temps d'assimiler ce qui vient d'être dit, elle ménage un suspens (suspens traduit généralement par des points de suspension dans le texte des écrivains) ou vise encore à mettre en évidence la suite de l'énoncé (elle est parfois appelée *pause de focalisation* ou encore *pause stylistique*), etc.

Si la pause constitue pour le locuteur l'occasion de reprendre son souffle et de respirer, il a été montré que ce moment de respiration n'est pas imposé au locuteur : celui-ci le « choisit » lors de sa production de parole (Autesserre *et al.* 1992). Ainsi, en général, on ne respire pas au milieu d'un mot, par exemple. Les contraintes physiques et physiologiques, si elles interviennent inévitablement², ne sont pas prépondérantes. Comme le dit Duez (1997 : 276) :

² Selon Grosjean et Deschamps (1975 : 157), qui comparent l'anglais et le français, « la respiration est un besoin physiologique qui se fait sentir à intervalles réguliers indépendamment de la langue envisagée ». Ils constatent également que « plus les suites sonores sont longues, plus le pourcentage de pauses silencieuses contenant une pause de respiration est grand » (*ibidem*).

La production de la parole est un processus dynamique qui est une fusion d'un ensemble d'actions physiologiques et cognitives. C'est un processus extrêmement complexe qui met en jeu toute une série d'organes et de muscles et qui recouvre différentes fonctions : respiration, recherche lexicale et planification, structuration grammaticale et mise en relief des idées.

Boomer et Dittman (1962) distinguent deux types de pauses silencieuses fonctionnellement différentes : la première est la *pause démarcative*, ou *pause grammaticale*, dont la présence participe de la structuration du discours. Cette pause démarcative ne se place donc pas n'importe où dans un énoncé ; elle participe pleinement à la structuration syntaxique et sémantique de celui-ci, ainsi qu'au regroupement des « mots » en segments cohérents. L'autre type de pause relève quant à lui de la disflue dans la mesure où il indique une hésitation du locuteur et constitue une marque du « travail de formulation » de celui-ci (Morel et Danon-Boileau 1998). Contrairement aux pauses démarcatives, celles-ci se placent à des endroits non prédictibles, notamment à l'intérieur de syntagmes.

La **pause démarcative** peut être illustrée par l'exemple suivant :

chaBR1 c'était un vieux monsieur // qui élevait des moutons // qui avait un grand jardin // qui était pensionné mineur // qui travaillait encore avec sa pension // et que j'étais souvent avec lui // et un jour il était allé voir un film de cinéma (bruit) (...)
[chaBR1r]

Comme on le voit dans cet extrait, les pauses isolent des groupes syntaxiques qui correspondent à des propositions : une proposition principale, séparée d'une relative par une pause longue, cette proposition relative étant elle-même séparée par une autre pause longue d'une autre proposition relative juxtaposée, etc. Dans cette transcription, on constate un fonctionnement très régulier de la pause longue, qui structure véritablement l'énoncé en regroupant les mots en groupes cohérents, d'un point de vue syntaxique et sémantique.

La **pause non démarcative** peut être illustrée par l'exemple suivant :

accCTO et est-ce que justement la profession des des parents ça peut influencer **sur la** / **façon** dont les enfants vont parler avec plus ou moins d'accent [accBF1r]

Ici, la pause intervient à l'intérieur d'un syntagme, entre le déterminant défini *la* et le nom qui suit. Elle ne permet pas d'isoler de quelconques constituants. On remarquera par ailleurs que le tour de parole n'a pas ici d'autre pause que cette pause qui peut être considérée comme une marque de disflue. Ce type de pause silencieuse non démarcative est fréquemment appelé dans la littérature sur le sujet *pause (silencieuse) d'hésitation* (Goldman-Eisler 1968 :10). En effet, tout comme le *euh*, ces marques considérées comme disfluentes ont été assimilées à des phénomènes d'hésitation de la part du locuteur.

Ces pauses ont fait l'objet de nombreuses études (principalement pour l'anglais, cf. Duez 1997 pour un état de l'art), dont le but est souvent de dégager des critères qui les distingueraient des pauses démarcatives (syntaxiques) : ces critères concernent la

distribution des différents types de pauses, souvent en lien avec leur durée et leur fréquence dans l'énoncé, voire encore avec le type de discours.

Les études sur la pause silencieuse et ses fonctions ne manquent donc pas, auxquelles s'ajoutent encore, par exemple, celles qui étudient le lien entre syntaxe et intonation (*cf.* Rossi 1981). Mais notre objectif, s'il consiste également à étudier les pauses selon qu'elles sont ou non structurantes du point de vue grammatical, se situe dans une perspective relativement différente des études antérieures que nous avons consultées, tant sur le français que sur l'anglais. En effet, s'il s'agit bien de faire émerger une éventuelle différence entre pause démarcative et pause non démarcative, nous le faisons non pas sur base de données objectives – nous entendons par là établies à l'aide d'appareils acoustiques fins qui mesureraient les variables temporelles du français parlé spontané, noteraient des indices prosodiques, etc. – mais bien à partir de données secondaires. Celles-ci sont constituées par un texte transcrit, dans lequel les pauses sont notées subjectivement par des transpositeurs humains qui n'ont recours qu'à leur oreille³.

De plus, la plupart des recherches citées se placent dans la perspective de cerner les marques d'hésitation (la pause étant une marque parmi d'autres), et ce point de vue cognitif n'est pas abordé dans notre travail.

2.2. La perception de la pause

La pause ayant une réalité physique (un blanc dans un continuum de signal sonore), on pourrait s'attendre à ce que cette réalité soit objectivable, et donc facilement décelable, et cela de manière identique, par les récepteurs, qu'ils soient auditeurs ou transpositeurs. Or, on constate qu'il n'en est rien et les tests de perception effectués par certains chercheurs concordent sur ce point.

La quasi-totalité des études sur le rôle de la subjectivité de l'auditeur dans la perception de la pause ont été faites sur l'anglais, mais les résultats de deux études sur le français que nous avons consultées vont dans le même sens⁴.

Ainsi, Candea (2000a : 123), dans un test de perception qu'elle effectue dans le cadre de sa thèse, constate qu'aucune des pauses silencieuses de son corpus n'a été notée par la totalité des 30 auditeurs auxquels elle a soumis ses données sonores.

³ Notons que les transcriptions se font maintenant avec le logiciel d'aide à la transcription Praat (Boersma et Weenink 2006) dans lequel figure le spectrogramme, c'est-à-dire une représentation graphique de la chaîne sonore qui symbolise la fréquence, l'intensité et la durée. Dans celle-ci, on perçoit visuellement les interruptions dans le continuum sonore. On peut faire l'hypothèse que le transpositeur qui se base sur ces indices indépendants de la forme phonique (et donc des mots et des découpages syntaxiques basés sur la compréhension de ceux-ci) marque des pauses là où elles sont physiquement présentes.

⁴ Voir Candea (2000a : 112 et sv.) pour un bref résumé d'autres tests de perception sur les pauses et les phénomènes dits d'hésitation.

Duez (1991) constate quant à elle que les pauses situées entre des constituants sont mieux perçues que les pauses intraconstituant⁵, la longueur de la pause ayant une répercussion directe sur son identification : plus une pause silencieuse est longue, mieux elle est perçue par les sujets. En ce qui concerne les pauses subjectives (c'est-à-dire les pauses perçues par les sujets-auditeurs qui ne correspondent pas à une interruption réelle du signal sonore), l'auteure note que celles-ci sont plus fréquemment identifiées à des frontières de syntagmes qu'à des frontières de propositions ou de phrases. La présence d'un *eu*h ou d'un allongement vocalique induit souvent la présence d'une pause subjective :

Dans la plupart des cas, les pauses subjectives correspondent à un allongement final, qui correspond à son tour à la réalisation d'une frontière syntaxique (Duez 1991 : 140).

Les résultats du test de Candea (2000a) vont dans le même sens : les frontières grammaticales favorisent la perception de pauses subjectives. De même, les pauses subjectives sont davantage perçues par les auditeurs à l'endroit où sont présentes des marques du travail de formulation.

Duez (1991), dans les tests de perception de phénomènes d'hésitation qu'elle met en place, évite de travailler sur des données transcrites, mais place ses sujets uniquement face à la source sonore. En effet, selon elle,

(...) la connaissance du texte écrit, et donc de sa structure syntaxique, peut aussi influencer sur la détection des pauses. Il est même possible qu'elle privilégie la détection des pauses grammaticales (Duez 1991 : 105).

Duez (1997 : 285) critique à cet égard la procédure de Martin et Strange (1968), qui font marquer les hésitations directement sur le texte transcrit. Les auteurs obtiennent néanmoins des résultats qui corroborent ceux d'autres tests de perception. Ils constatent en effet que les auditeurs déplacent à des frontières syntaxiques des hésitations présentes à l'intérieur de syntagmes.

Candea (2000a), dont le test de perception montre également que les pauses syntaxiques (réelles) sont mieux perçues que les pauses non syntaxiques, va dans le même sens que Duez, généralisant les propos de celle-ci au-delà du support écrit :

En fait, quelle que soit la procédure utilisée pour tester la perception des pauses, il semble que, de façon générale, tous les résultats issus de tests qui attirent explicitement l'attention des sujets sur le phénomène pausal concordent sur ce point : les frontières syntaxiques (limites de propositions grammaticales, limites de syntagmes, etc.) induisent la perception d'une pause subjective ou favorisent la perception d'une pause réelle située dans une telle configuration (Candea 200 : 113).

D'une certaine façon, leurs résultats vont dans le sens des remarques de Boomer et Dittman (1962), selon lesquels les pauses grammaticales (c'est-à-dire démarcatives) sont attendues par les auditeurs.

⁵ Notons par ailleurs que Duez, qui s'intéresse à la parole des hommes politiques, recense un certain nombre de pauses intraconstituant dans son corpus, pauses qui correspondent à des pauses stylistiques typiques de ce type de discours.

Comme nous l'avons dit, dans les transcriptions sur lesquelles nous travaillons, l'interruption du signal sonore perçue comme une pause est jugée intuitivement par le transcripateur, et non mesurée à l'aide d'appareils acoustiques. Il s'agit donc dans notre corpus d'une marque subjective⁶, laissée à l'appréciation du transcripateur⁷. Celui-ci prend en compte plusieurs facteurs, dont le principal est sans doute la vitesse d'articulation⁸. La pause est donc jugée relativement à la vitesse d'articulation et, ainsi, pour une même durée de blanc dans le continuum sonore, une pause d'une durée pourtant équivalente sera ou non marquée par le transcripateur selon que le locuteur parle plus ou moins vite.

Comme la notation de la pause est subjective, il nous semble possible de formuler l'hypothèse, à la suite des constatations de Bommer et Dittman (1962), Duez (1991 et 1997) et Candea (2000a), que les pauses présentes dans notre corpus ont été notées parce que justement perçues comme pertinentes par le transcripateur⁹, au regard de la syntaxe ou de la morphosyntaxe. Nous faisons l'hypothèse que les pauses ont été transcrites à des endroits préférentiels en termes de structuration de l'énoncé, de regroupement de segments homogènes syntaxiquement et sémantiquement. Et que donc, elles seraient de bonnes candidates à un découpage du texte.

2.3. Le degré des pauses : corrélation entre durée et fonction

Nous l'avons dit, les études sur les fonctions de la pause, démarcative vs non démarcative, ne manquent pas. Mais force est de constater que, selon les auteurs, les résultats des recherches divergent. Selon certains (Boomer 1965), la pause d'hésitation a une durée plus longue que la pause grammaticale. D'autres (cf. Duez 1997 : 280) ne voient pas de distinction significative entre la durée des pauses, selon qu'elles marquent une hésitation ou structurent le discours en constituants. Ainsi, pour Campione et Véronis (2004), la pause seule n'est jamais une marque d'hésitation, sauf si elle est très brève (en dessous de 200 ms). Pour remplir cette fonction d'hésitation, elle doit obligatoirement être associée à une autre marque : soit un allongement vocalique, soit un *euh*, soit la combinaison de l'allongement et du *euh*. Les auteurs rejoignent ici Danon-Boileau et Morel (1998), qui estiment qu'en dessous de ce seuil de 200 ms, la pause est en général de nature respiratoire.

⁶ Dans le sens où elles sont laissées à l'appréciation du transcripateur ; nous ne prétendons pas que toutes les pauses sont subjectives au sens de Duez (1991) et de Candea (2000a) (c'est-à-dire non présentes dans le signal sonore).

⁷ Toutes les transcriptions ont fait l'objet d'une validation par un second expert, et il est vrai que sur ce point, les jugements divergent régulièrement.

⁸ Lane et Grosjean (1973) ont montré « qu'un sujet qui articule rapidement ne présente pas automatiquement un temps de pause réduit et vice versa » (cité par Grosjean et Deschamps 1975 : 156).

⁹ On se garde néanmoins d'assimiler ici le transcripateur à un simple auditeur « naïf ». Mais les auditeurs dont il est question dans les recherches des auteurs cités ne sont pas non plus tout à fait naïfs dans la mesure où il leur était clairement demandé de se concentrer sur le phénomène pausal ou sur les marques d'hésitation.

Grosjean et Deschamps (1972), dans leur étude sur les variables temporelles du français spontané, s'intéressent à la distribution syntaxique des pauses non sonores. Ils ont étudié un corpus de 30 interviews d'une durée minimale de trois minutes, interviews reprises d'émissions radiophoniques¹⁰, et ne prennent en considération que les pauses dont la durée est supérieure à 0,25 secondes. Pour leur analyse de la distribution des pauses, les auteurs découpent les énoncés en phrases (« ensemble formé de SN et de SV ») et en syntagmes, et étudient la répartition des pauses en fonction de ce découpage. Sans faire de distinction entre les trois types de pauses qu'ils mentionnent (pause de respiration, pause stylistique et pause d'hésitation), ils constatent¹¹ que 60,46 % des pauses non sonores se situent en fin de phrase,

c'est-à-dire au point de rupture grammatical de la chaîne parlée le plus normal. (...) la pause en fin de phrase est tout naturellement considérée comme le point de rupture naturel qui permet la reprise d'air et (...) la pause à l'intérieur de la phrase aurait plutôt tendance à être une pause d'hésitation ou une marque stylistique (Grosjean et Deschamps 1972 : 146-147).

Les auteurs obtiennent des pauses qui se situent dans 23,36 % des cas à des « coupures grammaticales » (p. 146) (après ou avant un complément de phrase, entre SN et SV ou entre des mots coordonnés ou apposés). Restent enfin 16,18 % de l'ensemble des pauses qui se situent « à l'intérieur d'un groupe de sens (soit à l'intérieur de SN soit à l'intérieur de SV) » (p. 147). Le SV étant composé du verbe et de ses compléments, une pause entre un verbe et un complément d'objet est donc considérée comme marquant une rupture à l'intérieur d'un groupe de sens (nous verrons que nous ne considérons pas qu'il y a rupture dans ce cas). Les auteurs ne mentionnent aucunement les occurrences où la pause pourrait marquer l'abandon de la construction, sauf à considérer que les faux départs recensent tous ces cas¹².

Grosjean et Deschamps précisent par ailleurs que 74,64 % de leurs pauses sont des pauses de respiration (regroupées ou non avec une pause d'hésitation ou une pause stylistique), sans faire de lien explicite entre le découpage syntaxique et le type de pause (de respiration ou de non-respiration).

Des études antérieures, notamment sur la lecture oralisée, ont montré la corrélation qui existe entre la durée de la pause silencieuse et la position syntaxique de celle-ci. Ainsi par exemple, Vannier et *al.* (1999) distinguent trois positions dans le texte lu – intérieur de phrase, entre deux phrases et entre deux paragraphes – et constatent des moyennes croissantes de durée de pauses pour ces trois positions.

¹⁰ Dans leur étude de 1973 sur des descriptions de dessins humoristiques, ils obtiennent des résultats comparables de distribution syntaxique de la pause non sonore. Alors qu'ils qualifient la description de « langage hésitant » (Grosjean et Deschamps 1973 : 210) – car selon eux plus contraignant pour le locuteur que l'interview –, l'une des explications qu'ils avancent est que, dans le cas des descriptions, « les phrases (...) sont nettement moins complexes et plus courtes » (Grosjean et Deschamps 1973 : 213).

¹¹ Les auteurs présentent une répartition détaillée des différents cas au tableau IV de la page 148.

¹² Dans leur étude comparative français-anglais (1975), ils distinguent les faux départs repris des faux départs non repris.

À partir de données orales, Duez (1991) distingue quant à elle quatre positions – intraconstituant, interconstituant, fin de proposition et fin d'énoncé – et observe également une corrélation entre durée de la pause et frontière syntaxique : les pauses sont plus longues aux frontières syntaxiques majeures (propositions, phrases) qu'aux frontières de syntagmes. L'auteure observe également une variation importante de la durée et de la fréquence des pauses en fonction du type de parole, de la situation de communication. Ainsi, dans la parole des hommes politiques qu'elle étudie, elle constate une divergence entre d'une part les entrevues amicales et politiques et d'autre part les discours, ces derniers ayant une durée totale de pauses 50 % plus élevée que les premiers.

Campione et Véronis (2004), qui ont analysé un corpus de parole spontanée de 54 minutes (correspondant à 8500 mots), relèvent seulement 71 % de pauses démarcatives (celles-ci étant repérées grâce à des indices convergents (intonation, syntaxe, etc.)). Ce chiffre est faible si l'on considère que la fonction démarcative est la principale fonction de la pause. Selon ces chiffres, on peut s'attendre à ce que les pauses notées dans notre corpus ne soient pas de bonnes bases à un découpage du texte dans plus d'un cas sur quatre.

Nous l'avons dit ci-dessus, les auteurs constatent que les pauses très brèves (en dessous de 200 ms) ne sont jamais démarcatives. Ils en ont très peu dans le corpus : 24 sur 1163 pauses. En fait, s'ils ont également trois degrés de pauses, on s'aperçoit que ceux-ci ne sont pas comparables aux nôtres, et que nous ne pouvons donc pas extrapoler leurs résultats en ce qui concerne la longueur des pauses.

En fait, à la lecture des études antérieures sur le sujet, nous sommes confrontée à un double problème :

- nous n'avons aucune mesure de durée des pauses, simplement des degrés jugés intuitivement par les transcripateurs ;
- les auteurs consultés ne distinguent pas clairement *hésitation* et *structuration syntaxique*. Ce qui nous intéresse ici n'est pas d'aborder la pause en tant qu'éventuelle marque d' hésitation, appartenant au travail de formulation du locuteur, mais bien de l'analyser en termes de structuration syntaxique de l'énoncé. Or, une pause peut très bien noter une hésitation tout en structurant l'énoncé et en n'introduisant pas de rupture syntaxique de celui-ci. Mais généralement, les auteurs consultés ne font pas le départ entre ces deux points de vue. En tout état de cause, la pause, qu'elle signifie ou non l' hésitation du locuteur, ne nous intéresse que dans le cadre strict de la structuration syntaxique des énoncés.

2.4. La répartition des pauses dans le corpus

Les conventions VALIBEL (Dister *et al.* 2006) distinguent trois degrés de pauses silencieuses, qui sont évalués subjectivement par les transcrip-teurs :

- la pause brève (notée par une barre oblique : /)
- la pause longue (notée par deux barres obliques : //)
- la pause prolongée, qui correspond à un silence (notée entre parenthèses : *(silence)*).

Dans notre corpus, composé de 430 000 mots, la répartition des trois degrés de pauses est la suivante :

marque	nombre d'occurrences n=	pourcentage sur l'ensemble des pauses
silence	2178	7,1 %
pause longue	7377	23,9 %
pause brève	21 316	69 %

Tableau 1. Répartition des trois degrés de pauses

On constate que, parmi les trois types de pauses, les pauses brèves sont les plus nombreuses, avec 69 % de l'ensemble des pauses du corpus. Viennent ensuite les pauses longues et enfin les silences. Notons que les pauses brèves et longues auraient pu obtenir un score encore plus élevé si nos conventions préconisaient de les indiquer aussi en fin de tour de parole et non pas uniquement à l'intérieur de la prise de parole d'un locuteur unique¹³.

Les trois degrés de pauses confondus, cette indication d'une pause silencieuse se présente dans nos transcriptions en moyenne tous les 14,35 mots graphiques. Il s'agit donc d'une information qui apparait fréquemment dans les textes, et nous voudrions maintenant voir concrètement de quelle manière elle peut servir d'indice à un éventuel découpage du texte en unités plus petites.

Étant donné ce nombre important d'occurrences, il était totalement impossible de tirer de quelconques conclusions qui se seraient basées sur une relecture, qui ne pouvait qu'être plus ou moins attentive, de concordances. Nous avons donc choisi de nous baser sur l'étude de trois textes, afin de voir comment les pauses brèves et longues y étaient distribuées.

3. Étude sur un sous-corpus

¹³ Seuls les silences sont notés en fin de tour de parole.

Le sous-corpus que nous avons choisi est composé de trois textes : deux conversations informelles (blaAT1l et famVV1r) et une entrevue (accPH1r). Voici la répartition des pauses pour ces trois textes et pour le sous-corpus qu'ils constituent :

corpus	Nombre de mots	silence	pause longue	pause brève	longueur moyenne entre 2 pauses silencieuses, en nombre de mots
blaAT1l	8177	33	62	450	15
famVV1r	4220	19	77	194	14,55
accPH1r	8111	41	423	287	10,80
TOTAL	20 508	93	562	931	12,93

Tableau 2. Répartition des pauses dans le sous-corpus (valeur absolue)

On remarquera que accPH1r contient plus de pauses longues que de pauses brèves, inversement à la tendance générale du corpus global où les trois degrés de pauses connaissent un nombre décroissant d'occurrences, de la pause brève au silence, en passant par la pause longue.

Cette inversion de la tendance générale rencontrée dans accPH1r concerne en tout six textes dans notre corpus global, qui en compte 60.

La longueur moyenne entre deux pauses du corpus famVV1r (14,55) est très proche de celle de l'ensemble du corpus (14,35). Remarquons que dans ce calcul, nous globalisons à l'ensemble d'un texte, sans faire de distinction selon les locuteurs.

Le sous-corpus ainsi constitué comprend donc au total 1586 pauses, dont la répartition est la suivante selon les trois degrés de pauses :

silence	pause longue	pause brève
5,86 %	35,44 %	58,70 %

Tableau 3. Répartition des pauses dans le sous-corpus (pourcentage)

L'objectif spécifique de cette étude du sous-corpus est de voir si la pause, selon son degré, est présente à des endroits pertinents pour le découpage du texte transcrit :

1. La pause est-elle située à l'intérieur d'un constituant (suspend-elle momentanément le déroulement syntagmatique de l'énoncé, qui se poursuit ensuite « normalement ») ou bien se place-t-elle préférentiellement entre deux constituants ? Pour le dire autrement : la pause est-elle démarcative ou non ?

2. La pause constitue-t-elle une rupture dans la construction syntaxique de l'énoncé ? En d'autres termes, intervient-elle à un moment où le déroulement syntagmatique de l'énoncé est interrompu et laisse inachevée la structure syntaxique entamée ?
3. La pause a-t-elle une fonction différente selon qu'elle est ou non combinée à une autre marque ou qu'elle apparait dans certaines positions de l'énoncé ?
4. La pause a-t-elle un comportement différent selon son degré (pause brève, pause longue et silence) ?

3.1. Combinaison de la pause et d'autres marques

Parmi ce que nous considérons comme une autre marque ou une autre position pertinente dans l'énoncé, nous avons repris¹⁴ : les amorces de morphèmes, les autocorrections immédiates, les répétitions où la pause est interne¹⁵, les onomatopées et les ponctuants, les *euh*, les indications paraverbales et les passages incompréhensibles, les positions de fin et de début de chevauchement et de fin de tour de parole, cette position « fin de tour de parole » étant uniquement possible pour les silences.

Parmi les 1586 pauses du sous-corpus, on constate que 888 de ces marques sont contigües à la pause. Nous envisageons ici seulement des cooccurrences immédiates, et non des séquences de ces marques¹⁶. Voici les résultats des marques contigües, classées par ordre de fréquence décroissante et indépendamment du contexte gauche ou droit de la marque :

marque	n=
<i>euh</i> dit d'hésitation	170
ponctuants	146
fins de chevauchements	134
débuts de chevauchements	121
onomatopées	110
répétitions	83
fins de tour de parole	45
indications paraverbales	36
amorces	32
passages incompréhensibles	7
autocorrections	4

Tableau 4. Types de marques contigües à la pause

¹⁴ Voir Dister (2007) pour une définition de ces phénomènes.

¹⁵ Et non les répétitions qui précéderaient ou suivraient la pause.

¹⁶ Cela signifie que nous prenons en considération les marques strictement contigües.

Les marques ou positions considérées ici appellent immédiatement quatre remarques :

1. Nous ne pouvons prendre parmi les marques l'allongement vocalique, dont les études antérieures ont montré la fréquence de cooccurrence avec la pause, et le rôle dans la perception des pauses subjectives (Duez 1997 : 288).

2. L'ordre de fréquence de ces marques est intéressant, puisqu'on constate un classement différent dans la combinaison avec la pause. En effet, le *euh*¹⁷ est plus fréquent que les répétitions, ce qui n'est pas le cas de manière globale dans le corpus.

3. Certaines marques contraignent l'analyse que nous faisons de la pause en termes de position intra ou interconstituant, ainsi qu'en termes de rupture syntaxique de l'énoncé. Ainsi,

3.1. la position « fin de tour de parole » est de toute façon considérée comme une fin d'unité (Dister 2007), la notation du silence étant du point de vue du découpage en unités redondante ;

3.2. Les amorces de mots, les indications paraverbales et les passages incompréhensibles sont également considérés comme des endroits où l'analyse des grammaires doit s'arrêter (Dister 2007). La présence de la pause est donc ici également redondante par rapport aux choix de découpages déjà faits pour ces marques.

4. La pause silencieuse est plus fréquente en fin de chevauchement (134 occurrences) qu'en début de chevauchement (121 occurrences). Cela semble étonnant. En fait, une hypothèse qui peut être avancée pour les paroles superposées est que l'interlocuteur se saisit de la pause laissée par le locuteur pour tenter de prendre la parole, comme le suggère Oostdijk (2003 : 64) :

it (...) has been observed that in most communicative situations the time available for pausing is limited as there is the risk of losing the floor to another speaker or losing the audience's attention.

Si cette hypothèse s'avérait exacte, on devrait s'attendre à un plus grand nombre de pauses qui précèdent le début du chevauchement. Or, on constate pourtant que ce nombre est inférieur au nombre de pauses qui suivent la fin du chevauchement.

3.2. Contexte des marques combinées

contexte par rapport à la pause	<i>euh</i>	ponctuant	onomatopée	amorce
avant la pause	146	93	63	21
après la pause	24	53	47	11

Tableau 5. Contexte des marques combinées à la pause

¹⁷ Campione et Véronis (2004) notent également une forte interaction de la pause silencieuse et de la pause pleine : 33 % des pauses silencieuses sont en contact avec une pause pleine, et 58 % des pauses pleines cooccurrent avec une pause silencieuse. Ils comptent parmi la pause pleine l'allongement vocalique.

On le voit, pour les quatre marques envisagées, la position avant la pause est beaucoup plus fréquente qu'après la pause. Toutes marques confondues, celles-ci se placent dans 71 % des cas dans le contexte gauche de la pause, contre seulement 29 % des cas dans le contexte droit.

Selon Candea (2000a), la pause silencieuse seule et la pause silencieuse subséquente à une marque du travail de formulation – chez l'auteure : *euh*, allongement vocalique, répétition ou auto-correction immédiate – ont des fonctionnements très différents et elle y voit deux types de pauses. Pour elle, la configuration *marque de travail de formulation* + *pause silencieuse* forme une unité (comme si la pause était seule), ce qui n'est pas le cas de la combinaison *pause silencieuse* + *marque de travail de formulation*. Candea considère comme seul pertinent le contexte gauche de la pause pour classer celle-ci parmi les pauses non structurantes. Opposée à la pause non structurante, la pause structurante se définit formellement comme

- une pause silencieuse entre deux séquences de parole continue d'un même locuteur

ou comme

- une pause silencieuse précédée d'une marque du travail de formulation (à gauche donc, et non à droite comme pour la pause non structurante).

La durée de la pause silencieuse structurante est plus brève que celle de la pause subséquente à une marque de travail de formulation. Dans ces deux configurations, la pause définie formellement comme structurante sépare des constituants ; l'éventuelle présence d'une marque du travail de formulation subséquente à la pause indique le début du constituant suivant auquel elle appartient.

Ainsi, le terme *structurant* chez Candea n'a pas de correspondance stricte avec le rôle démarcatif tel que l'entend Bommer (1962) par exemple, et tel que nous l'entendons ici. Nous assimilons en fait *démarcatif* à *structurant du point de vue syntaxique* (excluant de démarcatif la fonction stylistique des pauses intraconstituant par exemple). Candea a donc des pauses qu'elle appelle structurantes en position intraconstituant. En fait, nous l'avons déjà dit, Candea s'intéresse aux phénomènes dits d'hésitation ; quant à nous, c'est l'aspect purement syntaxique des choses que nous souhaitons prendre en compte, ce qui rend les résultats des études antérieures souvent inutilisables pour nous.

De plus, la définition de ce qu'est l'unité (le constituant) syntaxique varie d'un auteur à l'autre, ce qui rend donc variable l'appréciation en termes de rupture syntaxique, du point de coupure grammatical (*cf.* principalement les divergences avec les études de Grosjean et Deschamps). Ainsi, Candea (2000a :170) considère comme une pause intraconstituant la pause placée entre un nom et son expansion, même si cette expansion est de type complément du nom (et pas seulement un adjectif épithète).

3.3. Pause marque de rupture et pause intraconstituant

Le tableau suivant reprend¹⁸, pour les trois degrés de pauses, le nombre d'occurrences (colonne n=) pour lesquelles la construction est soit rompue (le syntagme est laissé inachevé, colonne « rupture »), soit interrompue momentanément à l'intérieur d'un constituant puis achevée (colonne « intraconstituant »). Le pourcentage est calculé sur l'ensemble des pauses de même degré dans le corpus.

	rupture		intraconstituant	
	n=	%	n=	%
pause brève	37	3,97 %	100	10,74 %
pause longue	7	1,25 %	47	8,36 %
silence	3	3,23 %	2	2,15 %
total	47	2,96 %	149	9,39 %

Tableau 6. Répartition des pauses intraconstituant et des pauses à l'emplacement d'une rupture

Les pauses brèves sont les plus nombreuses à se placer non seulement à l'endroit d'une rupture syntaxique, mais également en position intraconstituant ; elles sont suivies de loin par les pauses longues et enfin par les silences, classement qui correspond à l'ordre de fréquence de ces pauses dans le corpus. Pour les trois degrés de pauses, les cas de rupture syntaxique sont moins fréquents que les occurrences de pauses placées à l'intérieur d'un constituant (au total, 47 cas de rupture et 149 cas de pauses intraconstituant).

Si l'on considère non pas le nombre d'occurrences mais la fréquence de chaque type de pause sur l'ensemble des pauses de même degré (la colonne pourcentage), on voit que l'analyse diverge sensiblement. En ce qui concerne la rupture syntaxique, les silences sont presque aussi nombreux que les pauses brèves (3,23 % contre 3,97 %), la pause longue ne se positionnant à l'emplacement d'une rupture syntaxique que dans 1,25 % de ses occurrences en corpus. Pour ce qui est de la pause intraconstituant, la fréquence reste la même, la pause brève se plaçant toujours au 1^{er} rang avec 10 % de ses occurrences qui arrivent à l'intérieur d'un même constituant.

3.4. La pause à l'emplacement d'une rupture syntaxique

En excluant les pauses cooccurentes à une autre marque qui constitue un arrêt dans notre procédure de découpage du texte (*cf.* ci-dessus), on recense 47 cas de rupture syntaxique à l'endroit d'une pause. Nous choisissons la formulation relativement neutre « à l'endroit d'une pause », car nous ne poserons pas d'hypothèse en ce qui

¹⁸ Nous ne considérons pas comme un point de rupture dans la linéarité syntaxique de l'énoncé les pauses contiguës à une marque déjà reconnue comme stoppant la construction, c'est-à-dire certaines amorces de morphèmes, les indications paraverbaux et les passages incompréhensibles. Ce cas de figure se présente à 75 reprises dans le corpus et les occurrences ne sont donc pas comptabilisées ici.

concerne la raison de cette rupture syntaxique (la pause en marque-t-elle la place ou en est-elle en partie la cause, par exemple).

Pour les indications *silence*, on voit que trois occurrences marquent une rupture syntaxique. Elles appartiennent toutes au texte blaAT1l, et deux ruptures se produisent en fin de tour de parole, après une conjonction¹⁹ :

blaAT1 ouais c'est c'est la folie quoi non mais c'est c'est c'est dingue d'être aussi euh / inconscient et puis / enfin tu vois j'ai l'impression que c'est je vivais que la nuit quoi tu vois // en même temps je crois que c'était pendant l'hiver donc à mon avis c'est cette image-là / qui qui m'est restée mais enfin non pas rien que l'hiver **mais (silence)**

blaAT1 en plus il y a un -| bébé mais non mais / je veux dire euh là c'est / je suis vraiment je suis vraiment bien avec lui quoi enfin c'est clair que **(silence)** oui monsieur oui madame (en parlant fort et en s'adressant à l'enregistreur)

La rupture syntaxique se produit à l'emplacement d'une pause longue à sept reprises. Ici aussi, toutes les occurrences apparaissent dans un même texte, accPH1r, dont on a vu qu'il comptait plus de pauses longues que de pauses brèves. Devant de tels regroupements des occurrences, on peut formuler deux hypothèses quant aux différences de résultats observées, même devant ce nombre d'occurrences réduit :

1. on est devant une variation interlocuteur ;
2. la variation est moins le fait des locuteurs que de la perception du transcripteur.

En fait, il nous semble plausible que les deux facteurs interviennent de façon concomitante, et nous n'avons pas eu recours au signal sonore pour vérifier, l'objectif étant de travailler uniquement sur base des données transcrites.

En ce qui concerne la pause brève, on constate qu'elle est davantage répartie sur l'ensemble des trois textes (6 pour famVV1r, 9 pour accPH1r et 22 pour blaAT1l).

Une typologie des cas de ruptures syntaxiques est très difficile à établir, et il semble bien, au vu de l'analyse de nos données, que celles-ci puissent revêtir des formes très variées, comme le montrent les exemples ci-dessous :

blaAT1 (...) ben si il arrive à à survivre à / à des cours du soir plus un travail à temps plein euh il arrivera **à / parce que** il y a beaucoup moins de cours euh du soir euh (silence) [blaAT1l]

blaAT1 ouais -| on allait chez elles ouais c'était gai / non c'était gai / c'était un peu nostalgique tu vois **c'est marrant de / il y avait** Nadia Lamantia tu vois [blaAT1l]

accCT0 et vous savez donner un exemple (silence) |- / ça n'est <accPH1> m a/ -| pas grave non **je / donnez** un exemple [accPH1r]

famVI1 merci (rire) / je n'ai pas été à Pattaya / quand je suis débar/ quand j'ai débarqué **de / / de** c'était où ? / du Laos / quand j'ai |- repris <famVV1> oui -| le bus pour revenir à / à Bangkok (...) [famVV1r]

¹⁹ Notons que 48 % des occurrences de la marque (*silence*) du sous-corpus se produisent en fin de tour de parole.

Parmi les quatre exemples repris ci-dessus, les deux premiers illustrent une interruption après une préposition. Dans l'exemple (1), une conjonction suit l'interruption, tandis que dans l'exemple (2) le locuteur redémarre avec un présentatif. Dans l'exemple (3), la coupure se produit après un pronom personnel sujet, avec une reprise par un impératif. Dans la définition que nous faisons du constituant et en fonction de nos procédures de levée d'ambiguïtés (Dister 2007, chap. 6), ces ruptures qui se produisent après une marque de début de groupe (une préposition, un pronom personnel sujet non ambigu) sont évidemment particulièrement gênantes.

Le dernier exemple montre une pause entre les deux termes d'une répétition de la préposition *de*, celle-ci étant suivie d'une auto-correction (la contraction *du*) avec, entre les deux, l'insertion d'une marque explicite de recherche dans la formulation *c'était où ?* Ces marques sont difficiles à répertorier de manière exhaustive, tant elles peuvent prendre des formes variées.

En fait, dans la tâche d'étiquetage, ces inachèvements semblent bien constituer le problème majeur, et cela d'autant plus qu'il est très difficile de les repérer formellement. En effet, on l'a vu ci-dessus, les cas de rupture ne sont pas préférentiellement marqués par la conjonction de la pause et d'une autre marque. Si c'est le cas pour l'exemple (4) avec la répétition, il ne s'agit pas d'une majorité de nos occurrences, puisqu'une minorité de cas (17 cas, soit 36,17 %) est concernée par une marque mixte. Les autres cas, presque deux fois plus nombreux, sont uniquement marqués par la présence de la pause seule. De plus, si la combinaison de la pause et de la répétition marque bien en ce cas précis une rupture de la construction de l'énoncé, il s'agit en fait d'une exception à la tendance générale, laquelle montre qu'il n'y a pas de rupture dans cette configuration.

La répartition des marques mixtes, elle aussi très variée, est la suivante :

- avant la pause : 2 *euh* et 3 ponctuants ;
- après la pause : 3 ponctuants ;
- 3 pauses entre les termes d'une répétition ;
- des séquences qui combinent la pause et 2 autres marques : 3 *euh* avant et un ponctuant après la pause ; 2 répétitions et un ponctuant après la pause ; une suite qui combine autocorrection, pause, *euh* et ponctuant.

En fait, au vu de nos données, il apparaît qu'on ne peut considérer l'accumulation comme un indice certain de rupture.

Inversement, on recense également de nombreux cas d'inachèvement de la structure entamée qui sont non « marqués », même par une pause. Cela se produit à deux reprises dans l'extrait ci-dessous, chaque fois après le pronom sujet *je*. Il nous semble que l'on pointe là une caractéristique langagière de la locutrice :

iljDV1 euh / d'abord **je en** règle générale **je on** a je trouve une assez mauvaise diction / enfin prononciation (...) [iljDV1r]

On est donc bien là, nous semble-t-il, devant une des pierres d'achoppement des traitements automatisés sur les transcriptions de l'oral.

3.5. La pause intraconstituant

Le tableau ci-dessus montre que 9,39 % des pauses du corpus, indépendamment de leur longueur, sont placées à l'intérieur d'un constituant. Ce chiffre est relativement élevé²⁰, et, dans l'optique d'étiquetage automatique qui est la nôtre, il constitue un vrai problème. En effet, si l'unité « constituant » telle que nous l'avons définie peut être coupée par une pause pratiquement une fois sur 10, cela rend le système nettement moins performant. Peut-être faut-il envisager de ne pas considérer la pause – du moins la pause brève et la pause longue qui ont respectivement 10,74 % et 8,36 % de leurs occurrences à l'intérieur d'un constituant – comme une marque qui permet d'isoler d'éventuels groupes syntaxiques homogènes. Le silence, lui, resterait alors le seul bon indice de découpage, puisqu'il se place préférentiellement à des frontières majeures (*cf.* ci-dessus, nos résultats qui corroboraient ceux de Deschamps et Grosjean 1972)

Voyons maintenant le détail de la répartition des 149 pauses intraconstituant. Nous avons relevé 11 catégories différentes concernées par ce phénomène. Dans notre corpus, la pause intervient entre :

- un déterminant et un nom : 2,01 % (3 occurrences) ;
- un nom et un adjectif : 3,35 % (5 occurrences) ;
- une préposition et un groupe nominal : 4,69 % (7 occurrences) ;
- un pronom sujet et un verbe : 1,34 % (2 occurrences) ;
- un pronom préverbal (autre que sujet) et un verbe : 1,34 % (2 occurrences) ;
- l'auxiliaire *être* et le participe passé : 2,68 % (4 occurrences) ;
- l'auxiliaire *avoir* et le participe passé : 0,67 % (1 occurrence) ;
- un pronom relatif et la proposition relative : 0,67 % (1 occurrence) ;
- une forme en *que*²¹ et la proposition : 28,19 % (42 occurrences) ;
- *et* conjonction de coordination et la suite coordonnée : 55,03 % (81 occurrences) ;
- une conjonction de subordination et la proposition : 0,67 % (1 occurrence).

La classification établie ci-dessus appelle certaines remarques :

1. nous considérons comme un constituant le groupe formé par un pronom préverbal sujet et par un verbe conjugué, alors que le sujet et le verbe forment deux syntagmes différents. Dans ces derniers cas, la pause qui les sépare n'est d'ailleurs pas

²⁰ Il l'est moins que chez Campione et Véronis (2004), mais nous avons vu que l'identification des pauses ne se fait pas de la même façon et que nos données ne sont pas comparables.

²¹ Nous avons ici regroupé sous la proforme *que* les *que* conjonction de subordination, les formes comme *parce que*, etc.

comptabilisée dans les données ci-dessus. Nous faisons un traitement différent de ces cas, car les grammaires qui traitent la séquence « pronom préverbal + verbe » sont très rentables (*cf.* Dister 2007, chap. 6), et cela d'autant plus dans des corpus oraux, où l'utilisation des pronoms connaît une fréquence particulièrement élevée (Dister 2007, chap. 2). C'est la raison pour laquelle nous considérons cette séquence comme un syntagme élémentaire, de même que le syntagme « pronom préverbal non-sujet + verbe », pour lequel nous avons aussi une pause intraconstituant ;

2. le groupe nominal est peu concerné par la rupture : 3 occurrences entre le déterminant et le nom et 8 entre le nom et l'adjectif ; s'ajoutent encore 7 occurrences qui séparent la préposition du GN qu'elle introduit ;

3. de même, le groupe verbal est peu concerné par la pause, avec 9 occurrences en tout ;

4. les cas de coupure les plus fréquents rencontrés concernent les mots introducteurs (conjonctions ou pronoms) et les propositions que ceux-ci introduisent, avec 84 % des occurrences.

Si l'on exclut ces derniers cas, il reste en fait 24 occurrences de ruptures intraconstituant, ce qui, sur l'ensemble du corpus testé, est finalement un chiffre très faible. Ainsi, hormis ces cas, on pourrait considérer que la pause, dans nos transcriptions, se place donc préférentiellement à des endroits pertinents, et permet d'isoler des groupes syntaxiques homogènes.

La distribution de ces 24 pauses intraconstituant, selon la durée de celles-ci, est la suivante :

pause brève	pause longue	silence
17	7	0

Tableau 7. Distribution des pauses intraconstituant

On le voit ici encore, la pause brève est la plus fréquemment concernée ; vient ensuite la pause longue, avec 7 occurrences. Aucun silence dans le corpus ici testé ne se positionne à l'intérieur d'un constituant. Nous avons vu ci-dessus, sur l'ensemble du corpus, que ces configurations pouvaient néanmoins se présenter, même si elles sont relativement rares.

En ce qui concerne la pause longue, 5 des 7 occurrences apparaissent dans l'entrevue accPH1r, dont on a vu que le nombre de pauses longues était supérieur au nombre de pauses brèves, et cela contrairement à la tendance générale du corpus global. On a néanmoins ce cas de figure dans chacun des deux autres textes, une fois entre l'auxiliaire *avoir* et le participe passé, et une fois entre le nom et l'adjectif.

Pour l'entrevue accPH1r, on notera tout de même que le comportement des pauses selon leur degré ne déroge pas à la tendance générale. En effet, le nombre de pauses

brèves intraconstituant est supérieur au nombre de pauses longues intraconstituant (8 occurrences contre 5).

Si nous reprenons notre hypothèse de départ, inspirée de Candea (2000a), selon laquelle la pause serait préférentiellement intraconstituant si elle est combinée à une marque de disflue, on constate que peu de cas répondent à ce critère formel. En effet, sur ces 23 occurrences, nous avons seulement quatre cas où la pause est combinée à une autre marque : deux occurrences concernent le *euh* (dont un cas cumulé avec une répétition), une séquence contient une onomatopée et une autre un ponctuant. Ces quatre marques précèdent la pause :

accPH1 oui / non ce n'est pas / parce que / le wallon **mm // grossier** c'est pas beau mais le wallon m // un beau wallon c'est |- beau ça // ah oui [accPH1r]

blaAT1 ouais c'est c'est la folie quoi non mais c'est c'est c'est dingue d'être aussi **euh / inconscient** et (...) [blaAT1l]

blaAT1 (...) non vous le ferez **pour euh / pour mardi** puis mardi ils ne l'avaient pas fait (...) [blaAT1l]

blaAT1 (...) parce que / j'arrive à la voir super noire **tu vois // vraiment noire** enfin (...) [blaAT1]

Conclusions

Les trois degrés de pauses notés dans les transcriptions obtiennent chacun des fréquences très différentes : les pauses brèves sont de loin les plus nombreuses, suivies par les pauses longues, puis par les silences. Ces pauses sont des notations subjectives du transcripteur, ce qui rend la comparaison de nos données avec celles d'autres chercheurs la plupart du temps impossible.

Nous avons constaté que ces pauses sont placées le plus souvent entre des groupes syntaxiques cohérents. Nous avons émis l'hypothèse, renforcée par les tests de perception d'autres chercheurs, que le transcripteur place préférentiellement les pauses à des endroits cohérents pour la structuration syntaxique et sémantique de l'énoncé. Néanmoins, si ces pauses apparaissent à des endroits pertinents pour la structuration en groupes syntaxiques, elles ne correspondent pas aux emplacements des marques typographiques de l'écrit telles que le point, la virgule ou le point virgule, puisqu'on les trouve entre un sujet et un verbe ou encore entre un verbe et un complément d'objet direct.

Lorsque ces pauses se placent à des endroits non pertinents en termes d'analyse syntaxique, nous avons distingué deux cas : le premier constitué des occurrences où la pause est intraconstituant, le deuxième où la pause est placée au lieu d'une rupture syntaxique (le syntagme qui la précède est inachevé). Tous degrés de pauses confondus, le premier groupe contient deux fois plus d'occurrences que le second ; la tendance s'inverse en ce qui concerne le silence, mais le nombre d'occurrences dont nous disposons est très réduit.

L'examen attentif des marques groupées ne montre rien de concluant. Pour ce qui est de la rupture syntaxique, les marques combinées sont même moins fréquentes que les cas où la pause apparaît seule.

Les pauses intraconstituant ont des distributions variables : on a vu que, dans le sous-corpus ici analysé, on a 11 configurations différentes, la plus fréquente étant entre une conjonction et la proposition introduite par cette conjonction.

Au vu des résultats que nous obtenons ci-dessus, on peut penser que la marque du silence est un bon candidat au découpage du texte en unités plus petites. En ce qui concerne les pauses brèves et longues, on voit qu'elles le sont dans la majorité des cas. Néanmoins, les occurrences où elles ne le sont pas s'avèrent problématiques en regard des contraintes liées à notre système de levée d'ambiguïtés.

Références

- AUTESSERRE D., NISHINUMA Y., DELRAN-BADO S. (1992), « Respiration et phonation en situation de dialogue oral », in *Séminaire Dialogue*, 15-16 avril 1992, Service reprographie CNRS, Dourdan : 19-30.
- BLANCHE-BENVENISTE Cl., JEANJEAN C. (1987), *Le français parlé. Transcription et édition*, Didier Érudition, Paris.
- BOERSMA P., WEENINK D. (2006), *Praat : doing phonetics by computer* (Version 4.5.08) <http://www.praat.org/>.
- BOOMER D. S. (1965), « Hesitations and grammatical encoding », in *Language and Speech* 8 : 148-158.
- BOOMER D. S., DITTMAN A. T. (1962), « Hesitation pauses and juncture pauses in speech », in *Language and Speech* 5 : 215-220.
- CAMPIONE E., VERONIS J. (2004), « Pauses et hésitations en français spontané », in *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès) : 109-112.
- CANDEA M. (2000a), *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané*, Université de Paris-3 Sorbonne-nouvelle, Thèse non publiée.
- CANDEA M. (2000b), « Les euh et les allongements dits "d'hésitation" : deux phénomènes soumis à certaines contraintes en français oral non lu », in *Actes des 23^{es} Journées d'Étude sur la Parole (JEP'2000)* (19-23 juin, Aussois).
- DISTER A. (2007), *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel*, Université catholique de Louvain, Thèse non publiée.
- DISTER A., FRANCARD M., GERON G., GIROUL V., HAMBYE Ph., SIMON A. C. et WILMET Régine (2006). « Conventions de transcription régissant la banque de données Valibel ». <http://valibel.fltr.ucl.ac.be/>
- DUEZ D. (1991), *La Pause dans la parole de l'homme politique*, Éditions du CNRS, Paris.
- DUEZ D. (1995), « Perception of hesitations in spontaneous French speech », in *Proceedings of the 23rd International Congress of Phonetic Sciences* vol. 2, Stockholm : 498-501.
- DUEZ D. (1997), « La signification des pauses dans la production et la perception de la parole », in *Revue Parole* 3-4 : 275-299.
- DUEZ D. (2001), « Signification des hésitations dans la production et la perception de la parole spontanée », in *Revue Parole* 17-18-19 : 113-137.

- GOLDMAN-EISLER Fr. (1968), *Psycholinguistics : experiments in spontaneous speech*, Academic Press.
- GROSJEAN Fr., DESCHAMPS A. (1972), « Analyse des variables temporelles du français spontané », in *Phonetica* 26 : 129-156.
- GROSJEAN Fr., DESCHAMPS A. (1973), « Analyse des variables temporelles du français spontané. II. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique) », in *Phonetica* 28 : 191-226.
- GROSJEAN Fr., DESCHAMPS A. (1975), « Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation », in *Phonetica* 31 : 144-184.
- LANE H., GROSJEAN Fr. (1973), « Perception of reading rate by listeners and speakers », in *Journal of Experimental Psychology* 97 (2) : 141-147.
- MARTIN J., STRANGE W.. (1968), « The perception of hesitation in spontaneous speech », in *Perception and Psychophysics* 53 : 1-15.
- MOREL M.-A., DANON-BOILEAU L. (1998), *Grammaire de l'intonation. L'exemple du français*, Ophrys, Paris.
- OOSTDIJK N. (2003), « Normalization and disfluencies in spoken language data », S. Granger et St. Petch-Tyson (Éds.), *Extending the scope of corpus-based research. New applications, new challenges*, Rodopi, Amsterdam-New York:59-70.
- ROSSI M. (1981), *L'Intonation : de l'acoustique à la sémantique*, Klincksieck, Paris.
- ZELLNER Br. (1998), *Caractérisation et prédiction du débit de parole en français. Une étude de cas*, Université de Lausanne, Thèse non publiée.