L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ?

Christophe Benzitoun

Equipe DELIC – Université de Provence
29, Av. Robert Schuman, 13100 Aix-en-Provence
christophebenzitoun@yahoo.fr

Résumé – Abstract

Dans cet article, nous présentons une typologie des phénomènes qui posent problème pour l'annotation syntaxique de corpus oraux. Nous montrons également que ces phénomènes, même s'ils y sont d'une fréquence moindre, sont loin d'être absents à l'écrit (ils peuvent même être tout à fait significatifs dans certains corpus : e-mails, chats, SMS…), et que leur prise en compte peut améliorer l'annotation et fournir un cadre intégré pour l'oral et l'écrit.

In this paper, we present a typology of the phenomena that create problems for the syntactic tagging of spoken corpora. We also show that these phenomena, although less frequent, are far from being absent in written language (they can even be quite significant in some corpora: e-mails, chats, SMS…). Taken them into account can improve the annotation and provide a unified analysis framework for both spoken and written data.

Mots-clefs – Keywords

Annotation syntaxique, corpus oraux, NFCE, annotation de référence
Syntactic annotation, spoken corpora, reference annotation

1 Introduction

Cette dernière décennie a vu l'émergence de l'oral comme objet de recherche à part entière aussi bien dans les descriptions linguistiques qu'en TALN. Nous allons nous intéresser ici au français parlé, plus particulièrement au "spontané", dans une perspective d'annotation manuelle des relations syntaxiques à partir de transcriptions humaines assistées par ordinateur (logiciel *Transcriber*). Les conventions de transcription sont celles de (DELIC, à paraître). Nous comparerons l'oral avec le français écrit non standard.

Rares sont encore les corpus oraux syntaxiquement annotés : on peut notamment relever le *Christine Project* (Sampson, 2003) et le *Switchboard Corpus* (Taylor et al., 2003) pour l'anglais. Ils sont totalement inexistants pour le français oral et d'ailleurs rarissimes pour l'écrit

(cf. Habert et al., 1997 ; Abeillé et al., 2001). Or, la constitution de tels corpus constitue un enjeu majeur, à la fois pour la communauté des linguistes (comparaison oral/écrit de certaines structures, extraction automatique de concordances plus précises…) et pour la communauté des chercheurs en TALN (entraînement des parseurs sur l'oral, dialogue homme-machine…).

Dans le cadre de la campagne d'évaluation EASY (Evaluation des Analyseurs SYntaxique) du projet Technolangue EVALDA, nous avons été amenés à nous interroger sur les problèmes posés par l'annotation syntaxique de corpus oraux afin de savoir si cette tâche représentait un problème spécifique, compte tenu du fait que les analyseurs seront évalués sur des corpus écrits et oraux authentiques. Nous présentons dans cet article une typologie retraçant une partie des problèmes rencontrés à l'oral. Nous montrerons que l'étude de l'oral "spontané" permet d'aborder la question du traitement des "Nouvelles Formes de Communication Ecrite" (NFCE) (e-mails, forums, chats, SMS…), écrits plus ou moins normés dont le Web et la téléphonie mobile constituent une demande colossale notamment en terme de filtrage et d'analyse de contenus.

La réflexion autour des outils, du formalisme et du standard d'annotation liés aux diverses sorties des analyseurs syntaxiques a été largement abordée dans une perspective d'évaluation des analyseurs syntaxiques (cf. Caroll et al., 2003, la conférence associée à TALN 2003 "Evaluation des analyseurs syntaxiques"). En revanche, la question concernant le choix des annotations de référence est beaucoup moins débattue (cf. Aït-Mokhtar et al., 2003, pour l'écrit) et dépasse largement le seul problème d'évaluation. C'est cette question que nous allons aborder ici.

Nous avons trois pré-requis méthodologiques : une analyse superficielle liée à l'orientation contemporaine vers des analyseurs syntaxiques robustes ; des structures syntaxiques en dépendances ; et la conservation de l'intégralité de l'information transcrite (amorces, répétitions, reformulations...) liée à la possibilité d'une identification ultérieure plus fine des intentions des locuteurs (Antoine et al., 2003 : 29) et au fait que, selon nous, l'analyse syntaxique commence avec la transcription fidèle des paroles, toute suppression constituant déjà une analyse syntaxique en soi (cf. Blanche-Benveniste, Jeanjean, 1986).

2 Phrase vs Unité Maximale (UM)

Le premier problème que l'on rencontre lorsque l'on travaille sur l'oral, c'est l'inexistence de la ponctuation et, par voie de conséquence, des "phrases" graphiques. (Blanche-Benveniste, Jeanjean, 1986) ainsi que (Leech et al., 1997) ont montré que le recours à la ponctuation n'est pas satisfaisant pour transcrire l'oral car celle-ci ne correspond pas à des marques clairement identifiables et sa notation constitue déjà, en elle-même, une analyse syntaxique implicite. Par ailleurs, (Campione, Véronis, 2002) ont montré que l'on ne pouvait pas considérer les pauses comme des marques de ponctuation. La plupart des projets d'annotations syntaxiques du français portent uniquement sur l'écrit et posent comme pré-requis de l'analyse la segmentation en "phrases" (Gendner et al., 2003) qui n'est autre qu'une segmentation en "phrases" graphiques délimitées en amont par une majuscule et en aval par une ponctuation forte (Abeillé et al., 2003 : 170) avec quelques exceptions telles que les incises.

Cette position semble totalement inadaptée aux NFCE (ex. 1) et 2) ou à l'oral. Cette entité graphiquement délimitée est supposée représenter la plus grande unité syntaxique, c'est-à-dire

qu'en partant de n'importe lequel des constituants, on est capable de remonter soit à la tête de la phrase (généralement le prédicat en grammaire de dépendances) soit au nœud phrase (en grammaire de constituants). Pour la rendre opératoire à l'oral, il est donc nécessaire de vérifier ses propriétés syntaxiques afin de voir s'il est possible de la définir sans l'aide de la ponctuation. Or, on constate qu'une phrase graphique peut inclure un grand nombre d'unités non dépendantes (ex. 1) ou segmenter des unités dépendantes (ex. 2), si bien que l'on ne peut en faire une unité syntaxique à part entière.

1) *Les armes c'est comme les voitures,ca coute des vies humaines chaque annee, mais c'est bien de les avoir quand meme, voudriez vous vous passer des voitures,non et bien les armes c'est pareil,si les citoyens etaient armes les criminels ne seraient plus aussi nombreux,beaucoup d'entre eux sont des laches,une balle ca fait reflechir,les bons sentiments ca les fait rigoler,une balle c'est le debut de la sagesse,je sais c'est triste je compatis avec vous,mais la realite c'est ca,pouvoir ce defendre quelque soit le prix a payer,sinon on va avoir droit a la jungle et ce systeme aussi il tue des innocents chaque annee.* [écrit][1]

2) *C'est une grande épreuve pour toute la famille et je l'accompagne encore aujourd'hui quotidiennement.* ***Dans ce combat permanent pour la recherche de l'équilibre glycémique****.* [écrit]

Notons que l'accumulation d'unités sans lien syntaxique dans une même phrase graphique et la segmentation d'unités construites ne sont pas des procédés propres aux NFCE mais que l'on en trouve dans d'autres écrits plus normatifs : en littérature, par exemple (cf. Vialatte cité par Blanche-Benveniste, 2002).

De plus, la ponctuation peut marquer autre chose que des fins de phrases :

3) ***Trois...Anglais!*** *se sont totalement dévêtus mardi devant Buckingham Palace, répondant ainsi à l'appel lancé par autocollants dans le métro londonien à venir manifester ce jour à 14:00 pour le droit de se promener tout nu.* [écrit]

Au vu de ces exemples, il semble déjà difficile de définir une unité "phrase" à l'écrit. Il est donc inenvisageable de la transposer à l'oral. En conséquence, il nous faut une unité de substitution et comme l'a montré (Kleiber, 2003) à propos des unités proposées par Berrendonner, la tâche est loin d'être aisée si l'on ne veut pas retomber dans les contradictions mêmes que l'on dénonce.

Nous avons donc choisi de suivre (Blanche-Benveniste, 2002) qui part des constructions verbales et non des "phrases". Nous définirons nos Unités Maximales (UM) comme étant des constructions verbales, nominales, adjectivales ou adverbiales regroupant un élément tête ainsi que tous les éléments qui sont dans sa dépendance ou qui lui sont "associés" (au sens de Blanche-Benveniste, 1990). Ainsi, notre UM ne précèdera pas l'analyse des dépendances mais elle résultera de la mise en relation syntaxique de toutes les unités entre elles, ce qui, dans une perspective d'annotation syntaxique, se justifie pleinement. Nous analyserons donc l'exemple 2) comme étant formé d'une seule UM, malgré la présence du point, et l'exemple 4) de deux UM.

---

[1] Références des corpus : [écrit] : forums sur Internet ; [oral] : corpus oraux "spontanés".

4) *en principe il me fait toujours les mêmes + sauf le le /où, ou/ le dimanche où il fait euh les gâteaux traditionnels mille-feuilles euh baba au rhum Paris-Brest et tout ça + |$_{UM1}$ sinon dans la semaine j'ai toujours les mêmes + |$_{UM2}$* [oral]

La recherche d'une UM à l'oral, liée au fait qu'il n'y a pas de ponctuation, permet donc de remettre en perspective l'UM postulée à l'écrit, à savoir la "phrase". En effet, si l'on conserve la "phrase graphique", on ne voit pas trop comment faire de l'exemple 1) un seul îlot de dépendances et faire de la deuxième phrase de l'exemple 2) une unité syntaxique autonome.

Nous regroupons donc dans une même unité ce que (Biber et al., 1999) appellent "clausal units" (pour les UM à tête verbale) et "non-clausal units" (pour les UM à tête autre que verbale) et nous rejoignons (Wagner et Pinchon, 1962 : 501-2) qui utilisent "phrase" dans l'acception restreinte de construction verbale, excepté que nous ne séparons pas les UM à tête verbale et non verbale.

3 Marques liées à la mise en discours

Une objection que l'on fait souvent à l'adoption d'un cadre commun d'unités pour l'écrit et pour l'oral est que ce dernier présente des formes quasiment imprévisibles du fait de sa production "on line" où se mêlent faits de langue et faits de parole. Ces faits sont liés à ce que qualifient (Blanche-Benveniste et al., 1990) de "modes de production de l'oral" et (Morel et Danon-Boileau, 1998) de "marques du travail de formulation". Il s'agit notamment des répétitions, reformulations, amorces, etc. qui sont des traces de la production du discours au même titre qu'un brouillon, à l'écrit (Blanche-Benveniste et al., 1990 : 17 ; Blanche-Benveniste, Jeanjean, 1986 : 155-161). En fait, nous pouvons montrer que ce phénomène n'est pas spécifique à l'oral car l'on retrouve ces marques dans des textes écrits tirés des NFCE.

Nous aurions pu éliminer ces traces de la mise en discours, qui, le plus souvent ne modifient pas la structure syntaxique de l'énoncé mais en compliquent seulement la réalisation linéaire. Nous pensons qu'il est intéressant de les conserver afin de ne pas présumer de l'analyse syntaxique qui pourrait en être faite. De plus, contrairement à ce que nous pourrions penser, les répétitions (ex. 5), 6) et 7) ainsi que les reformulations (ex. 8) et 9) sont des procédés certes rares à l'écrit mais pas inexistants. Encore faut-il préciser sur quel type d'écrit ou quel type d'oral on travaille car il y a peu de chances que l'on en trouve dans des situations d'oral très contraintes (discours juridiques, politiques etc.) ou d'écrits oralisés.

3.1 Les répétitions

Outre la distinction, qui peut poser problème, entre répétitions de "langue" :

5) *Avec des conseils, des explications, des solutions, des témoignages, des livres, un Forum, des blagues le tout de **très très** bonne qualité.* [écrit]

et répétitions liées à la mise en discours :

6) *la les traitements les produits sont beaucoup plus efficaces et finalement **on + on** utilise de moins grandes quantités de produit* [oral]

on remarque que les répétitions sont, par définition contiguës et, comme le fait remarquer

(Antoine et al., 2003) s'appuyant sur (Blanche-Benveniste, 1997 : 47), elles sont généralement limitées aux chunks c'est-à-dire à des constituants minimaux, comme la plupart des phénomènes répertoriés dans cette partie.

Nous trouvons difficilement des cas de répétitions liées à la mise en discours dans les écrits standards, hormis dans les dialogues ou les citations. A première vue, il s'agirait donc d'un phénomène propre à l'oral mais que l'on peut trouver à l'écrit dans les cas où l'écrit représente l'oral (dialogue, discours rapporté, citation).

En fait, on trouve des exemples de répétitions dans les NFCE :

7) *Oui ceux de mon Université sont soient fils de riches investisseurs ou bien fils **de de** politiciens ayant réussi.* [écrit]

De plus, des concepteurs de logiciels de traitement de textes ont éprouvé le besoin d'intégrer une fonction prenant en charge la correction de ce type de phénomène dans leurs correcteurs orthographiques (c'est le cas, par exemple, dans le logiciel *Microsoft Word*), ce qui laisse penser que l'on produit des répétitions en tapant du texte. Les répétitions ne se limitent donc pas aux seules productions orales, même si, dans certains registres, elles sont beaucoup plus représentées qu'à l'écrit. Nous allons voir si les reformulations sont aussi présentes à l'écrit.

3.2 Les reformulations

8) *les enseignants à l'époque n'avaient peut-être pas + euh comme aujourd'hui + euh + l'ha- l'ha- l'ha- **l'habitude** enfin la la **la vocation** hein + euh disons **l'envie** + euh de travailler + euh sur ces thèmes-là* [oral]

9) *Au plaisir de vous **rire**, euh, pardon, **lire** ...* [écrit]

Les reformulations ressemblent beaucoup aux énumérations, si bien qu'il est souvent possible de les confondre. Dans l'énoncé 6), par exemple, nous ne disposons pas d'indices permettant d'analyser *les traitements les produits* comme une reformulation ou une énumération. En revanche, nous disposons parfois d'indices formels permettant de les distinguer. Par exemple, *l'habitude enfin la vocation disons l'envie* (ex. 8) peut être considérée comme étant une reformulation car *enfin* et *disons* orientent clairement vers une interprétation dans ces termes. Ce problème d'ambiguïté entre reformulation et énumération a été pointé notamment par (Levelt, 1983).

La prise en compte de cette distinction est envisageable car nous disposons parfois d'indices formels tels que les marqueurs d'énumération et de reformulation ou les accords. Mais nous avons fait le choix de ne pas la conserver car il ne s'agit pas là d'un problème syntaxique à proprement parler mais plutôt d'un problème d'interprétation, la reformulation mettant en jeu le même cadre syntaxique que l'énumération. Il s'agit en fait de trois utilisations différentes d'un même procédé syntaxique concernant l'axe paradigmatique. Dans les trois cas (répétition, énumération et reformulation), on liste des éléments dans une même place syntaxique. Nous suivons donc la perspective de (Blanche-Benveniste et al., 1990 : 20) :

[…] nous ne distinguerons pas entre les phénomènes qui paraissent créés par la volonté des locuteurs et ceux qui paraissent leur échapper ; nous traiterons de la même façon des phénomènes apparemment involontaires comme bredouillages, hésitations, maladresses,

reprises, et d'autres qui semblent intentionnels comme : répétitions intensives, variations stylistiques et autres. On verra que ce point de vue […] a l'avantage de suivre une ligne d'analyse grammaticale unifiée.

Nous avons donc fait le choix de regrouper tous les procédés permettant de lister des éléments dans une même place syntaxique sous l'étiquette "liste".

3.3 Inserts

En ce qui concerne les mots du type *hein, bon,* que l'on désigne par des termes aussi divers que "particules discursives", "marqueurs discursifs"… nous allons adopter la terminologie de (Biber et al., 1999 : 93-4 et 1082-3) qui parlent d' "inserts". Les "inserts" sont définis comme étant des mots pouvant constituer des UM à eux seuls et ne pouvant pas entrer dans une relation syntaxique avec un autre élément. Ils sont généralement attachés prosodiquement à un ensemble plus vaste.

10) *L3 oui mais c'est pas définitif* ***hein*** *madame c'est pas encore définitif* [oral]

11) *Elle se veut rhétorique,* ***hein****, votre question?* [écrit]

Les inserts peuvent être ambigus car ils ont des homonymes. Par exemple, *quoi* peut être insert ou pronom, *bon* peut aussi être adjectif. Dans les corpus, on trouve de nombreux cas d'ambiguïté qui, dans une perspective de traitement automatique[2], restent très compliqués à désambiguïser. Dans l'exemple suivant, on ne sait pas s'il s'agit du pronom ou de l'insert.

12) *L1 je ne sais plus* ***quoi*** *+ bon alors comment j'appliquais c'est ça* [oral]

Les inserts ne sont pas propres à l'oral. On en trouve de nombreuses attestations par écrit notamment dans les forums ou les chats (cf. ex. 11).

3.4 Inachèvements et amorces

- les inachèvements

13) *L1 hé quelques années hein + ça allait que j'étais fille bon ben le soir je rentrais j'avais rien à faire eh* ***mais les personnes qui étaient mariées*** [oral]

14) *Moi j'accepte vos excuses,* ***mais les malades*** ... [écrit]

- les amorces

15) *et + mais en* ***m-*** *donc dans un souci d'égalité de* ***d'ac-*** *que tout le monde soit accepté et en même temps enfin* [oral]

16) *N'empêche que moi je la fais remplir par ma môman tellement ça me casse les* ***c...*** [écrit]

---

[2] Ces cas d'ambiguïté sont aussi difficiles à résoudre dans une perspective linguistique.

Dans tous ces cas, la structure syntaxique en dépendances est achevée et doit être marquée comme telle. C'est la réalisation de la place syntaxique qui est laissée en suspens (avec des effets de recherche lexicale ou d'anaphore discursive). Dans les écrits standards, on trouve d'ailleurs des ellipses conventionnelles qui sont plus codifiées mais qui relèvent du même principe (ex. 17). Le "spontané" ne fait qu'étendre le domaine de ces structures à lexique incomplètement réalisé.

17) *Ce n'est pas de forte volonté politique qu'il s'agit seulement, mais bien de définir quelles politiques défendre et dans l'intérêt* ***de qui****....* [écrit]

4 Marqueurs de relation

Dans la partie précédente, nous avons essentiellement évoqué des phénomènes non envisagés dans d'autres systèmes d'annotation syntaxique. Dans cette partie, nous proposons un traitement original des mots que l'on retrouve traditionnellement sous les termes prépositions, conjonctions et adverbes.

Contrairement à l'habitude consistant à faire de certains indices morphologiques des marques de dépendance telle que les "conjonctions de subordination" qui marqueraient une relation de "subordination", nous pensons que ces items marquent en fait une relation dont la nature nous est révélée par le contexte. C'est ainsi que *parce que*, contrairement à ce qui est admis, n'est pas une marque de "subordination", au sens syntaxique du terme, mais peut tout aussi bien marquer une relation de dépendance entre constituants (ex. 18) qu'un enchaînement sur un contexte extralinguistique (ex. 19) (cf. Debaisieux, 1994). De même, *et* peut débuter un tour de parole (ex. 19).

18) *aujourd'hui on peut se marier* ***parce que ben on a envie de payer moins d'impôts*** *on peut se marier* ***parce qu'on a besoin d'un prêt*** *+ et qu'il faut être ben euh monsieur madame* [oral]

19) *L2 c'est > foutu bon les Bretons vous avez vu les ils a- sont en plus ils sont maladroits ces pauvres Bretons*
L1 oui ça c'est sûr
L2 bon alors voilà < mais le le
L1 bon >
L2 ***parce qu'on ne fait rien pour cela quoi***
L1 ***et*** *concrètement pour vous par exemple vous pouvez nous parler un peu de la manière dont vous essayez de mener le combat pour préserver cet héritage culturel* [oral]

Ainsi, nous ferons de ces mots des "marqueurs de relation" sans spécifier la relation qu'ils marquent. C'est le cas des mots en gras dans l'exemple suivant.

20) *on est un petit peu décalé* ***vis à vis de*** *ça* ***mais*** *ben oui je suis jeune j'ai vingt-cinq ans +* ***et puis alors*** *c'est* ***comme*** *ça ça arrive + je trouve* ***que*** *peut-être* ***que*** *ça me manquait* ***quand*** *j'étais petit* [oral]

Ces mots peuvent marquer qu'il y a une relation entre deux unités quelle qu'en soit la nature hiérarchique. Le fait qu'un même mot puisse marquer des relations de nature différentes n'est pas un phénomène propre à l'oral. On en trouve dans les NFCE.

21) *Ah bon,* ***parce que*** *le codage numérique protège contre le brouillage hertzien ?* [écrit]

Nous aurions pu penser que les prépositions constitueraient un cas particulier mais il n'en est rien. On a tous entendu plusieurs dizaines de fois, après avoir pris une baguette de pain, la question *avec ceci* dans laquelle la préposition n'est pas rattachée ou on a déjà mis ou vu à la fin d'une lettre *Dans l'attente d'une réponse de votre part*.

Par ailleurs, la frontière entre insert et marqueur de relation peut parfois être mince. Certains mots peuvent avoir un emploi d'insert ou de marqueur de relation. Pour le traitement des cas ambigus, il faut observer que les inserts ont éventuellement une valeur démarcative et ne sont jamais relationnels alors que les marqueurs de relation cumulent valeurs relationnelle et démarcative.

Ces marques de relation à différents niveaux ne sont pas marginales ni surprenantes dans les langues. On en trouve dans un peu tous les types d'écrits et d'oraux. Elles sont seulement marginalisées dans les formes les plus conventionnelles enseignées à l'école.

5 Problèmes d'analyse qui restent à résoudre

Tous les cas précédents trouvent, on l'a vu, une analyse dans un cadre syntaxique descriptif couvrant l'ensemble des usages de la langue. Il reste des cas qui résistent encore à l'analyse. Nous en évoquons deux dans cette partie.

5.1 Les places remplies par une autre catégorie que celle attendue

Il s'agit de complémentations "non canoniques" réalisées par l'intermédiaire de remarques métalinguistiques qui annoncent une réalisation lexicale pouvant faire problème (ex. 23) ou l'absence d'un élément attendu (ex. 22) dans lequel il manque la préposition *de* en principe sélectionnée par le verbe *m'occuper*). Il est clair que la réalisation en catégories de la place objet du verbe *dire*, dans l'exemples 23), n'est pas canonique et qu'il est difficile de prévoir toutes les formes d'introducteurs de ce type.

22) *je m'occupais euh* ***tout ce qui était transmission*** [oral]

23) *t'as raison, Elie, d'ailleurs, si dieu n'existait pas, il faudrait l'inventer - comme disait* ***je sais plus qui*** *(voltaire, maybe ?)* [écrit]

5.2 Les incises

24) *euh on considérait que former + les hommes* ***et c'est toujours euh + en en en vigueur ça hein*** *+ former les les en- les les enfants d'aujourd'hui + c'est aussi former les hommes de demain hein* [oral]

25) *Pour ce qui concerne le SMIC et le chômage, il est vrai que plus le salaire minimum est élevé* ***-ce qui bien entendu ne répond pas à la question de savoir quel devrait être son niveau "optimum" afin de maximiser l'emplois-*** *moins on n'embauche de mains d'oeuvres "peu" qualifiées.* [écrit]

On observe que toutes les contraintes sur la combinatoire des catégories tombent si l'on inclut les incises dans les analyses. Ainsi, on peut séparer un clitique d'un verbe par une incise :

26) *Vous signale que l'an dernier **il (JMV)** l'etait encore (cette année je ne sais pas)* [écrit]

6 Conclusion et perspectives

A l'époque où la demande d'applications de TALN pour le Web - source hétérogène dans laquelle les écrits sont plus ou moins normés – la téléphonie mobile ou l'oral "spontané" est de plus en plus grande, nous devons envisager de nouveaux problèmes que l'on n'avait peut-être pas remarqués faute de représentativité statistique dans les corpus étudiés jusqu'à présent. Il est vrai que, pour des écrits standards, la faible représentativité (ou l'absence) de ces phénomènes laisse la possibilité de les ignorer sans affecter considérablement le rappel ou la précision. En revanche, pour l'oral "spontané" ou les NFCE, il n'est pas envisageable de ne pas les prendre en considération.

Nous avons montré que l'on ne pouvait pas opposer de manière simpliste écrit et oral, surtout quand on traite des NFCE, qui peuvent être considérées comme de l'écrit "spontané". Comme l'ont dit (Blanche-Benveniste, Jeanjean, 1986 : 154) à propos de la position qui consiste à opposer oral et écrit :

Cette position est pourtant totalement injustifiée ; si on le présente sous cette forme c'est une supercherie ; car cela revient à comparer de l'oral spontané à de l'écrit élaboré.

En comparant écrit et oral "spontanés", on se rend compte qu'il y a finalement de nombreuses similitudes : non recours à la ponctuation pour délimiter des UM, absence généralisée de la négation *ne* (ex. 27), répétitions, reformulations, particules discursives, etc.

27) *Il a pas fait beau aujourd'hui, beaucoup de nuages et de vent.* [écrit]

L'opposition n'est donc pas entre écrit et oral mais entre planifié et non planifié. Les marques attribuées à tort à l'oral sont en réalité des marques liées à la mise en discours avec un temps d'élaboration court. En effet, on peut supposer que plus l'élaboration est faible, plus l'on trouvera ces marques. Le fait que l'on retrouve les mêmes phénomènes dans les NFCE et à l'oral "spontané" montre bien que cette hypothèse est justifiée.

Nous sommes bien conscient que les propositions faites ici ne sont pas utilisables dans une perspective d'évaluation d'analyseurs syntaxiques pour la raison évidente que les systèmes évalués n'opèrent pas ces distinctions. Mais du moins avons-nous montré que, pour ce qui est de l'annotation syntaxique, nous pouvons traiter l'oral et l'écrit de manière unifiée moyennant une plus grande prise en compte des usages non standards. Cette orientation peut être aussi intéressante pour les analyseurs syntaxiques à large couverture.

Il nous reste encore à tester plus systématiquement notre approche sur des corpus plus grands et plus diversifiés afin de créer ou supprimer des étiquettes et aboutir ainsi à un système d'annotation de référence suffisamment général pour permettre d'étiqueter n'importe quel type de corpus. Actuellement, nous effectuons des tests destinés à voir si le découpage en UM est faisable à grande échelle et reproductible. Cette démarche nous semble nécessaire pour

répondre à la demande technologique d'un média (le Web) dans lequel tous les types de corpus sont potentiellement présents et dont on ne connaît pas a priori la nature.

Références

Abeillé A., Clément L., Toussenel F. (2003), Building a treebank for french, *Treebanks : Building and Using Parsed Corpora*, Abeillé (éd.), Kluwer Academic Publishers, pp. 165-187.
Abeillé A., Clément L., Kinyon A., Toussenel F. (2001), Un corpus de français arboré : quelques interrogations, Actes de la *Conférence TALN'2001*, Tome 1, pp. 33-42.
Aït-Mokhtar S., Hagège C., Sándor A. (2003), Problèmes d'intersubjectivité dans l'évaluation des analyseurs syntaxiques, Actes de la *Conférence TALN*, Tome2, pp. 57-66.
Antoine J.-Y., Goulian J., Villaneau, J. (2003), Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée, Actes de la *Conférence TALN'2003*, Tome 1, pp. 25-34.
Berrendonner A. (2002), Les deux syntaxes, *Verbum*, XXIV, 1-2, pp. 23-35.
Blanche-Benveniste Cl. (2002), Phrase et construction verbale, *Verbum*, XXIV, 1-2, 7-22.
Blanche-Benveniste Cl., Bilger M., Rouget C., van den Eynde K. (1990), *Le français parlé : études grammaticales*, Paris, éditions du CNRS.
Blanche-Benveniste Cl., Jeanjean C. (1986), *Le français parlé. Edition et transcription*, Paris, Didier-Erudition.
Biber D., Johansson S., Conrad S., Finegan E. (1999), *Longman Grammar of Spoken and Written English.* Longman.
Campione E., Véronis J. (2002), Etude des relations entre pauses et ponctuations pour la synthèse de la parole à partir de texte, Actes de la *Conférence TALN'2002*, pp. 175-184.
Caroll J., Minnen G., Briscoe T. (2003), Parser evaluation, *Treebanks : Building and Using Parsed Corpora*, Abeillé (éd.), Kluwer Academic Publishers, pp. 299-316.
Debaisieux J.-M. (1994), *Le fonctionnement de* parce que *en français parlé contemporain : Description linguistique et implications didactiques*, Thèse, Université de Nancy II.
DELIC (à paraître), Conventions de transcription, *Recherches sur le français parlé*, N°19.
Gendner V., Vilnat A. (2003), *Les annotations syntaxiques de référence PEAS*, Version 1.4.
Habert B., Nazarenko A., Salem A. (1997), *Les linguistiques de corpus*, Armand Colin.
Kleiber G. (2003), Faut-il dire *adieu* à la phrase ?, *l'information grammaticale*, n°98, juin.
Leech G., McEnery T., Wynne A. (1997), Further levels of annotation, *Corpus Annotation*, Garside, Leech & McEnery (éds.).
Levelt W.J.M. (1983), Monitoring and self-repair in Speech, *Cognition*, 14, pp. 41-104.
Morel M.-A., Danon-Boileau L. (1998), *Grammaire de l'intonation : l'exemple du français*, Paris, Ophrys.
Sampson G. (2003), Thoughts on two decades of drawing trees, *Treebanks : Building and Using Parsed Corpora*, Abeillé (éd.), Kluwer Academic Publishers, pp. 23-41.
Taylor A., Marcus M., Santorini B. (2003), The Penn treebank : an overview, *Treebanks : Building and Using Parsed Corpora*, Abeillé (éd.), Kluwer Academic Publishers, pp. 5-22.
Wagner R.L., Pinchon J. (1962), *Grammaire du français classique et moderne*, Hachette Université.