Un corpus français arboré : quelques interrogations

Anne Abeillé (1), Lionel Clément (1), Alexandra Kinyon (1)(2), François Toussenel (1)
(1) UFRL, Université Paris 7
(2) IRCS - Upenn
{abeille, clement, kinyon, ftoussen}@linguist.jussieu.fr

Résumé - Abstract

Dans cet article nous présentons les premiers résultats de l'exploitation d'un Corpus français arboré (Abeillé *et al.*, 2001). Le corpus comprend 1 million de mots entièrement annotés et validé pour les parties du discours, la morphologie, les mots composés et les lemmes, et partiellement annotés pour les constituants syntaxiques. Il comprend des extraits de journaux parus entre 1989 et 1993 et écrits par divers auteurs, et couvre différents thèmes (économie, littérature, politique, etc.). Après avoir expliqué comment ce corpus a été construit, et comment l'exploiter à l'aide d'un outil de recherche spécifique, nous exposerons quelques résultats linguistiques concernant les fréquences et les préférences lexicales et syntaxiques. Nous expliquerons pourquoi nous pensons que certains de ces résultats sont pertinents en linguistique théorique et en psycholinguistique.

This paper presents the first linguistic results exploiting a new treebank for French (Abeillé *et al.*, 2001). The corpus comprises 1 million words fully annotated and disambiguated for parts of speech, inflectional morphology, compounds and lemmas, and partially annotated with syntactic constituents. It is made of extracts from newspapers ranging from 1989 to 1993 and written by different authors, and covers a variety of subjects (economy, literature, politics, etc.). After explaining how this corpus was built, and how it can be used with a specific search tool, we present some linguistic results obtained when searching the corpus for lexical or syntactic frequencies and preferences, and explain why we think some of these results are relevant both for theoretical linguistics and psycholinguistics.

Mots-clefs – Keywords : Corpus arboré, corpus journalistique, français, syntaxe. Treebank, French, Newspaper corpora, syntax.

1 La construction du corpus annoté

1.1 Motivation

Comme le notent (Habert *et al.*, 1997), aucun corpus arboré n'est actuellement disponible pour le français. Dans le cadre des récents projets français (GRACE, CLIF) et européens (Multext, Parole), des corpus littéraires (Frantext) et journalistiques (Le Monde) commencent à

être disponibles. Mais ceux qui sont annotés le sont seulement pour les catégories morpho-syntaxiques, et sans validation humaine systématique.

Le corpus journalistique d'environ 1 million de mots annoté systématiquement pour la morpho-syntaxe et la syntaxe (avec contrôle de qualité) doit pouvoir servir aussi bien aux linguistes informaticiens qu'aux linguistes traditionnels comme aux psycholinguistes.

Le but de l'annotation morpho-syntaxique est de fournir une ressource pour entraîner un segmenteur ou un étiqueteur automatique, extraire de façon automatique des dictionnaires de formes, d'évaluer des étiqueteurs existants.

Le but de l'annotation syntaxique est de pouvoir entraîner des analyseurs automatiques, extraire de façon automatique des grammaires et des dictionnaires de contexte, d'évaluer des analyseurs existants, mais aussi d'affiner nos connaissances sur des constructions souvent peu décrites, et également de mesurer l'écart entre le possible et l'attesté (par exemple, quels adverbes trouve-t-on en tête de phrase ou en tête de groupe nominal ? Quels verbes trouve-t-on au passif ? Les sujets phrastiques sont-ils toujours extraposés ?).

Nous souhaitons présenter un type d'annotations paramétrable, permettant à l'utilisateur de prendre en compte ou d'ignorer nos mots composés, de visualiser tout ou partie des informations annotées.

1.2 Choix du corpus

Les textes que nous avons choisi d'étiqueter sont extraits d'un ensemble d'articles du journal Le Monde de 1989 à 1993 compilés par Linguistic Data Consortium (LDC). Nous avons sélectionné 1 million de mots (875 000 mots sans ponctuation, 920 000 mots avec les mots composés regroupés, 17 000 lemmes différents, 33 000 phrases) en puisant de façon aléatoire dans l'ensemble de la base pour faire varier les dates de parution des articles, leurs auteurs et leurs thèmes. Sans être représentatif des différentes variétés de français, le corpus couvre différents domaines puisque les pages culturelles, économiques ou sportives n'ont pas été exclues.

1.3 Les choix d'annotation

Nous souhaitons que le corpus annoté puisse servir aussi bien aux informaticiens (à des fins d'évaluation ou d'entraînement de parseur, d'extraction de lexique ou de grammaire) qu'aux linguistes ou aux psycholinguistes (pour connaître les distributions fines de certains mots ou de certaines catégories, ou connaître la fréquence relative de certaines constructions, etc.). Contrairement à d'autres projets (Prague Dependency Treebank, Penn Treebank), nous ne souhaitons pas appliquer fidèlement telle ou telle théorie syntaxique, forcément éphémère, mais contribuer à l'émergence d'un standard de découpage en constituants, sans doute un peu grossier mais suffisamment consensuel, et traduisible dans différents cadres théoriques.

Nous distinguons les choix d'étiquetage (des mots ou des syntagmes) des choix de découpage (des mots ou des syntagmes).

1.3.1 Les choix d'étiquetage

Nous avons 14 catégories lexicales présentes dans le corpus, pour les mots simples ou composés :

A (adjectif), Adv (adverbe), D (déterminant), CC (conj de coordination), CL (pronom personnel clitique= forme faible), CS (conj de subordination), NC (nom commun), NP (nom propre), P

(préposition), PRO (pronom non clitique), V (verbe), I (interjection), ET (mot étranger dont on ne peut deviner la catégorie en contexte), PONCT (ponctuation).

Nous avons retenu 12 types de syntagmes : syntagmes adjectival (AP), adverbial (AdP), coordonné (COORD), nominal (NP), prépositionnel (PP) ; propositions participiale (VPpart), infinitive (VPinf), relative (Srel), subordonnée (Ssub), conjuguée interne (Sint) ; noyau verbal (verbes avec les clitiques, auxiliaires, faire : VN) ; et phrase indépendante (SENT).

Les pronoms clitiques, les déterminants, les ponctuations, les mots étrangers, ne projettent pas de syntagmes.

Nous n'avons pas de syntagme verbal (VP), car en français, la séquence postverbale inclut aussi bien des compléments que des circonstants (1) ou des sujets inversés (2) :

(1) Les actionnaires décideront certainement une augmentation de capital.

(2) Les actions qu'a mises IBM sur le marché.

Donc, soit VP englobe tout et il est inutile, soit il n'inclut que les compléments et il est discontinu. Pour distinguer compléments et circonstants, on notera des fonctions, dans une seconde phase d'annotation (non pertinente ici).

Nous n'avons pas de syntagme déterminatif (DP) non plus. Les prédéterminants (adverbes : *environ*, *presque* etc. ou prépositions composées : *près de*...) ne sont pas enchâssés, les déterminants coordonnés donnent lieu à un sous-constituant COORD :

<NP> presque:Adv tous:A les:D enfants:NC </NP>
<NP> près-de:P trois-cents:D personnes </NP>
<NP> deux:D <COORD> ou:CC trois:D </COORD> enfants:NC </NP>

Le jeu d'étiquettes[1] est assez fin pour marquer la catégorie, la sous-catégorie et la morphologie flexionnelle. Il est principalement inspiré des projets Multext et GRACE (Lecomte & Paroubek, 1996) et suit les recommandations du projet EAGLES (von Rekowski, 1996).

1.3.2 Les choix de découpage

Pour le découpage en mots, nous utilisons un dictionnaire de mots composés, qui est maximal pour les mots composés grammaticaux (les moins controversés) et minimal pour les mots composés lexicaux (qui sont les plus flous). Nous suivons en général les critères de G. Gross (voir (Gross, 1996)).

Le marquage en constituants majeurs est la première étape d'un marquage syntaxique qui comprendra ensuite le marquage des principales fonctions grammaticales.

Certains choix que nous avons dû prendre concernent un certain nombre de cas réputés difficiles, celui des constituants discontinus, des éléments vides, des syntagmes exocentriques (sans tête) ou unaires, des ambiguités résiduelles.

Pas de constituants discontinus On fait l'hypothèse qu'on n'a pas de constituants discontinus ni de syntagmes croisés. Toutefois, l'annotation des fonctions grammaticales permettra de marquer liens de dépendance entre des éléments à distance.

On ne note pas non plus de catégorie vide pour les constituants absents ou déplacés. C'est à un autre niveau qu'on pourra noter une fonction à distance (dans : *que veux-tu voir ? que* est objet de *voir* et non de *veux*) ou une double fonction (dans : *je le laisse entrer*, *le* est à la fois objet de *laisser* et sujet *d'entrer*).

[1] La liste contient 202 étiquettes différentes.

Les syntagmes exocentriques On ne note pas de catégorie vide pour les syntagmes elliptiques ou "sans têtes".

Pour l'étiquetage lexical, on a évité de recatégoriser les mots qui n'avaient pas leur fonction canonique car il en serait résulté un grand arbitraire au niveau lexical. Tous les adjectifs peuvent être employés comme têtes de groupe nominal et il serait artificiel de noter comme ambigus (N ou Adj) seulement ceux rencontrés dans cet emploi dans le corpus. Les noms peuvent aussi être employés comme épithètes ou attributs (*un ingénieur maison*, *il est très famille*).

On a donc des NP sans Nom, et on n'ajoute pas un N vide qui serait artificiel. On a des phrases sans verbe (relatives ou subordonnées) pour les tours elliptiques (*que toi*, ou *dont trois idiots*).

On ne créé pas de syntagmes adjectivaux à tête nominale (sans adjectif), car on autorise les groupes nominaux à avoir la fonction épithète ou attribut. On ne crée pas d'adverbiales à tête adjectivale non plus, car on autorise les adjectivales à être mobiles dans la phrase (c'est à dire qu'on découpe: *laver plus blanc*, comme *partir content*, ou *voter utile*).

Pas d'ambiguïté résiduelle On a fait l'expérience qu'une lecture suffisamment attentive permettait généralement de désambiguer les phrases du corpus. Dans certains cas, il ne faut pas hésiter à faire appel à des connaissances encyclopédiques.

Dans certains cas, l'interprétation est exactement la même et on a le choix entre deux découpages. Si les tests syntaxiques marchent aussi bien pour les deux (pronominalisation, clivée et *c'est ... que*, etc.), on opte pour celui qui comporte le moins d'enchâssements (principe du rattachement minimum).

C'est le cas par exemple pour la plupart des constructions à verbe support, dans *Jean <VN> a commis </VN> <NP> une agression </NP> <PP> contre <NP> Marie </NP> </PP>* on peut inclure ou non le PP (*contre Marie*) dans le NP (*une agression*), sans différence d'interprétation, on choisit donc de ne pas l'inclure.

Les syntagmes unaires A des fins de simplification et de lisibilité, on essaie de limiter les syntagmes unaires.

On considère que certaines catégories peuvent projeter des syntagmes à elles toutes seules. On dit qu'on a affaire à des syntagmes unaires. C'est le cas des noms propres, des pronoms (non clitiques), des verbes (sauf participe passé), des adjectifs, mais ce n'est pas systématique.

A chaque fois on doit se demander si on peut remplacer le mot seul par une séquence de mots tout en gardant le même découpage syntagmatique. Les noms employés en apposition correspondent à des NP, mais pas les noms employés comme épithète.

1.4 Méthodologie

Nous procédons en deux étapes :

- une étape d'annotation morpho-syntaxique, qui comprend segmentation en mots et en phrase, désambiguation catégorielle et morphologique (tagging), regroupement des mots composés et lemmatisation,
- une étape d'annotation syntaxique (qui comprend le marquage en constituants et l'assignation de fonctions grammaticales à ces constituants).

A chaque étape, nous distinguons un marquage automatique et une validation/correction humaine systématique, afin de minimiser le risque d'erreurs.

La première étape est aujourd'hui achevée. Seule la première phase de la seconde étape (marquage en constituants) est actuellement opérationnelle.

1.4.1 Validation du corpus étiqueté

Un guide d'annotation très précis a été écrit et complété durant la correction manuelle du corpus (Abeillé & Clément, 1997). L'intégralité du corpus à été relu et validé deux fois par deux personnes différentes. Et des petits programmes ont été utilisés pour corriger semi-automatiquement les erreurs résiduelles.

Pour les mots composés, nous avons utilisé le système INTEX (Silberztein, 1993), en ôtant nombre de mots composés lexicaux et en ajoutant nombre de mots composés grammaticaux.

Le corpus ne contient aucune ambiguïté morphologique résiduelle.

1.4.2 Validation du corpus parsé

Le corpus étiqueté corrigé a été analysé automatiquement par l'analyseur de surface développé par A. Kinyon (adapté pour le format et les choix d'annotation par F. Toussenel) et corrigé à la main par des annotateurs.

L'analyseur de surface est incrémental, déterministe, à base de règles[2] et inspiré de celui d'Abney 1992 pour l'importance qu'il accorde aux mots fonctionnels (introducteurs de syntagmes) mais cependant différent de ce dernier en ce qu'il autorise certains enchâssements.

Le taux de réussite de l'analyseur de surface est estimé à 92% de frontières ouvrantes correctes, et 62% de frontières fermantes correctes. Ce qui est cohérent avec notre parti-pris d'attachement minimal des syntagmes.

2 Quelques résultats

Nous montrons d'abord, sur quelques statistiques globales, comment les différents niveaux d'annotations, permettent d'enrichir les calculs de fréquences sur corpus.

Puis, en comparant certaines fréquences relatives pour des formes possiblement ambiguës, nous dégageons quelques principes de préférence, qui sont nouveaux ou confirment des principes observés par ailleurs en linguistique générale ou en psycholinguistique.

2.1 Fréquences

Les fréquences lexicales du français ont souvent été calculées à partir de données brutes (Catach 84, Julliand 70). Comme le montre l'exemple de Silberztein 1993, de tels calculs sont nécessairement erronés du fait de la proportion élevée de formes ambiguës.

Voyons comment la désambiguation des parties du discours opérée sur notre corpus améliore ces calculs.

Si nous trions ces formes par ordre de fréquence, nous obtenons la liste de la seconde colonne (tableau 1) comme les formes les plus courantes, qui comprennent seulement des mots fonctionnels (prépositions, déterminants, conjonctions) et qui sont comparables avec ce que d'autres auteurs trouvent dans différents corpus français. Mais la plupart de ces formes sont en fait

[2] L'analyseur de surface est basé sur un automate déterministe à pile.

ambiguës : *de* peut être une préposition ou un déterminant, *le* peut être un déterminant ou un pronom, *en* peut être une préposition ou un pronom clitique. Si on s'intéresse aux mots les plus courants dans le corpus, il est donc nécessaire, d'une part de distinguer ces formes ambiguës et, d'autre part de rassembler différentes flexions du même mot (*d'* et *de* pour la préposition DE, *le, la, les, l'* pour le déterminant LE, etc.).

En faisant cela et en triant les formes par lemmes (désambiguïsés), nous obtenons la liste dans la troisième colonne qui est tout à fait différente. Le mot le plus courant est le déterminant LE et quelques verbes (être, avoir) se trouvent parmi les 10 mots les plus fréquents.

Ordre de fréquence	**Par forme**	**Par lemme**		**Par partie du discours**	
1er	de (Prép ou Dét)	LE (le,la,les,l')	Det	Noms	24,5% (20% NC, 4,5% NP)
2e	le (Dét or CL)	de (de,d')	Prep	Déterminants	16,8%
3e	les (Dét or CL)	à	Prep	Prépositions	14,6%
4e	la (Dét or CL)	un (un, une, des, de, d')	Det	Ponctuations	13%
5e	à	être (suis, est etc)	V	Verbes	11,4%
6e	l' (Dét or CL)	et	CC	Adjectifs	6,5%
7e	et	avoir (ai, a etc)	V	Adverbes	4%
8e	en (Prép or CL)	il (il, ils, elle, elles)	CL	Conjonctions	3,3% (2,3% CC, 1% CS)
9e	un	en	Prep	Clitiques (pronoms faibles)	2,8%
10e				Autres pronoms	1,8%

Table 1: Fréquences lexicales par forme, par lemme et par catégorie (partie du discours)

Si maintenant nous trions les catégories elles-même (quatrième colonne du tableau), la répartition des catégories montre que la catégorie la plus représentée est le nom (24,5% soit près du quart des occurrences), les verbes sont moitié moins nombreux et les adjectifs eux-mêmes moitié moins nombreux que les verbes. Contrairement à ce que laissent croire les calculs sur les fréquences lexicales (où de et le arrivent toujours en tête), les mots grammaticaux ne sont pas les plus fréquents du corpus.

2.2 La préférence lexicale pour les mots composés

Lors de la phase de segmentation en mots, nous avons d'abord vérifié les préférences pour les mots composés bien connues (des psycholinguistes, cf (Gibs, 1985)). Nous avons pris les séquences qui sont éventuellement ambiguës entre les mots composés et les séquences en mots simples et avons calculé leur nombre d'occurrences respectifs. Voici des exemples de telles paires :

en fait : adverbe composé ou pronom clitique en suivi du verbe faire **d'ailleurs** : adverbe composé ou préposition d' suivie du nom ailleurs

Quelques résultats sont montrés dans le tableau 2.

Forme	**Nombre d'occurrences en mots composés**	**Nombre d'occurrences en mots simples**
d'abord	154 (97 %) Adv	5 (3%) : Prep NC
alors que	231 (96%): CS	8 (4%): Adv CS
plus de	305 (60%) Prep	(40%) Adv Prep or Det
il y a	221 (57%) : Prep	(43%) : CL CL V
le plus	123 (39%) Adv	(61%) Det Adv
sur ce	0 (Adv)	65 (100%) Prep Det

Table 2: Proportion relative des catégories des mots simples et composés

La préférence est attestée (plus de 93% des occurrences sont un mot composé en moyenne) mais dépend des catégories impliquées. Pour les mots composés nominaux et verbaux (comprenant généralement des noms, des verbes et des adjectifs), l'interprétation en mots composés concerne presque 100% des occurrences. Pour les mots composés adverbiaux, la préférence est moindre, et il y a des exceptions comme "sur ce" ou "le plus" dans le tableau 2. Cette différence peut être expliquée par une préférence pour les catégories grammaticales (clitique, déterminant, préposition... voir plus bas) associées avec les mots impliqués dans l'interprétation décomposée.

Nous vérifions que la préférence pour l'interprétation en mot composé est une préférence lexicale parce que le nombre total d'occurrences de mots composés dans le corpus est bien plus bas que celui des mots simples (50614=6,2% versus 765953=93,8%, sans compter la ponctuation).

2.3 La préférence lexicale pour les catégories grammaticales

Quand on considère les formes ambiguës syntaxiquement, les probabilités des différentes parties du discours sont généralement très inégales (cf (Church, 1988)).

Indépendamment des préférences syntaxiques qui peuvent être associées à telle ou telle unité lexicale (par exemple *ferme* est plutôt verbe que nom ou adjectif en contexte de langue générale), nous avons cherché des principes de préférence plus généraux, qui peuvent être utiles pour développer des étiqueteurs en partie du discours automatiques mais qui peuvent aussi mettre en lumière des stratégies humaines d'analyse.

Au niveau de l'étiquetage (ou de la désambiguation des parties du discours), nous avons trouvé une forte préférence pour les catégories grammaticales par rapport aux catégories lexicales. Nous entendons par catégories grammaticales les listes fermées de mots fonctionnels (déterminants, prépositions, pronoms clitiques et autres pronoms, conjonctions de subordination et de coordination), et par catégories lexicales les verbes, noms, adverbes et adjectifs. Nous avons pris l'ensemble des formes ambiguës entre ces deux classes et avons étudié les fréquences respectives de leurs occurrences. Voici quelques exemples :

CAR : Conjonction de coordination ou nom commun masculin singulier
OUTRE : Préposition ou nom commun féminin singulier
ENTRE : Préposition ou verbe entrer

En tout, nous avons trouvé une proportion écrasante d'emplois de catégories grammaticales (plus que 95% en moyenne, parfois 100%). Quelques uns font exception ; leur forme lexicalisée est particulièrement rare ou réservée à une langue de spécialité (dans pluriel de *dan*, *par* nom commun – terme de golf, *sur* adjectif synonyme d'acide, etc.)

Cette préférence est confirmée par des interrogations d'autres corpus, en particulier de corpus transcrits (Claire Blanche-Benveniste, communication personnelle). Dans la mesure où elle est stable, elle peut être considérée comme représentative des stratégies de désambiguation humaine.

Formes ambiguës	Nombre total d'occ.	Nombre d'occurrences	
		lexicale	grammaticale
car	235	5 (2,1%) Nom	230 (97,8%) C coord
cela	284	1 (0,3%) Verbe	283 (99,7%) Pronom
dans	5341	0 (0%) Nom	5341 (100%) Preposition
devant	285	33 (11,5%) Verbe	252 (88,4%) Preposition
entre	1195	23 (1,9%) Verbe	1172 (98%) Preposition
envers	25	3 (12%) Nom	22 (88%) Preposition
la	24471	1 (0.0%) Nom	24470 (100.0%)
lui	763	0 (0%) part. de luire	763 (100%) pronom clitique et pronom fort.
or	189	30 (15.9%) Nom masc	159 (84.1%) Conjonction de coordination
si	989	0 (0%) Nom masc	989 (100%) Conjonction de subordination, Adverbe
son	2427	10 (0.4%) Nom masc	2417 (99.6%) Déterminant
sous	359	25 (6.9%) nom masc pluriel	334 (93.1%) préposition
ton	31	22 (70.9) nom masc sing	9 (29.1%) déterminant

Table 3: Fréquence relative par catégorie de quelques formes ambiguës

2.4 Préférences fonctionnelles et structurales

2.4.1 La hiérarchie d'accessibilité des fonctions

Si on regarde la fréquence relative des pronoms relatifs, dont la fonction est marquée par la morphologie, on obtient les résultats suivants (sur le corpus entier) :

Pronoms relatifs :

Sujet (qui sans préposition)	6291	61%
Objet direct (que ou qu')	1565	15,2%
Génitif (dont)	1076	10,4%
Locatif (où)	782	7,6%
Objet indirect (prép+qui, quoi, lequel)	539	5,2%
autres		0,3%

Cette répartition rappelle celle trouvée par (Keenan & Hawkins, 1987) sur des textes journalistiques anglais (sujet 46%, objet direct 24%, objet indirect 15%, génitif 5%, autres 10%), ce que confirme la hiérarchie d'accessibilité relative universelle de Keenan et Comrie. En français, la préférence pour les relatifs sujet est bien plus forte que pour l'anglais, et la fréquence relative du génitif (*dont*) est également plus grande, peut-être à cause de l'emploi fréquent de subordonnées relatives en *dont* avec un pronom résomptif dans les journaux français (*Un problème dont on sait qu'il est difficile*).

Ces fréquences relatives confirment les resultats obtenus par ailleurs en psycholinguistique par (Holmes & O'Regan, 1981) comparant les relatives en *qui* et en *que* (taux de compréhension et mouvement oculaire).

2.4.2 Préférences pour l'attachement des propositions relatives

Nous avons recherché les occurrences de propositions relatives dans le contexte suivant : N1 (adj) prep (det) (adj) N2 (adj) relative. Sur 370 relatives trouvées, 52,2% sont attachées au premier nom, et on trouve la préférence inverse (44%) pour les relatives courtes (< 6 mots).

Ces résultats confirment les travaux de (Zagar *et al.*, 1997) et de (Pynte, 1998) sur la préférence pour l'attachement haut en français (à la différence de l'anglais) et le rôle de la longueur de la relative.

Bien qu'il faille être prudent sur la reproductibilité de ce type de résultats, ceux-ci sont prometteurs pour la contribution de la linguistique de corpus à la psycholinguistique (contra (Gibson & Schütze, 1999)).

Conclusion

Nous avons présenté un corpus annoté syntaxiquement pour le français, pleinement désambigué et validé manuellement, et quelques investigations préliminaires. Certaines de ces investigations ont confirmé des fréquences et des préférences lexicales bien connues, d'autres ont apporté la lumière sur de nouvelles fréquences et de nouvelles préférences qui devraient être confirmées sur d'autres corpus.

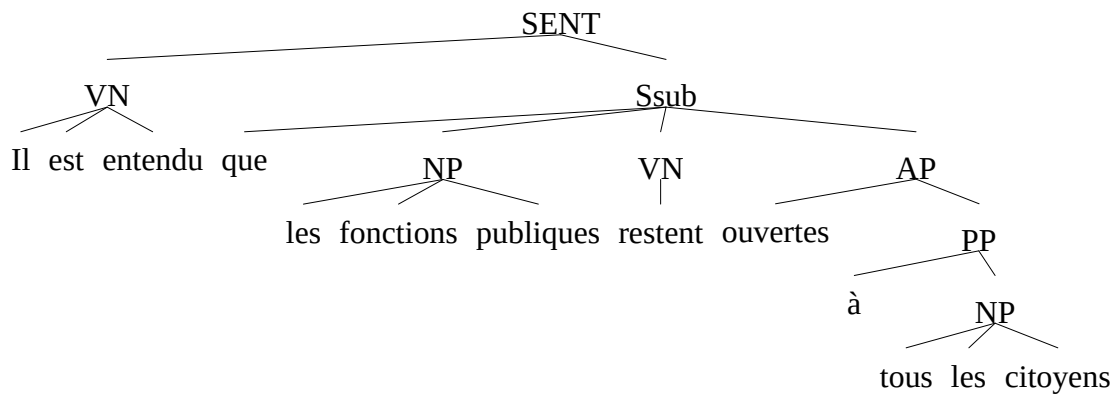
D'autres enquêtes pourront être menées, en collaboration avec des linguistes ou des psycholinguistes. Des comparaisons devront aussi être faites avec d'autres corpus arborés pour d'autres langues.

Des annotations futures impliquent l'assignation d'une fonction grammaticale pour chaque syntagme majeur. Ceci permettra davantage d'investigations, par exemple sur l'inversion du sujet.

Le corpus est distribué en tant que ressource linguistique et est déjà utilisé par quelques équipes en France et ailleurs.

A Échantillon du corpus arboré

```
<SENT>
   <VN> <w lemma="il" ei="CL3ms" ee="CL-suj-3ms" cat="CL" subcat="suj" mph="3ms">Il</w>
      <w lemma="être" ei="VP3s" ee="V–P3s" cat="V" subcat="" mph="P3s">est</w>
      <w lemma="entendre" ei="VKms" ee="V–Kms" cat="V" subcat="" mph="Kms">entendu</w>
   </VN>
   <Ssub>
      <w lemma="que" ei="CS" ee="C-S" cat="C" subcat="S">que</w>
      <NP> <w lemma="le" ei="Dfp" ee="D-def-fp" cat="D" subcat="def" mph="fp">les</w>
         <w lemma="fonction publique" ei="NCfp" ee="N-C-fp" cat="N" subcat="C" mph="fp">
            <w catint="N">fonctions</w> <w catint="A">publiques</w> </w>
      </NP>
      <VN> <w lemma="rester" ei="VP3p" ee="V–P3p" cat="V" subcat="" mph="P3p">restent</w> </VN>
      <AP> <w lemma="ouvert" ei="Afp" ee="A-qual-fp" cat="A" subcat="qual" mph="fp">ouvertes</w>
         <PP> <w lemma="à" ei="P" ee="P" cat="P">à</w>
            <NP> <w lemma="tout" ei="Amp" ee="A-ind-mp" cat="A" subcat="ind" mph="mp">tous</w>
               <w lemma="le" ei="Dmp" ee="D-def-mp" cat="D" subcat="def" mph="mp">les</w>
               <w lemma="citoyen" ei="NCmp" ee="N-C-mp" cat="N" subcat="C" mph="mp">citoyens</w>
            </NP></PP></AP>
   </Ssub>
   <w lemma="." ei="PONCTS" ee="PONCT-S" cat="PONCT" subcat="S">.</w>
</SENT>
```



Références

ABEILLÉ A., CANDITO M.-H. & KINYON A. (1999). Ftag : current status and parsing scheme. In *VEXTAL'99*, Venise.

ABEILLÉ A. & CLÉMENT L. (1997). *Désambiguation morphosyntaxique; 1 Les mots simples; 2 Les mots composés.* TALANA, Université Paris 7.

ABEILLÉ A., CLÉMENT L. & KINYON A. (2001). *Building a Treebank for French*, In *Building and using syntactically annotated corpora*. Kluwer Academic Publishers.

ABNEY S. (1990). *Parsing by Chunks*, volume Principle-based Parsing. Kluwer, berwick *et al.* edition.

CHURCH K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *2nd ANLP Conference*, p. 136–143, Austin.

CLÉMENT L. & KINYON A. (2000). Chunking, marking and searching a morpho-syntactically annotated corpus for french. In *ACIDCA*, Monastir, Tunisia.

GIBS R. (1985). On the process of understanding idioms. *Journal of Psycholingistic Research*, **14**.

GIBSON E. & SCHÜTZE C. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language*, **40**, 263–279.

GROSS G. (1996). *Les expressions figées en français.* Ophrys.

HABERT B., NAZARENKO A. & SALEM A. (1997). *Les linguistiques de corpus.* Armand Colin.

HOLMES V. & O'REGAN J. (1981). Eye fixation patterns during the reading of relative clause sentences. *Journal of Verbal Learning and Verbal Behaviour*, **20**, 417–430.

KEENAN & HAWKINS (1987). The psychological validity of the accessibility hierarchy. In KEENAN, Ed., *Universal Grammar*. Routledge London.

LECOMTE J. & PAROUBEK P. (1996). *Le catégoriseur d'E Brill : mise en oeuvre d'une version entraînée pour le français.* Rapport interne, INaLF, Nancy.

PYNTE J. (1998). The time-course of attachment decisions: Evidence from french. *Syntax and Semantics*, **31**, 227–245.

SILBERZTEIN M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

VON REKOWSKI U. (1996). Elm-fr : Specifications for french morphosyntax, lexicon specification and classification guidelines. EAGLES document.

ZAGAR D., PYNTE J. & RATIVEAU S. (1997). Evidence for early-closure attachment on first-pass reading times in french. *The Quarterly Journal of Experimental Psychology*, **50**, 421–438.