

Les corpus de français parlé Elicop : consultation et exploitation

PIET MERTENS

K.U.Leuven

Abstract

Cet article présente sommairement la collection de corpus de français parlé Elicop et décrit les travaux réalisés pour en faciliter l'accès et la consultation par des utilisateurs non initiés. Il aborde successivement la normalisation des formats et des annotations, l'enrichissement du corpus grâce à l'étiquetage grammatical et à la lemmatisation, et enfin la consultation ciblée des corpus à partir des propriétés lexicales, morphologiques et syntaxiques. Ces outils de consultation sont accessibles par le réseau Internet.

1 Introduction

Les corpus de langue parlée, transcriptions fidèles de conversations authentiques, constituent une fabuleuse source d'informations pour l'étude du français, des particularités de l'oral, de la phonétique, de la conversation et de ses procédés, pour la sociolinguistique et même pour la recherche sur l'apprentissage d'une langue étrangère, notamment si les locuteurs enregistrés sont des apprenants. La préoccupation didactique de Mark Debrock l'a amené à réunir, à constituer et à exploiter de tels corpus dans le cadre de deux projets de recherche, l'un sur le français parlé (Elilap, "Étude linguistique de la langue parlée", 1980-1983), l'autre sur le français des néerlandophones (LanCom, "Langue et Communication", 1993-2001). L'essor de l'informatique et la généralisation du réseau Internet permettent de mettre ces ressources à la disposition de tous, sans la moindre barrière technique. Cependant, pour y arriver, une foule de problèmes techniques doivent être réglés. D'abord il faut réunir les corpus hétérogènes dans une collection à format uniforme (normalisé). Ensuite, il s'agit de mettre au point des logiciels pour leur consultation et leur enrichissement. Enfin, il reste à réaliser des interfaces utilisateur conviviales pour l'interrogation des corpus par l'Internet. Le résultat de cet effort, mis sur pied par Mark Debrock, a été baptisé Elicop (Étude linguistique de la communication parlée).

Dans cette contribution, nous faisons le bilan des résultats réalisés au fil des années, en mettant l'accent sur la consultation à la fois conviviale et sophistiquée des corpus, sans doute l'aspect le plus intéressant pour l'utilisateur.

2 La collection de corpus ELICOP

La collection Elicop rassemble des corpus de français parlé réalisés au cours de plus de 30 années. Ces corpus portent sur des enregistrements sonores ou vidéo réalisés entre 1968 et 2001. Il s'agit des corpus suivants: "Étude sociolinguistique sur Orléans" (1968 - 1971), "Le livre parlé de Tours" (1974), "Voix d'Auvergne" (1976) et "Langue et Communication" (1993-2000).

Une partie importante de ce matériau sonore a été transcrite sous forme textuelle (transcription graphique) et dans une moindre mesure sous forme de transcription phonétique. Ce travail a été effectué dans le cadre de projets de recherche menés

aux universités de Leuven¹ et d'Amsterdam². Le tableau 1 ci-dessous énumère quelques données sur la taille des corpus transcrits repris dans Elicop. Des informations plus détaillées (sur la nature des corpus, les circonstances de leur création, la situation de communication, les consignes d'enregistrement, etc.) sont disponibles sur le site de Elicop ou dans DEBROCK *et al.* (2000).

	transcription		
	graphique	graphique	phonétique
	heures	mots	heures
Orléans	80	902.631	11
Tours	4	36.508	0
Auvergne	17	176.665	0
total	101	1.115.804	11
Lancom FF	2	28.732	0
Lancom FB	14	19.366	0
Lancom FN	1h30	111.574	0
total	17h30	159.672	0
	118h30	1.275.476	11

Tableau 1. Inventaire des corpus de la collection Elicop

3 La normalisation des corpus.

Comme, très souvent, ces corpus ont été constitués en vue de l'étude phonétique, la plupart des transcriptions originales comportent des annotations de phénomènes de prononciation : pauses, hésitations, présence ou absence de la liaison, enchaînements consonantiques, réalisation ou omission du e-muet ou de la détente consonantique, sons escamotés, allongements marqués, timbre

1 Il s'agit d'une part du projet connu sous le sigle ELILAP, "Etude linguistique de la langue parlée", effectué entre 1980 et 1983, auquel participaient Josse De Kock, Mark Debrock, Nicole Delbecque et Ellen Bas, et d'autre part du projet Lancom, "Language et communication", auquel ont participé Mark Debrock, Danièle Flament-Boistrancourt, Piet Mertens, Veerle Brosens, Fred Truyen, ainsi que de nombreux étudiants de philologie romane.

2 Vrije Universiteit Amsterdam

vocalique particulier, contours intonatifs caractéristiques, et ainsi de suite. Lorsque ces annotations se présentent comme des écarts par rapport à l'orthographe normale, elles posent des problèmes épineux, surtout si les signes d'annotation apparaissent à l'intérieur des mots. En effet, l'utilisateur doit alors maîtriser le moindre détail du système d'annotation non seulement pour retrouver les occurrences des phénomènes phonétiques notés, mais aussi pour une tâche aussi banale que la constitution de la concordance.

Les corpus prévoient généralement un système de renvoi basé sur la numérotation des lignes ou des interventions (tours de parole) et sur l'identité du locuteur. Le corpus LanCom, qui vise l'étude des erreurs de langue chez des apprenants du français langue étrangère, prévoit en outre une annotation des erreurs de langue. On est donc en présence de plusieurs types d'annotation. L'enrichissement des corpus par un étiquetage grammatical (cf. infra) rajoutera un niveau d'annotation supplémentaire.

Qui plus est, les conventions de transcription varient considérablement d'un corpus à l'autre. Certaines différences, comme la représentation des caractères à signes diacritiques (à, â, é, è, ê, ç ...), s'expliquent par les contraintes informatiques de l'époque et leur normalisation pose peu de problèmes. Face à la multiplicité des annotations (particularités phonétiques, identification du passage et du locuteur, analyse d'erreurs, étiquetage grammatical) et des conventions de transcription, il est indispensable d'élaborer une nouvelle convention unique, qui évite les problèmes cités, puis de normaliser l'ensemble des corpus dans ce format uniforme.

Dans la convention d'annotation adoptée dans Elicop, tout élément d'annotation, quel que soit sa nature, prend la forme d'une balise au format SGML. Nous ne nous attarderons pas ici aux détails techniques de cette norme. Pour la suite il suffit de préciser que ces balises se reconnaissent facilement aux crochets pointus qui les entourent. Leur intérêt se situe entre autres dans les points suivants : il est très aisé (d'un point de vue informatique) de séparer le texte brut des balises, et dès lors de faire des recherches sur la forme textuelle seulement ; l'ajout ultérieur de balises ne pose aucun problème ; contrairement à d'autres types d'annotation, le décodage des balises n'est pas lié à la place qu'elles occupent dans le fichier.

4 La concordance et l'utilisation d'expressions régulières

Depuis les débuts de l'utilisation des corpus, le concordancier constitue l'outil essentiel pour la consultation des corpus. Pour une forme introduite par l'utilisateur, ce logiciel fournit la liste des occurrences dans le corpus, en affichant pour chacune le contexte dans lequel elle apparaît et éventuellement un renvoi permettant de localiser l'occurrence dans le corpus. On trouve ci-dessous [Figure 1] un extrait de la concordance pour le mot “disons”; en tête de ligne figure un chiffre qui correspond au numéro de la ligne dans le fichier traité.

142	assez importante et euh	disons que j' ai le mal autour de moi
162	il y a beaucoup de jalousie	disons que chacun essaie de pousser le
181	cette différence souvent euh	disons qu' on euh comment je pourrais
225	disons que ils me comprennent	disons ils acceptent ils comprennent
225	ils comprennent pas euh	disons que ma mère elle est elle est
299	mais en fin de compte euh	disons que ce sont des évangélistes
311	argot entre eux mais oui	disons que voyez euh il y a des drogués
382	mais en fin de compte euh	disons que on peut disons euh avoir
382	compte euh disons que on peut	disons euh avoir déjà notre vie qui
409	peu près votre réponse ben	disons que moi je veux marcher toujours
416	sans avoir la certitude	disons que j' ai envie de partir hein
416	ai envie de partir hein mais	disons que pour avoir cette certitude
455	quinzième siècle mais enfin	disons que les les plus vieilles
560	mais ça c' est uniquement	disons au niveau de l' animation
573	les cours de l' amour puisque	disons les cours d' amour arrivaient
584	secondaires si l' on veut	disons parce qu' il ne peut pas y avoir
590	eh bien une entreprise ou	disons contre leurs propriétaires eh
608	mais il faut que enfin	disons faut qu' il y ait mis des
618	euh je pense que le	disons au niveau de la rue si vous
652	pendant d' une façon active	disons en championnat pendant plus de
671	ah les jours fériés et on a	disons on procure des loisirs aux
994	que euh les femmes euh	disons sont aussi intelligentes que les
1275	enfin je suis clemontoise	disons voilà et vous avez eu la
1361	dernière ma petite dernière	disons qui a vingt -huit ans disons c'
1481	c' est pratiquantes	disons euh tous les dimanches elles se
1529	dans un milieu plus aisé	disons mais peut -être qu' il y a ça
1867	entre quatre et cinq quoi	disons nous il y en a qui sortent à
2977	a donné aux euh gens de	disons de qu' une formation assez

Figure 1. Extrait de concordance pour la forme “disons”.

Comme chaque mot du lexique présente plusieurs formes fléchies, du moins pour la plupart des catégories grammaticales, la concordance d'un lemme nécessite des requêtes pour chacune de ces formes, ce qui est assez fastidieux.

La facilité d'utilisation du concordancier augmente considérablement par l'emploi, à la place de la simple forme littérale, d'une *expression régulière*. Il s'agit d'une formule qui permet de spécifier plusieurs chaînes de caractères à la fois et dès lors plusieurs formes. En dehors de l'alphabet normal, une expression régulière fait intervenir des signes supplémentaires avec des fonctions précises : elles permettent d'indiquer des alternatives pour une même position dans la chaîne, d'indiquer le caractère facultatif d'une ou de plusieurs positions, et ainsi de suite. Ainsi, le point '.' indique un caractère quelconque: l'expression "p.rte" couvre les chaînes "porte", "perte", "parte", ainsi que toutes les chaînes où ces séquences apparaissent: "porterons", "reporter", "département", et ainsi de suite. Les crochets gauche et droit entourent les caractères alternatifs pour une même position dans la chaîne. Par exemple, l'expression "soi[st]" couvre les formes "sois" et "soit" (ainsi que toutes les chaînes comportant ces séquences). D'autres signes indiquent le nombre de caractères ou, d'une manière générale, la longueur d'une chaîne : l'élément suivi du plus '+' peut se répéter plusieurs fois, celui suivi de l'astérisque '*' peut apparaître plusieurs fois ou être absent. L'expression "sot+e*" couvre ainsi les formes "sot", "sote", "sott" et "sotte". Ce dernier exemple montre également la possibilité de combiner plusieurs signes spéciaux dans une même expression. Certains systèmes permettent d'explicitement les frontières de mot dans l'expression, ce qui permet d'affiner la requête et dès lors de réduire le nombre de fausses alertes. L'emploi d'expressions régulières ne se limite d'ailleurs pas à la recherche de mots isolés ; elles peuvent tout aussi bien être utilisées pour spécifier des suites de mots, puisqu'elles peuvent être appliquées à toute séquence de caractères, comportant éventuellement des blancs et des signes de ponctuation.

Les expressions régulières constituent ainsi un moyen performant pour faciliter l'utilisation du concordancier. Cependant, dans la pratique, à moins d'être des informaticiens ou des passionnés de l'informatique, les utilisateurs ont des difficultés évidentes à s'en servir ou à s'en servir efficacement.

5 Les corpus étiquetés

Dans le concordancier les recherches d'occurrences s'effectuent toujours sur le texte, c'est-à-dire sur la forme orthographique. Or, en syntaxe les questions ne portent pas essentiellement (ou du moins pas exclusivement) sur des mots

particuliers, mais plutôt sur des catégories de mots ou sur des entités plus grandes, tels que les constituants ou les constructions. Il nous faut en conséquence un outil d'interrogation qui permette de formuler des requêtes en ces termes-là. En outre, comme toute construction syntaxique comporte plusieurs éléments successifs, l'outil doit également permettre de spécifier (les propriétés de) plusieurs éléments successifs. L'objectif est donc d'avoir la possibilité de formuler des requêtes portant sur la forme des mots, leur lemme, et leur catégorie grammaticale, et cela pour une chaîne de plusieurs éléments ou *tokens* (le token correspond à un élément du texte délimité par un blanc, par un séparateur, par un signe de ponctuation ou par certains caractères particuliers tel que l'apostrophe; il s'agit de mots au sens traditionnel, de signes de ponctuation ou d'annotations).

Afin de réaliser cet objectif, le corpus textuel doit être enrichi d'informations lexicales et grammaticales. Vu la taille des corpus, ce travail ne peut pas se faire à la main. Des logiciels spécifiques ont été conçus pour cette tâche ; ils sont appelés étiqueteurs morpho-syntaxiques (*part-of-speech tagger*). C'est après avoir présenté les étiquettes et le résultat obtenu, que nous expliquerons le fonctionnement de l'étiqueteur lui-même.

L'inventaire des étiquettes

Tout système d'étiquetage adopte un inventaire d'étiquettes particulier (dans le jargon on parle de *tagset*). D'une manière générale cet inventaire varie d'un système à l'autre, en fonction des choix linguistiques, du détail de l'analyse morphologique et de la langue traitée. Si on se limite aux catégories grammaticales majeures (verbe, nom, adjectif, adverbe, déterminant, préposition, pronom, conjonction, interjection), on obtient une dizaine d'étiquettes ; quand on ajoute la catégorie mineure (par exemple: déterminant article, démonstratif, indéfini, possessif,...), soit deux critères, on passe à une trentaine d'étiquettes ; enfin, pour l'ensemble des traits morphologiques du français (mode, temps, personne, nombre, genre, ...) il en faut plus de deux cents. Dans le cas du français, il y a eu des propositions de standardisation de l'inventaire d'étiquettes (par exemple dans les projets Multext¹ et Grace¹), mais elles sont largement tributaires de l'analyse

1 Le projet Multext visait la standardisation des données, des outils et des ressources linguistiques en vue de faciliter leur réutilisation dans le traitement du langage naturel pour l'analyse de corpus et pour la réalisation d'applications. Le site <http://www.lpl.univ-aix.fr/projects/multext/> fournit toutes les informations sur le projet ainsi que sur les conventions d'étiquetage grammatical proposées.

syntaxique et morphologique adoptée, le plus souvent celle de la grammaire scolaire. Le projet Elicop propose son propre inventaire, cependant très proche de la norme Grace. Par ailleurs, la constitution de l'inventaire des étiquettes soulève une foule de problèmes d'analyse linguistique qu'on a l'habitude d'oublier.

Le tableau ci-dessous [Tableau 2] montre le nombre d'étiquettes retenues selon qu'elles se limitent aux catégories majeures, aux catégories majeures et mineures ou qu'elles explicitent l'ensemble des traits morphologiques.

	catégorie majeure	+ catégorie mineure	maximum de traits
verbe	V	6	103
nom	N	2	8
adjectif	A	3	24
adverbe	R	2	2
déterminant	D	7	36
pronom	P	6	24
préposition	S	3	3
conjonction	C	2	2
numéral	U	1	1
interjection	I	1	1
résidu	X	1	1
punctuation	F	1	1
extra-lexical	?	1	1
Nombre total d'étiquettes	13	36	207

Tableau 2 Nombre d'étiquettes dans le *tagset*, en fonction des critères retenus.

Nous présentons ci-dessous [Figure 2] un extrait de corpus étiqueté, à titre d'illustration. Chaque ligne du corpus étiqueté correspond à un élément du texte d'entrée ; il s'agit soit de mots, soit d'annotations. Les dernières se notent sous forme de balise. Sur une même ligne, l'élément de l'entrée sera suivi de son

1 L'action GRACE visait l'évaluation de plusieurs analyseurs morpho-syntactiques et syntaxiques du français. La première session d'évaluation a porté sur les étiqueteurs pour le français. Les résultats et les conventions pour les étiquettes sont disponibles à l'adresse suivante: <http://www.limsi.fr/TLP/grace/>

lemme et d'une étiquette représentant la catégorie grammaticale et les propriétés morphologiques (selon la convention mentionnée plus haut). Bien sûr ces éléments manquent pour les balises. L'ordre des mots et des annotations est maintenu scrupuleusement. L'extrait illustre également le traitement des locutions ("alors que") : les parties de la locution apparaissent sur des lignes séparées afin de préserver les annotations éventuelles entre elles. Ce traitement est appliqué à l'ensemble des locutions, qu'elles soient conjonctives ("bien que"), adverbiales ("sans doute"), prépositives ("aux alentours de"), ou autres.

```

<sp who="F05" nr=24>  sgml
mais                  mais:Cc
<ph_noliasion>       sgml
euh                   euh:I
disons               dire:Vmm-1p
que                  que:Cs
<ph_long>           sgml
<ph_pause l=1>      sgml
la                   le:Da-fs
véritable            véritable:Afpfs
rencontre           rencontre:Ncfs
avec                 avec:Sp
dieu                 dieu:Ncms
<ph_pause l=1>      sgml
faut                 falloir:Vmip3s
la                   le:Pp3fs
faire                faire:Vmn---
consciemment        consciemment:Rg
<ph_pause l=1>      sgml
hein                 hein:I
<ph_pause l=1>      sgml
alors                 alors que (1/2):Cs
que                  alors que (2/2):Cs
<ph_long>           sgml
<ph_pause l=1>      sgml
malheureusement     malheureusement:Rg
dans                 dans:Sp
le                   le:Da-ms
dans                 dans:Sp
certaines           certains:Di-fp
religions           religion:Ncfp
<ph_pause l=1>      sgml
on                   on:Pp3-s
baptise             baptiser:Vmip3s
un                   un:Da-ms
<ph_liaison>        sgml
enfant              enfant:Ncms
<ph_pause l=1>      sgml
</sp>                sgml

```

Figure 2 Extrait d'un corpus étiqueté

Le nombre élevé d'étiquettes (207) nous empêche de les énumérer toutes ici. Quelques exemples permettront d'illustrer leur interprétation. La forme "disons" reçoit l'étiquette "dire:Vmm-1p" qui indique qu'il s'agit d'une forme du verbe "dire" employé comme verbe (V) principal (m, pour *main verb*) à l'impératif (m), à la première personne du pluriel (1p). La forme "baptise" du lemme "baptiser" est un verbe principal (m) employé au présent (p) de l'indicatif (i), à la troisième personne du singulier (3s), ce qui donne l'étiquette "Vmip3s". "Véritable" est un adjectif (A) qualificatif (f) positif (p) féminin (f) singulier (s).

6. Méthodes pour l'étiquetage grammatical

Etant donné la facilité apparente avec laquelle l'être humain effectue ce genre d'étiquetage, on a tendance à sous-estimer sa complexité réelle d'un point de vue informatique. Celle-ci dépend de la langue individuelle (de sa morphologie flexionnelle, de son orthographe, etc.) et de l'inventaire d'étiquettes utilisé. Dans un corpus de français écrit, presque la moitié (43%) des mots (formes) sont ambigus : ils présentent plusieurs analyses possibles. HABERT *et al.* (1997:165-166) citent les travaux de EL BÈZE & SPRIET (1995:58) qui ont mesuré le nombre d'étiquettes possibles pour un même mot : "[...] une très grosse part de l'ambiguïté syntaxique est détenue par un petit nombre de mots fréquents [...]. De plus, ces mots sont essentiellement des mots outils. [...] 30 % de l'ambiguïté est détenue par les 8 mots ambigus les plus fréquents [dans le corpus utilisé par les auteurs il s'agit de "la, le, l', les, en, un, une, a"] (50 % par les 36 premiers) mais il faut traiter 1825 formes différentes pour lever 90 % de l'ambiguïté."

Les étiqueteurs morpho-syntaxiques comportent en général plusieurs modules ou étapes de traitement, à savoir l'analyseur morphologique, le lemmatiseur et la désambiguïsation. Alors que les deux premières étapes portent sur les éléments isolés, la dernière fait intervenir leur contexte. La première étape détermine les propriétés morphologiques (mode, temps, personne, nombre, genre, etc.) à partir des désinences, processus conduisant le plus souvent à plusieurs candidats. Le lemmatiseur, ensuite, identifie les lemmes possibles conduisant éventuellement à des candidats additionnels. Enfin, la désambiguïsation porte sur les séquences de candidats; il consiste à choisir parmi les candidats pour un seul mot celui qui convient dans le contexte où celui-ci apparaît. Cet aspect varie considérablement d'un système à l'autre. Tantôt il est basé sur la probabilité des séquences de

catégories grammaticales, de formes, de traits morphologiques, ou d'une combinaison de ces aspects, et cela dans un contexte local autour de chaque élément à étiqueter. Tantôt il fait intervenir une forme d'analyse syntaxique visant à reconnaître des constituants ou des phrases entières (analyse syntaxique partielle ou complète). (Pour une présentation des approches, voir PAROUBEK & RAJMAN (2000).)

Dans le cadre du projet Elicop, nous avons réalisé un système d'étiquetage grammatical qui effectue une analyse syntaxique de complexité variable, en fonction de la couverture de la grammaire utilisée : partant de la séquence des mots lemmatisés dotés de leurs propriétés morphologiques (ou plus exactement du réseau de candidats), on identifie des constituants de plus en plus grands, pouvant aller jusqu'à l'ensemble de la phrase, dans le meilleur des cas. L'analyse syntaxique fait appel à l'analyseur Vertex¹ et à une grammaire syntagmatique pour le français. Cette étape est précédée de la lemmatisation et de l'analyse morphologique fournie par le système Morlex². Ces deux logiciels, nous les avons déjà mis au point auparavant.

À la suite de ce traitement, on obtient un corpus étiqueté (enrichi d'étiquettes morphosyntaxiques) de façon automatique.

-
- 1 L'analyseur Vertex est un analyseur tabulaire (*chart*) de type ascendant, qui part des mots dans la phrase à analyser pour former des constituants de plus en plus grands. Il permet une séparation nette entre l'analyseur et la grammaire qu'on lui fournit. Il suffit ainsi de remplacer la grammaire pour pouvoir appliquer l'analyseur à une autre langue. La grammaire d'unification associée à chaque constituant (y compris les mots fléchis) a une structure de traits, ce qui permet de paramétrer les propriétés des constituants et de réduire le nombre de règles. En cela le système Vertex se rapproche du système PATR-II. Il s'en distingue par les points suivants: le fonctionnement bidirectionnel à partir de la tête syntaxique du constituant (plutôt que de gauche à droite), la possibilité de préciser dans une règle de la grammaire le caractère facultatif d'un sous-constituant, la possibilité d'ajouter à une règle des contraintes portant sur l'unification de structures de traits et/ou sur quelques fonctions prédéfinies vérifiant la présence d'un sous-constituant, par exemple. Ces propriétés visent à augmenter le pouvoir expressif de la grammaire et permettent une réduction du nombre de règles.
Cf. <http://bach.arts.kuleuven.ac.be/pmertens/vertex/>
 - 2 La base de données lexicales Morlex comporte plus de 33000 lemmes français. Elle s'accompagne de logiciels de génération et de lemmatisation. La génération permet d'obtenir toutes les formes fléchies d'un lemme (par exemple de conjuguer un verbe). La lemmatisation permet d'obtenir le lemme et les traits morphologiques pour une forme fléchie donnée.
Cf. <http://bach.arts.kuleuven.ac.be/pmertens/morlex/>

Le taux d'identification correcte est de l'ordre de 96 %, ce qui est comparable aux résultats obtenus par d'autres systèmes (cf. VÉRONIS (2000: 117)). Cependant, il varie considérablement selon la nature du corpus (langue écrite ou orale) et selon le nombre d'étiquettes utilisées (ou la partie de l'étiquette prise en considération). Si l'étiquette se limite à la catégorie grammaticale, le résultat sera bien sûr meilleur que quand on tient compte également des autres propriétés morphologiques. Plus le nombre d'étiquettes est élevé, plus il y a de chances que l'étiquette choisie soit fautive. Le taux d'erreur varie aussi suivant que l'on tient compte de l'identité du lemme ou pas.

Les fichiers du corpus Elicop, étiquetés automatiquement par le système décrit ci-dessus, sont ensuite validés manuellement, avant d'être consultés par les utilisateurs. En principe l'étiquetage peut s'effectuer automatiquement au moment même de la consultation des corpus, comme cela se fait d'ailleurs dans le système Intex (conçu par Max Silberstein, cf. SILBERZTEIN 1993), qui doit alors être installé sur la machine de l'utilisateur. Dans le cas d'un corpus comme Elicop, accessible par le réseau Internet, l'étiquetage préalable garantit des délais minimaux lors de la consultation. L'utilisation d'Internet nous libère des problèmes de la distribution des logiciels ; elle permet la centralisation des corpus et garantit donc une mise-à-jour aisée qui profitera immédiatement à l'ensemble des utilisateurs. La validation manuelle se justifie par l'objectif de la recherche linguistique, et par les utilisations répétées.

7 L'interface utilisateur

Grâce à l'étiquetage grammatical le corpus ainsi enrichi se prête désormais à des interrogations portant sur des aspects proprement linguistiques. Il s'agit maintenant de fournir un outil convivial qui permette à l'utilisateur profane – ou béotien – d'exploiter ces possibilités de façon transparente, sans avoir à connaître la façon dont ces informations sont représentées dans le corpus. Bon nombre de recherches d'occurrences font intervenir des séquences de plusieurs mots. Là aussi, il faut concevoir une façon d'indiquer ces séquences indépendamment de la façon dont elles sont notées dans le corpus.

L'outil d'interrogation de corpus que nous avons réalisé prend la forme d'un tableau dans lequel on remplit une ou plusieurs cases pour préciser les propriétés des éléments recherchés. Chaque colonne de ce tableau correspond à un mot (ou

plus précisément à un *token*) dans le corpus. Si on recherche les occurrences d'un seul élément, on se sert des cases de la colonne de gauche. Si en revanche on recherche une séquence de trois mots, on remplira des cases dans les trois premières colonnes. Les rangées correspondent aux types de propriétés. Pour chaque mot, l'utilisateur spécifie un ou plusieurs aspects, à savoir la catégorie grammaticale, le lemme ou la forme graphique. Pour la catégorie grammaticale, on sélectionne, parmi les possibilités indiquées, une catégorie majeure ou mineure. Pour le lemme et la forme graphique on entre soit la forme complète, soit une expression régulière qui sera alors comparée à la forme ou au lemme du corpus. Le formulaire permet ainsi de préciser les propriétés d'un ou de plusieurs mots successifs recherchés. Enfin on sélectionne le corpus à utiliser, le nombre d'occurrences souhaité et la taille (en mots) du contexte à afficher. Après l'envoi de la requête, un index des occurrences s'affiche. Il suffit ensuite de cliquer sur un élément de l'index pour obtenir l'occurrence choisie, dans son contexte. Pour passer à l'occurrence suivante, on clique sur un autre élément de l'index.

Pour obtenir la liste des emplois du conditionnel, par exemple, on sélectionne l'item "verbe au conditionnel" dans la première colonne du tableau, dans la case correspondant à la catégorie grammaticale. Les autres cases resteront vides, puisque dans cette requête il n'y a pas de contrainte sur la nature du lemme, ni sur la forme graphique.

Si on veut se limiter aux occurrences du verbe "pouvoir" au conditionnel, on ajoute à la requête précédente l'infinitif "pouvoir" dans la case réservée au lemme.

Imaginons une étude sur les adverbes déterminant un autre adverbe ; on commencerait probablement par une requête du type Adv + Adv, ce qui revient à sélectionner la catégorie adverbe dans les deux premières colonnes du tableau et à laisser vides les autres cases.

L'exemple ci-dessous [Figure 3] illustre la recherche de séquences à trois éléments, respectivement un déterminant, un adjectif et un nom, dans l'ordre donné. On obtient les occurrences des adjectifs utilisés en antéposition.

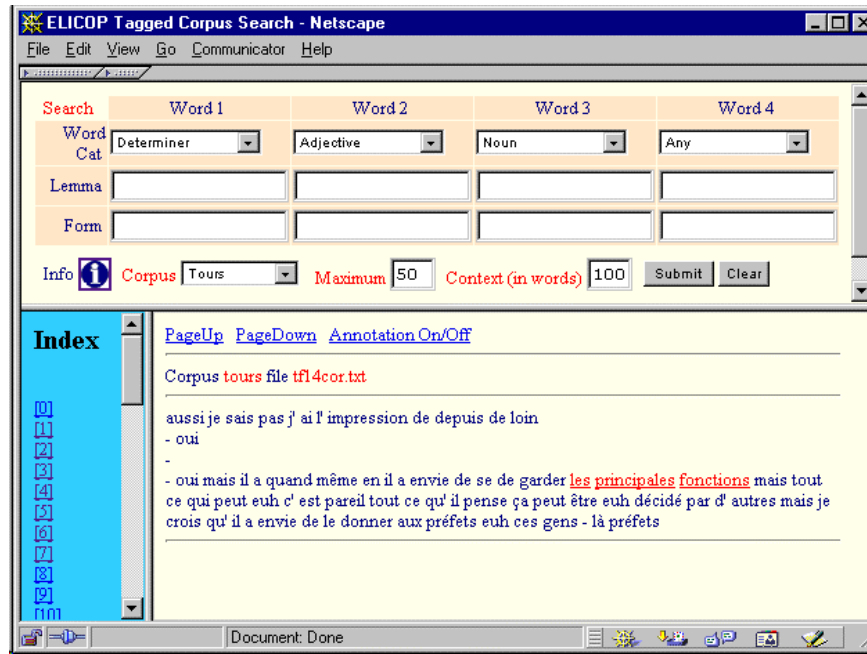


Figure 3 Illustration de l'outil de consultation de corpus étiqueté

Cet outil peut donc remplacer le concordancier classique ou étendu (utilisant les expressions régulières) dans la majorité des cas ; mais surtout il permet en même temps de formuler des requêtes beaucoup plus ciblées.

Dans la version décrite ci-dessous, l'outil d'interrogation n'exploite qu'une partie de l'information contenue dans les étiquettes. Cela s'explique une fois de plus par le nombre élevé d'étiquettes. Il serait cependant aisé de prendre en considération l'ensemble des informations morphologiques, par exemple en remplaçant la sélection à partir d'un menu par une spécification explicite ou par une expression régulière portant sur l'étiquette.

8 Conclusion et perspectives

Après une présentation très sommaire de la collection de corpus de français parlé Elicop, nous avons situé les difficultés à résoudre pour en faciliter l'accès et la

consultation par des utilisateurs non initiés à la linguistique informatique. Face à la diversité des formats et des annotations, une normalisation des corpus s'imposait. Celle-ci a été conçue pour préserver au maximum l'annotation déjà enregistrée dans les formats originaux. Ensuite le corpus a été enrichi par des informations linguistiques, grâce à un étiquetage grammatical et à une lemmatisation automatiques. Cette étape facilite la consultation ciblée des corpus sur le plan lexical, morphologique et syntaxique. Enfin, plusieurs outils de consultation ont été réalisés qui sont accessibles par le réseau Internet.

Parmi les extensions futures on peut envisager – suivant un ordre de complexité technique – l'ajout de corpus supplémentaires (existants ou nouveaux), la consultation des transcriptions phonétiques, l'écoute interactive de l'enregistrement sonore original de chaque occurrence, et le calcul automatique d'une transcription de l'intonation.

Dans leur livre, HABERT *et al.* (1997: 7) formulent les enjeux de la linguistique de corpus comme suit : “Ce qui est neuf, ce n'est pas l'utilisation des corpus électroniques. [...] La nouveauté réside dans l'enrichissement des corpus, l'accroissement de leur taille et dans l'accessibilité effective des corpus et des outils. D'abord, les corpus ne sont plus des suites de mots “nus”, c'est-à-dire de simples chaînes de caractères, mais ils sont *annotés* (ou encore *enrichis*).” La collection Elicop et ses outils de consultation semblent bien répondre à ces critères.

Références

- DEBROCK MARK & PIET MERTENS & FRED TRUYEN & VEERLE BROSENS (2000a):
“ELICOP, Etude Linguistique de la COmmunication Parlée : Constitution et exploitation d’un corpus de français parlé automatisé” Preprint nr. 172, K.U.Leuven, Departement Linguïstiek, 40 pp.
- DEBROCK MARK & PIET MERTENS & FRED TRUYEN & VEERLE BROSENS (2000b):
“ELICOP, Etude Linguistique de la COmmunication Parlée : Constitution et exploitation d’un corpus de français parlé automatisé (Annexes)” Preprint nr. 173, K.U.Leuven, Departement Linguïstiek, 78 pp.
- HABERT, BENOÎT; NAZARENKO, ADELIN & SALEM, ANDRÉ (1997) : *Les linguistiques de corpus*. Paris: A. Colin.
- PAROUBEK, PATRICK & RAJMAN, MARTIN (2000) : “Etiquetage morpho-syntaxique.” in: PIERREL, JEAN-MARIE (ed.) *Ingénierie des langues*. p. 131-150. Paris : Hermes.
- SILBERZTEIN, MAX D. (1993) : *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Collection “Informatique Linguistique”. Paris : Masson.
- VÉRONIS, JEAN (2000) : “Annotation automatique de corpus : panorama et état de la technique”. in: PIERREL, JEAN-MARIE (ed.) *Ingénierie des langues*. p. 111-130. Paris : Hermes.