Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe

Jacques Vergne

GREYC - Université de Caen
BP 5186 - 14032 Caen cedex
Jacques.Vergne@info.unicaen.fr
http://www.info.unicaen.fr/~jvergne

Résumé – Abstract

Nous présentons une méthode d'analyse descendante et calculatoire. La démarche d'analyse est descendante du document à la proposition, en passant par la phrase. Le prototype présenté prend en entrée des documents en anglais, français, italien, espagnol, ou allemand. Il segmente les phrases en propositions, et calcule les relations sujet-verbe dans les propositions. Il est calculatoire, car il exécute un petit nombre d'opérations sur les données. Il utilise **très peu de ressources** (environ 200 mots et locutions par langue), et le traitement de la phrase fait environ 60 Ko de Perl, ressources lexicales comprises. La méthode présentée se situe dans le cadre d'une recherche plus générale du Groupe Syntaxe et Ingénierie Multilingue du GREYC sur l'exploration de solutions minimales et multilingues, ajustées à une tâche donnée, exploitant peu de propriétés linguistiques profondes, la généricité allant de pair avec l'efficacité.

We present a method for top-down and calculatory parsing. The prototype we present is top-down from the document to the clause, through the sentence. Its inputs are documents in English, French, Italian, Spanish, or German. It tokenises sentences into clauses, and computes subject-verb links inside clauses. It is calculatory, as it executes few operations on data. It uses **very few resources** (about 200 words or locutions per natural language), and the sentence processing size is about 60 Kb Perl, including lexical resources. This method takes place in the frame of more general researches of the "Groupe Syntaxe et Ingénierie Multilingue du GREYC" into exploring minimal and multilingual solutions, close fitted to a given task, exploiting few deep linguistic properties, presuming that genericity implies efficiency.

Mots Clés – Keywords

analyse syntaxique, analyse descendante, analyse calculatoire, corpus multilingues
top-down parsing, calculatory parsing, multilingual corpora

1 Objectifs et cadre de cette recherche

Cette recherche se situe dans le cadre des travaux du groupe Syntaxe et Ingénierie Multilingue du GREYC. Nous cherchons et explorons des solutions minimales, c'est-à-dire pour une tâche donnée, ajuster et minimiser les moyens utilisés, trouver les moyens suffisants et nécessaires, avec l'ambition de qualité analogue ou meilleure que d'autres solutions actuelles. Nous pensons que la minimalité d'une solution est un signe de la bonne utilisation de quelques propriétés linguistiques profondes du matériau traité (seulement celles qui servent aux calculs de la tâche donnée) [1]. Des telles solutions minimales ont des algorithmes simples, de complexité pratique linéaire en temps; nous les appelons "calculatoires" car elles explicitent des opérations sur les données, au lieu d'explorer une combinatoire d'attributs issue de ressources lexicales quasi-exhaustives et de filtrer cette combinatoire par des grammaires formelles [2]. Ces solutions minimales utilisent **très peu de ressources** spécifiques à une langue donnée (**pas de dictionnaire** quasi-exhaustif), et ainsi facilitent les applications multilingues.

La conception et le développement d'applications multilingues a un double intérêt linguistique et informatique : du point de vue linguistique, on rejoint la préoccupation des linguistes d'abstraire la variété apparente des langues pour faire émerger des propriétés linguistiques indépendantes des langues; du point de vue informatique, on est conduit à dissocier le générique à plusieurs langues du spécifique à une langue; et d'un point de vue applicatif, il paraît difficile de ne pas avoir l'ambition de pouvoir traiter la variété des langues présentes sur Internet (remarquons que les moteurs de recherche généralistes ont cette ambition).

Enfin, la découverte des propriétés linguistiques, l'élaboration et la mise au point et les tests de ces solutions nécessitent un important travail sur corpus multilingues, corpus amplement disponibles aujourd'hui sur Internet [3].

2 Une tâche classique

Pour des raisons expérimentale et pédagogique, nous avons choisi une tâche limitée et (apparemment) simple : détecter et relier sujets et verbes, tâche classique qui permettra aux lecteurs des comparaisons avec des méthodes existantes.

Le prototype de notre "analyseur-relieur" est descendant du document à la proposition, en passant par la phrase. Il est multilingue (anglais, français, italien, espagnol, allemand). Il est

---

[1] Un autre exemple de ce type de travail, mais au niveau du document entier, est la solution minimale mise au point pour la détection automatique de la citation et du discours rapporté (Lucas, Giguet, Vergne, 2001).

[2] C'était l'objet de notre conférence invitée à TALN 2001 (Vergne, 2001) : *Analyse syntaxique automatique de langues : du combinatoire au calculatoire*.

[3] Nos corpus sont constitués d'articles de presse : articles en français de Ouest-France et du Courrier International, en anglais du Washington Post, en allemand du Spiegel, en espagnol de El Mundo, et en italien de La Stampa. Ces articles sont téléchargés du site de ces journaux dans le format html. On trouve par exemple de très nombreuses adresses sur le site de l'hebdomadaire "Courrier International" :
`http://www.courrierinternational.com/kiosk/kiosq.htm`

écrit en Perl (très bon outil de prototypage pour des applications de traitements de documents, en particulier grâce aux expressions régulières), et la partie "analyse de phrase" fait 40 Ko, avec des **ressources spécifiques aux 5 langues de 20 Ko environ au total**.

Un pré-traitement (qui sera décrit dans un article ultérieur) extrait le texte à analyser du fichier html (par comparaison des tables d'un ensemble de fichiers téléchargés d'une même source et extraction des cellules apax), puis il le représente sous la forme d'une unique chaîne de caractères, le translittère en un même jeu de caractères (iso 8859-1 pour l'instant, iso 10646/Unicode plus tard), en diagnostique la langue, et enfin le découpe en phrases, segment d'entrée de l'"analyseur-relieur".

3 Caractéristiques de la méthode

Appelons "**grain**" une classe de segments d'un document : on peut alors parler du grain-mot, du grain-phrase, du grain-paragraphe, etc. Ces grains peuvent former des hiérarchies inclusives, comme dans un fichier XML (ou en physique de la matière, ou en astrophysique), et appartenir à des hiérarchies différentes non réductibles à une seule hiérarchie inclusive : on peut distinguer les grains physiques (le donné de la typographie) des grains calculés au cours du processus d'analyse.

On peut alors caractériser le processus d'**analyse** comme un processus de passage des grains physiques (en entrée) aux grains calculés (en sortie) : traditionnellement, les grains physiques d'entrée sont la phrase et le mot, et le grain calculé en sortie est le syntagme.

Toute analyse commence par une segmentation physique de la donnée de départ (le document) en grains physiques d'entrée. Ensuite deux opérations sont possibles : soit le processus consiste à regrouper des grains physiques (processus appelé *analyse montante*), par exemple regrouper des mots en syntagmes, soit à découper des grains physiques (processus appelé *analyse descendante*), par exemple découper des phrases en propositions. Notons que les termes "montante" et "descendante" font implicitement allusion à un haut et un bas d'une hiérarchie des grains où l'unique gros grain (le document par exemple) est placé en haut (on peut y voir une trace de l'étymologie du mot hiérarchie), ce qui donne conventionnellement un arbre avec la racine en haut.

La méthode présentée est une méthode d'**analyse descendante** :

- le grain physique de départ est la graphie de phrase, obtenue par segmentation de la graphie du document entier par expressions régulières sur la ponctuation de fin de phrase ;
- le grain calculé en sortie est la proposition [4] et c'est le plus petit grain représenté dans une structure répétitive de propositions ; la proposition est l'espace de calcul, et sa graphie en

[4] Eva Ejerhed, dans (Ejerhed, 1996) a utilisé cette stratégie d'analyse (mais **avec** dictionnaire). À notre connaissance, elle n'a plus publié ultérieurement sur ce sujet, car, en 2000, son "procédé" était sur le point d'être breveté, pour être utilisé par la société de *Language technology* Hapax (`www.hapax.com`) qu'elle a créée. On trouvera un autre exemple de cette stratégie d'analyse sur l'égyptien ancien dans (Rosmorduc, 1994).

sortie est balisée sur les chunks sujets et les chunks verbes [5]. On ne tokenise pas jusqu'au grain chunk ni a fortiori jusqu'au grain mot [6] (ils ne sont pas représentés dans une structure répétitive de chunks ou de mots), mais les concepts de chunk et de mot sont utilisés dans le traitement : chunks et mots sont "vus d'en haut" du point de vue de la proposition (à l'aide d'un balisage posé par des expressions régulières).

C'est une **analyse de corpus multilingues** qui nécessite un diagnostic de la langue du document (supposé monolingue) pour choisir correctement les ressources spécifiques à cette langue. Ce diagnostic est simplement effectué par comptage des occurrences des graphies des ressources lexicales (spécifiques à chaque langue) dans le document; la langue du document est celle qui recueille le score le plus élevé. Dans la mesure du possible, ce qui est spécifique à une langue se situe dans les ressources, et le traitement reste le plus générique possible (il y a actuellement quelques exceptions, surtout pour l'allemand qui permet d'utiliser des marques de fin de propositions qui lui sont propres).

4 Stratégie générale du processus d'analyse

Une graphie de phrase est d'abord soumise à un "processus standard" qui, pour environ la moitié des phrases, suffit pour segmenter correctement en propositions et pour relier sujets et verbes. Le résultat de ce processus est ensuite évalué, pour détecter la nécessité éventuelle d'un "post-traitement" de segmentation plus fine et/ou de mise en relation d'un couple sujet - verbe séparés par une proposition enchâssée.

4.1 Le processus standard

On part d'un document segmenté en phrases par le pré-traitement décrit en section 2. Des balises de début de proposition (conjonctions de subordination et pronoms relatifs), de début de chunks (prépositions et déterminants), d'auxiliaires conjugués (avoir, être et modaux) et de pronoms sujets, sont posées dans la graphie de la phrase au moyen d'expressions régulières (en prenant soin des frontières des mots et locutions : espace, ponctuation). Ces mots et locutions constituent, avec les terminaisons verbales et les clitiques (Déjean, 1998), **l'intégralité des ressources lexicales** spécifiques à une langue (environ **200 mots ou locutions par langue**).

Puis la phrase est découpée sur les débuts de proposition, d'où des segments nommés "*proto-propositions*", qui constituent un grain jouant un rôle d'intermédiaire de calcul [7].

---

[5] Le concept de "chunk" a été fondé par (Abney, 1991) sur les formes orales **et** écrites des langues; voir ensuite les thèses issues du Groupe Syntaxe : (Déjean, 1998), (Giguet, 1998), (Vannier, 1999), et (Lebarbé, 2002), ainsi que (Vergne, 2000) et (Vergne, 2001) page 26. Voir aussi le "bunsetsu" traditionnel japonais.

[6] Un aspect de la minimalité de la solution consiste aussi à ne pas tokeniser jusqu'au grain mot.

[7] Cette "proto-proposition" est ainsi nommée car elle n'est validée comme proposition qu'à la fin du traitement. Elle est analogue au "segment" de Thomas Lebarbé (Lebarbé, 2002).

Propriété exploitée : **chaque** proposition a **un** verbe et **un** sujet, qui sont le plus souvent connexes (et donc placés dans la même proto-proposition), mais quelquefois non connexes, car séparés par une proposition enchâssée (dans ce cas, sujet et verbe sont dans deux proto-propositions différentes, ce qui nécessitera un post-traitement de mise en relation).

Ces proto-propositions sont étudiées une par une, et un traitement particulier est déclenché selon le nombre d'auxiliaires et le nombre de pronoms sujets (personnel ou relatif) [8] :

	aucun auxiliaire	**1 auxiliaire**	**> 1 auxiliaires**
aucun pronom sujet	**chercher** sujet et verbe accordés [3]	**chercher** un sujet accordé [2]	**couper** en 2 proto-propositions [4]
1 pronom sujet	**chercher** un verbe accordé [1]	pas de traitement [0]	**couper** en 2 proto-propositions [4]
> 1 pronoms sujets	**couper** en 2 proto-propositions [4]	**couper** en 2 proto-propositions [4]	**couper** en 2 proto-propositions [4]

Figure 1 : Traitements déclenchés selon les attributs de la proto-proposition au cours du processus standard

Voici des exemples des différents cas de traitement déclenché selon les attributs de la proto-proposition, chacune subissant un seul des 5 traitements (les [numéros] réfèrent au tableau ci-dessus) :

[0] 1 pronom sujet et 1 auxiliaire => on a un verbe accordé avec un pronom sujet [9] :

*<[>|<pp>**Il**◇|<V>**a**◇fourni◇ **à**◇l'Europe ◇<d>**une**◇protection</d>◇très◇forte◇*
* **contre**◇une◇multitude ◇ **d'**événements ◇*

(sujet et verbe reliés sont marqués par le signe |)

[8] Environ la moitié des proto-propositions contiennent un auxiliaire, et environ un quart contiennent un pronom sujet. La fréquence élevée de ces marques en fait une très bonne base de départ des calculs.

[9] Pour faciliter la compréhension des résultats du calcul, la proto-proposition (grain et espace du calcul du processus standard) est encadrée, les marques servant aux calculs sont en gras (dont certains espaces), le reste du texte est en grisé, et les espaces sont marqués par le signe : ◇ .
Conventions des balises : *<[>* : début de proto-proposition marqué par le début de phrase,
*<pp>**il**◇* : **p**ronom **p**ersonnel sujet,
*<V>**a**◇fourni* : début de chunk **V**erbal conjugué,
* **à**◇l'Europe ◇* : chunk **p**répositionnel (partie masquée entre les balises),
*<d>**une**◇protection</d>◇très◇forte◇* : chunk nominal commençant par un **d**éterminant (partie masquée),
*<[cs>**parce**◇**qu'**</cs>◇* : début de proto-proposition marqué par une **c**onjonction de **s**ubordination ou un pronom relatif non sujet.

• on vérifie qu'ils font partie du même chunk verbal (=> aucun traitement), sinon on a 2 chunks verbaux et on est ramené au cas [4]

[1] 1 pronom sujet et pas d'auxiliaire => recherche d'un verbe accordé pour ce pronom sujet

*<[>Jetzt◇|<V>üb**t**◇|<pp>**er**◇sich◇ **im**◇Verschenken ◇*

• pronom et verbe font partie du même chunk verbal et le verbe est contigu au pronom

[2] pas de pronom sujet et 1 auxiliaire => recherche d'un sujet accordé à cet auxiliaire

*<[>|<d>**One**◇factor</d>◇restraining◇previous◇military◇action◇|<V>**was**◇<d>**an**◇ emphasis</d>◇ **on**◇zero ◇casualties◇*

• on cherche un groupe sujet commençant par un déterminant ou constitué d'un nom propre; en général, on le trouve avant, sauf en allemand, ou en français dans certaines subordonnées

[3] ni pronom sujet ni auxiliaire => recherche d'un sujet et d'un verbe conjugué accordés

• sujet et verbe accordés sont recherchés par essais de motifs comprenant un couple déterminant **et** terminaison verbale : accord au singulier (par exemple *L' -end* en français, *die -e*, *das -t* en allemand, *La -ó* en espagnol), au pluriel (*ces -ent* en français, *die -en* en allemand), ou de nombre indéterminé (par exemple en anglais : *the -ed*); la recherche de verbe à l'aide de sa terminaison s'accompagne d'un masquage temporaire des chunks nominaux et prépositionnels pour éviter d'y chercher un verbe; ce masquage est complet en allemand grâce à la majuscule des noms, et partiel dans les 4 autres langues :

*<[>|<d>**L'**euro</d>◇|<V>r**end**◇déjà◇ **d'**éminents ◇services◇*

*<[> **Dans**◇les◇deux ◇cas◇|<d>**ces**◇systèmes</d>◇ **d'**armes ◇ |<V>dispos**ent**◇ **de**◇radars *

*<[>|<d>**La**◇Bolsa</d>◇ **de**◇Tokio ◇|<V>cerr**ó**◇ayer◇ **a**◇su◇nivel ◇más ◇ bajo◇ **en**◇17 ◇años◇*

*<[>|<d>**Questo**◇tema</d>◇|<V>rischi**a**◇ **di**◇essere ◇<d>**la**◇questione</d>◇ sociale◇ **del**◇futuro *

• en anglais, on exploite la propriété (au présent) de la coupure entre sujet et verbe contigus, marquée *-non s / -s* au singulier ou *-s / -non s* au pluriel :

*<[cc>**But**◇|<d>**the**◇Pentagon</d>◇mov**e**◇|<V>represent**s**◇<d>**the**◇first</d>◇significa nt◇ federal◇call-up◇*

*<[>|<d>**The**◇cost**s**</d>◇|<V>moun**t**◇rapidly,◇*

• en allemand, la frontière chunk nominal - chunk verbal est marquée par le motif : mot à initiale majuscule / mot en minuscule ; voici un premier exemple : l'ordre sujet-verbe-complément qui fonctionne comme en français :

<[>|<d>**Das**◇*Sternbild*</d>◇*nämlich*◇|<V>*steh***t**◇ **in**◇*dieser*◇*Jahreszeit* ◇
besonders◇*tief*◇ **am**◇*Himmel* ◇

• et un deuxième exemple : l'ordre complément-verbe-sujet très courant dans la principale en allemand :

<[> **Bis**◇*Ende*◇*Oktober* ◇|<V>*schließ***t**◇*sich*◇|<d>**der**◇*Reigen*</d>◇
in◇*Connecticut* ,◇*Massachusetts*◇<cc>**und**◇*Rhode*◇*Island*◇

• mais on trouve aussi l'ordre verbe-sujet dans certaines subordonnées en français [10] :

<[cs>**dont**◇</cs>|<V>**ne**◇*dispos***ent**◇*pas*◇|<d>**les**◇*moyens*</d>◇*civils*◇
actuellement◇ **en**◇*service* ◇

[4] plus d'un pronom sujet ou plus d'un auxiliaire => segmentation ultérieure

• dans le post-traitement, il faudra couper cette proto-proposition en 2, en recherchant une coupure ;

Le traitement de la phrase est terminé si chaque proto-proposition a son couple sujet - verbe reliés, ce qui en fait une proposition, et si au moins une proposition principale a été détectée. C'est le cas pour environ la moitié des phrases. Si ces conditions ne sont pas réunies, le post-traitement est déclenché.

4.2 Le post-traitement éventuel

Le post-traitement a deux fonctions non exclusives :

1. segmenter en 2 les proto-propositions où 2 verbes ont été détectés, ou qu'il faut couper pour isoler la proposition principale,
2. relier sujet et verbe situés dans 2 proto-propositions différentes séparées par une subordonnée enchâssée, par le processus "ping-pong" : "ping" du candidat sujet, "pong" du candidat verbe - voir en 5.2 de (Vergne, 2001) page 26.

Remarquons qu'on retrouve les 2 opérations fondamentales de l'analyse : segmenter ou bien réunir (dans ce cas, réunir des segments non connexes).

Processus : chaque proto-proposition est de nouveau étudiée dans l'ordre d'apparition dans la phrase :

1. *segmentation* : si la proto-proposition a été marquée au cours du processus standard comme devant être découpée (cas [4] du processus standard), un point de coupure est recherché, et, s'il est trouvé, la proto-proposition est remplacée par 2 proto-propositions qui subissent alors de nouveau chacune le processus standard, et éventuellement de nouveau le post-traitement ; ces segmentations n'ont pas été faites au début du processus

[10] Voir à ce sujet une étude sur corpus dans (Vergne, 1998).

standard, car le nouveau point de coupure est contraint par l'étude de la proto-proposition dans le processus standard et la limitation de l'espace de recherche ;

2. *mise en relation* : si la proto-proposition est non résolue (elle n'a pas son couple sujet - verbe) :

   - soit elle contient un sujet possible, il est alors mis en attente ("ping" du sujet);

   - soit il y a un candidat sujet en attente, et la proto-proposition a un verbe possible accordé, alors ces deux proto-propositions sont reliées car elles forment ensemble une proposition non connexe autour d'une ou plusieurs propositions enchâssées ("pong" du verbe).

4.2.1 Exemples de segmentation d'une proto-proposition

• Coupure sur le *début d'une proposition coordonnée*, marqué par le motif coordination-déterminant (processus standard puis post-traitement) [11] :

*0 : <[><d>**Toutes**◇ces</d>◇personnes◇<V>**ont**◇reçu◇<d>**un**◇traitement</d>◇ <cc>**et**◇<d>**aucune**◇</d><V>**n'a**◇développé◇<d>**la**◇maladie</d> [nbV=2]*
1 : <[.>.

*0 : <[>|<d>**Toutes**◇ces</d>◇personnes◇|<V>**ont**◇reçu◇<d>**un**◇traitement</d>◇ [nbV=1 saturS=1]*
*1 : <[cc>**et**◇|<d>**aucune**◇</d>|<V>**n'a**◇développé◇<d>**la**◇maladie</d> [nbV=1 saturS=1]*
2 : <[.>.

• Coupure sur le *début de la proposition citative*, marqué par le motif virgule-déterminant (processus standard puis post-traitement) :

*0 : <[><d>**Several**◇thousand</d>◇reservists◇ **with**◇"specialized ◇skills"◇ <V>**could**◇be◇called◇up◇ **in**◇the◇next ◇few◇days**,**◇<d>**the**◇official</d>◇ <V>**said** [nbV=2]*
1 : <[.>.

*0 : <[>|<d>**Several**◇thousand</d>◇reservists◇ **with**◇"specialized ◇skills"◇ |<V>**could**◇be◇called◇up◇ **in**◇the◇next ◇few◇days**,** [nbV=1 saturS=1]*

[11] Convention des balises et des attributs : *[nbV=2]* : 2 verbes détectés (=> segmentation du post-traitement), *[nbV=1 saturS=1]* : 1 verbe détecté, saturation de la valence sujet = vraie, *<[cc>**et**◇* : début de proto-proposition avec une **c**onjonction de **c**oordination.

*1 : <[>|<d>**the**◇official</d>◇|<V>**said*** [nbV=1 saturS=1]*
2 : <[.>.

• Coupure double : sur le *début de la proposition principale*, marqué par le motif virgule-déterminant, et sur le *début de la citative inversée*, marqué par le motif guillemet-verbe (processus standard puis post-traitement) :

*0 : <[>"◇ **Avec**◇des ◇frappes◇ [nbV=0]*
*1 : <[cs>**comme**◇</cs>celles-ci**,**◇<d>**les**◇taliban</d>◇<V>**sont**◇rassurés◇**"**◇<V>**a**◇ironisé◇<d>**le**◇commandant</d>◇Djan◇Akhamat,◇numéro◇<d>deux◇</d> **d es** ◇forces ◇militaires◇ **du**◇Nord ◇ **dans**◇la◇plaine ◇ **de**◇Shomali [nbV=2]*
2 : <[.>.

*0 : <[>"◇ **Avec**◇des ◇frappes◇ [nbV=0]*
*1 : <[cs>**comme**◇</cs>celles-ci**,**◇ [nbV=0 saturS=0]*
*2 : <[>|<d>**les**◇taliban</d>◇|<V>**sont**◇rassurés◇**"**◇ [nbV=1 saturS=1]*
*3 : <[>|<V>**a**◇ironisé◇|<d>**le**◇commandant</d>◇Djan◇Akhamat,◇numéro◇<d>deux◇ </d> **des**◇forces ◇militaires◇ **du**◇Nord ◇ **dans**◇la◇plaine ◇ **de**◇Shomali [nbV=1 saturS=1]*
4 : <[.>.

4.2.2 Exemple de mise en relation de 2 proto-propositions

• Le processus standard a diagnostiqué qu'il n'y a pas de verbe dans la proto-proposition 0, ni dans la proto-proposition 2, autour de la (proto-)proposition relative 1 enchâssée (NB : **die** est un **p**ronom **r**elatif sujet - balisé <[pr> - car il ne précède pas de nom, marqué par une majuscule en allemand) :

*0 : <[><d>**Eine**◇junge◇Südafrikanerin</d>,◇ [nbV=**0**]*
*1 : <[pr>**die**◇ </pr>1969◇<d>**ein**◇neues◇Herz</d>◇|<V>erhiel**t**,◇ [nbV=1 saturS=1]*
*2 : <[>überlebte◇damit◇zwölf◇Jahre [nbV=**0**]*

3 : <[.>.

• Le post-traitement provoque ensuite la mise en relation en mettant la proto-proposition 0 en attente comme candidat sujet singulier, puis en le reliant à la proto-proposition 3 qui comporte un verbe singulier sans sujet (on se sert de l'accord au singulier du couple *Eine -e*) :

0 *: <[>|<d>**Eine**◇junge◇Südafrikanerin</d>,◇* *[S_en_attente=**0** **lienS=2** nbV=0]* [12]
*1 : <[pr>**die**◇ </pr>1969◇<d>**ein**◇neues◇Herz</d>◇|<V>erhiel**t**,◇* *[nbV=1 saturS=1]*
2 *: <[>|<V>überlebt**e**◇damit◇zwölf◇Jahre* ***[lienS=0** nbV=1 saturS=1]*
3 : <[.>.

4.2.3 Segmentation d'une proto-proposition avec mise en relation de 2 proto-propositions

• Coupure sur le *début du chunk verbal de la principale* (motif : verbe), avec mise en relation sujet-verbe autour d'une relative enchâssée (processus standard puis post-traitement) :

*0 : <[><d>**La**◇protection</d>* *[nbV=**0**]*
*1 : <[cs>**que**◇</cs><pp>**nous**◇voulons◇assurer◇<V>**est**◇<d>**une**◇protection</d>◇* * **d'**ensemble * *[nbV=**2**]*
2 : <[.>.

0 *: <[>|<d>**La**◇protection</d>◇* *[S_en_attente=**0** **lienS=2** nbV=0]*
*1 : <[cs>**que**◇</cs><pp>**nous**◇voulons◇assurer◇* *[nbV=1 saturS=1]*
2 *: <[>|<V>**est**◇<d>**une**◇protection</d>◇ **d'**ensemble * ***[lienS=0** nbV=1 saturS=1]*
3 : <[.>.

5 Conclusion

Nous avons présenté une méthode pour l'analyse descendante et calculatoire de corpus multilingues, appliquée au calcul des relations sujet - verbe dans les propositions. Les algorithmes utilisés consistent en des traitements répétitifs sur le grain minimal représenté (ici la proto-proposition); ces traitements n'utilisent que les deux opérations de base des

[12] Conventions des attributs : *[S_en_attente=**0** **lienS=2** nbV=0]* : cette proto-proposition est un sujet qui n'est plus en attente et qui a été relié à son verbe situé en proto-proposition 2.

expressions régulières : la recherche de motif et le remplacement de motif dans le grain minimal représenté. Cette méthode comporte la caractéristique très originale de n'utiliser que **très peu de ressources** lexicales et morphologiques (seulement 200 mots, locutions et morphèmes par langue), ce qui facilite l'ajout d'une nouvelle langue. Cette nouvelle voie est intéressante et prometteuse : ces moyens très légers ont déjà abouti à des résultats de qualité régulière et de bon niveau. Bien sûr, il reste encore un important travail d'évaluation comparative. Une limite actuelle est le fait que l'analyseur cherche systématiquement une relation sujet-verbe, d'où des erreurs dans certaines phrases nominales. Nous prévoyons d'élargir nos travaux à des langues que nous ne connaissons pas, en dissociant le chercheur-expérimentateur de l'évaluateur-locuteur de la langue étudiée. Les résultats déjà acquis nous confirment l'intérêt d'ajuster au plus serré les moyens à la tâche, ce qui permet d'éviter d'utiliser des outils trop généraux ou trop volumineux, qui exploitent des propriétés et des données linguistiques disproportionnées à la tâche. Ainsi, nous replaçons l'étude linguistique préliminaire en première position, étude qui doit faire émerger les *quelques* propriétés nécessaires à *cette* tâche. Plus ces propriétés sont générales, profondes et indépendantes des langues, plus la solution est robuste, rapide et facile à mettre en œuvre.

Références

Abney S. (1991), Parsing by Chunks, In : Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*, Dordrecht, Kluwer Academic Publishers.
(`http://www.sfs.nphil.uni-tuebingen.de/~abney/Abney_90e.ps.gz`)

Déjean H. (1998), *Concepts et algorithmes pour la découverte des structures formelles des langues*, Thèse de doctorat de l'Université de Caen.
(`http://www.info.unicaen.fr/~dejean/these/`)

Ejerhed E. (1996), Finite state segmentation of discourse into clauses, Actes de *ECAI'96 Workshop Extended finite state models of language*, A. Kornai (Ed.), .24-33.
(`http://www.kornai.com/ECAI/ejerhed.html`)

Giguet E. (1998), *Méthode pour l'analyse automatique de structures formelle sur documents multilingues*, Thèse de doctorat de l'Université de Caen.
(`http://www.info.unicaen.fr/~giguet/these/`)

Lebarbé T. (2002), *Hiérarchie Inclusive des Unités Linguistiques en Analyse Syntaxique Coopérative ; Un "segment" entre chunk et phrase dans le traitement linguistique par système multi-agents*, Thèse de doctorat de l'Université de Caen (soutenance prévue en mai 2002).

Lucas N., Giguet E., Vergne J. (2001), Détection automatique de la citation et du discours rapporté dans les textes informatifs, Actes de *Le discours rapporté dans tous ses états : Question de frontières*, Bruxelles, à paraître.

Rosmorduc, S. (1994). *Analyse morpho-syntaxique de textes non ponctués, application aux textes hiéroglyphiques*, Thèse de doctorat de l'École normale supérieure de Cachan.

Vannier G. (1999), *Étude des contributions des structures textuelles et syntaxiques pour la prosodie : application à un système de synthèse vocale à partir du texte*, Thèse de doctorat de l'Université de Caen.

(http://www.info.unicaen.fr/~vannier/Pages/Travail/these.htm)

Vergne J. (1998), Entre arbre de dépendance et ordre linéaire, les deux processus de transformation : linéarisation, puis reconstruction de l'arbre, *Les Cahiers de Grammaire*, n° 23, pp. 95-136.
(http://www.info.unicaen.fr/~jvergne/Cahiers_de_Grammaire_Vergne.pdf)

Vergne J. (2000), *Trends in Robust Parsing*, tutoriel du CoLing 2000, Nancy, Sarrebrück.
(http://www.info.unicaen.fr/~jvergne/tutorialColing2000.html)

Vergne J. (2001), Analyse syntaxique automatique de langues : du combinatoire au calculatoire, Actes de *TALN 2001*, 15-29.
(http://www.info.unicaen.fr/~jvergne/TALN2001_JV.ppt.zip)