Validation des calculs de relations de dépendance une expérience sur le “corpus Internet”

Thomas LEBARBÉ

Laboratoire LIDILEM EA609
Université Stendhal - Grenoble 3

Journée d’étude de l’ATALA
Le Web comme ressource pour le TAL
11 mars 2006

Plan de l'exposé

Préambule

- Cadre de l'analyse syntaxique
  - générale, corpus tout-venant
  - français / anglais principalement
- Ressources légères, minimales (méthode 'Vergne')
  - mots grammaticaux
  - marques morphologiques
  - pas de lexique exhaustif
- Déduction contextuelles
  - système à base de règles
- Principe de hiérarchie inclusive :
  - token (unités grpahiques plutôt que linguistiques)
  - chunk (token+) [Abney92]
  - segment (chunk+) [Lebarbé02]
  - phrase (segment+)

Définition et propriétés du segment

Définition :

Il existe un **segment** *de phrase composé de mots, chunks et/ou propositions dont les extrémités sont délimitées par des bornes.*

Par **borne**, *nous entendons les éléments qui marquent une rupture dans la linéarité de la phrase et dont la fonction est d'entretenir la logique et la structure phrastique : les conjonctions (de coordination et de subordination) et les pronoms relatifs que nous considérons comme bornes de début de segment; et les ponctuations que nous considérons comme bornes de fin.*

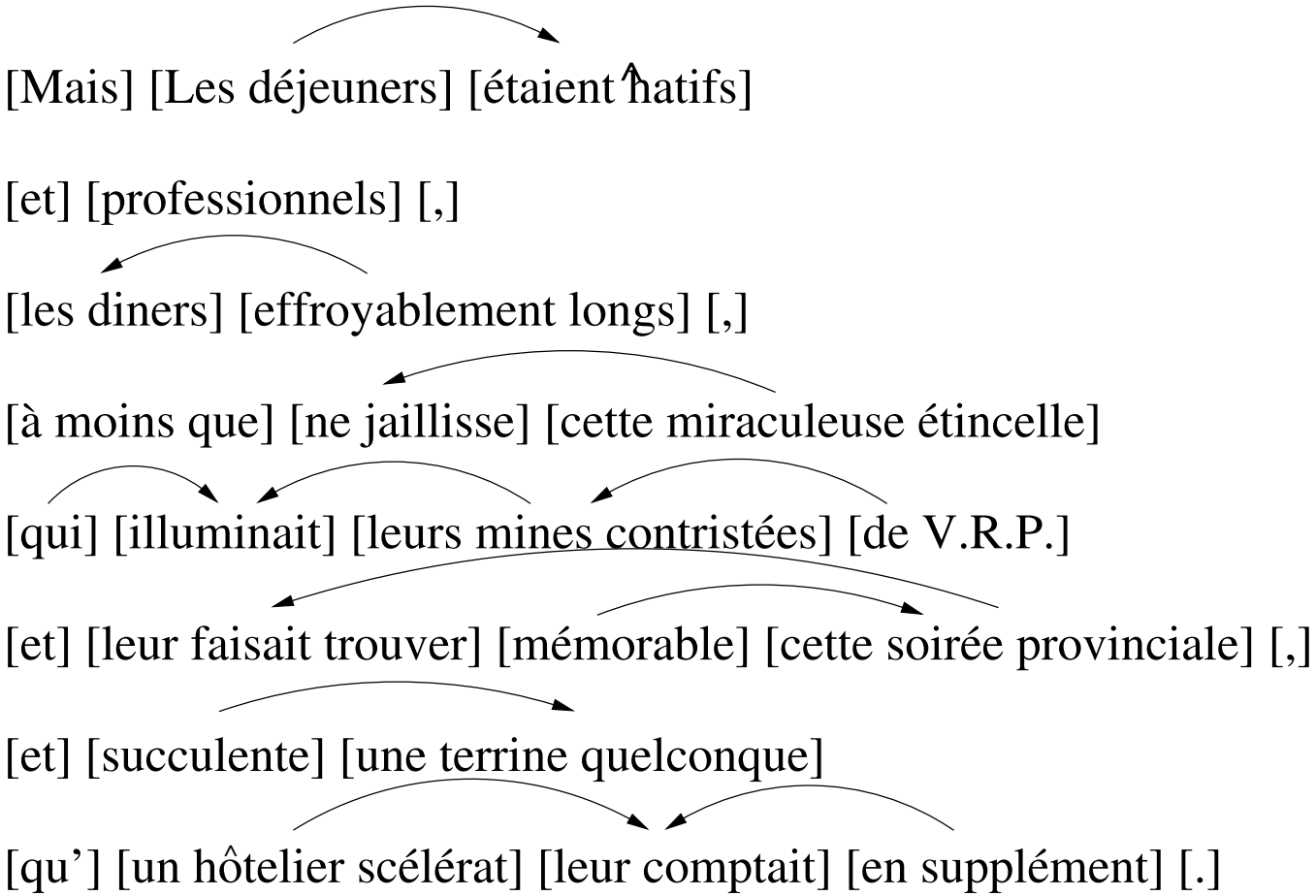
Propriété intrinsèque :

Les chunks internes au segment sont reliés par des relations de contiguïté quand le segment ne contient pas de verbe conjugué, sinon, les chunks sont organisés en deux branches unaires autour du chunk verbal.

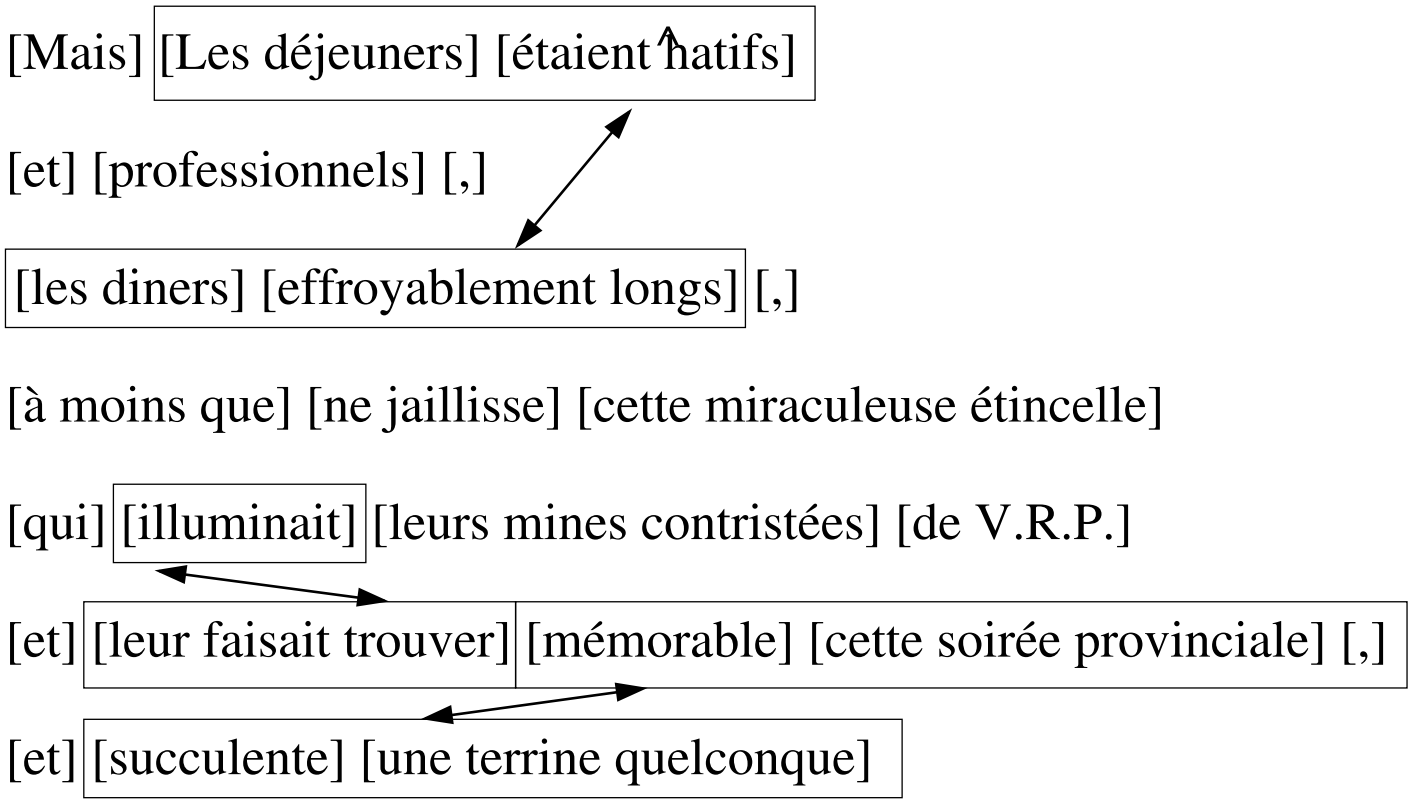
Propriété extrinsèque :

On peut observer des similarité structurelles entre segments qui jouent un rôle de grain d'observation et d'analyse de la phrase.

Exemple de propriété intrinsèque du segment



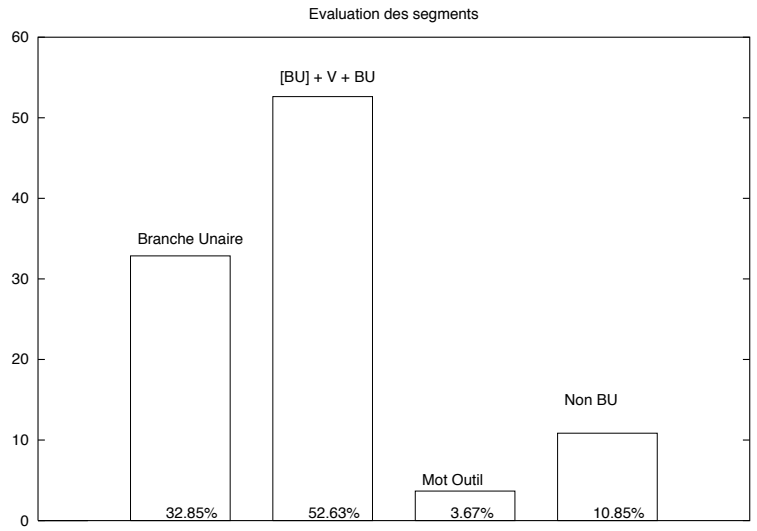
Exemple de propriété extrinsèque du segment



Limites du modèle du segment

Corpus : Le Monde, Vocable

	BU	[BU] + V + BU	Mot outil	Non BU	GLOBAL
Nb occ.	448	717	50	148	1363
Proportions	32,85%	52,63%	3,67%	10,85%	100%
Taille moy.	1,78	3,62	1,00	5,22	3,09
Taille max	7	9	1	10	10
Taille min	1	1	1	2	1
Écart type	0,95	1,34	0	1,68	1,71



Près de 11% ne correspondent pas à la règle par défaut

Hypothèse de coprésence :

Hypothèse :

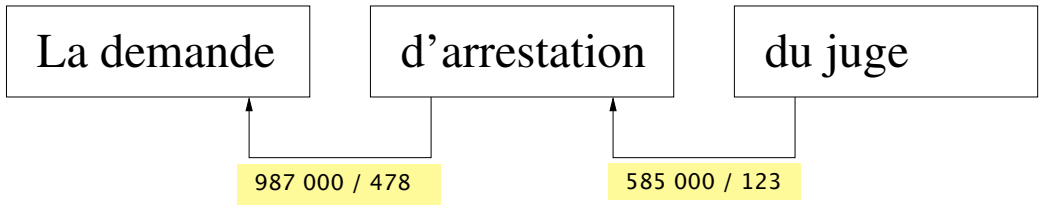
Un chunk prépositionnel A est plus probablement rattaché syntaxiquement à un autre chunk nominal ou prépositionnel B qui le précéde plutôt qu'à un autre C si la paire A–B est plus fortement coprésente que la paire A–C dans un corpus de grande taille.

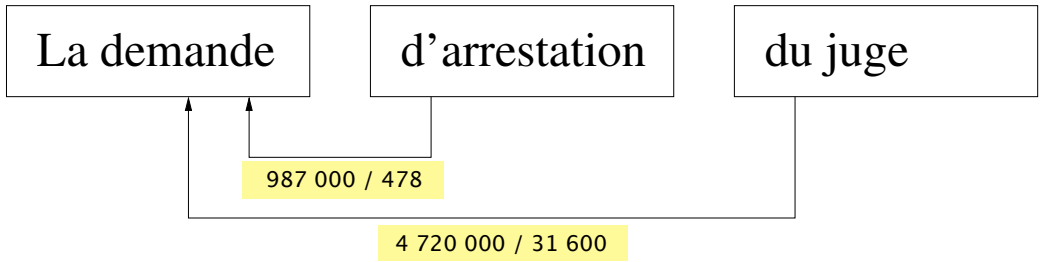
Variante :

Un chunk prépositionnel A est plus probablement rattaché syntaxiquement à un autre chunk nominal ou prépositionnel B qui le précéde plutôt qu'à un autre C si la **chaîne exacte** *AB est plus fortement coprésente que la* **chaîne exacte** *AC dans un corpus de grande taille.*

Exemple :

- extrait du corpus “Le Monde”
- requêtes Google : nombre de pages
- coprésence / chaînes strictes





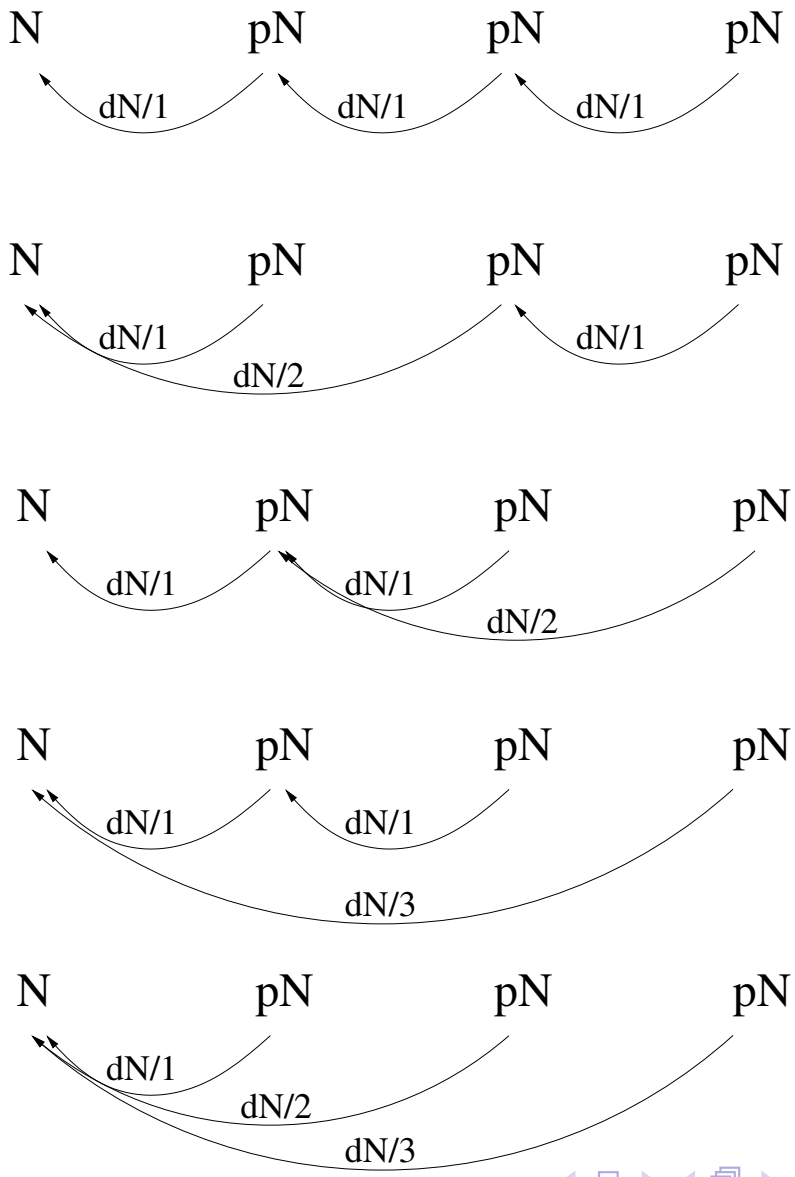
Chaînes strictes vs. Co-présence

- **Chaînes strictes**
  - ↗ Correspond exactement au texte traité
  - ↘ Problème de surspécification (“Juge” ou “Juge Garzon”)
  - ↘ Petit nombre d’occurrences → représentativité statistique
- **Co-présence**
  - ↗ Peu de résultats nuls
    (*“Aucun document ne correspond aux termes de recherche spécifiés”*)
  - ↘ Position dans le document / page
  - ↘ Pas d’ordre d’apparition
    (“la demande du juge” ou “le juge demande”)

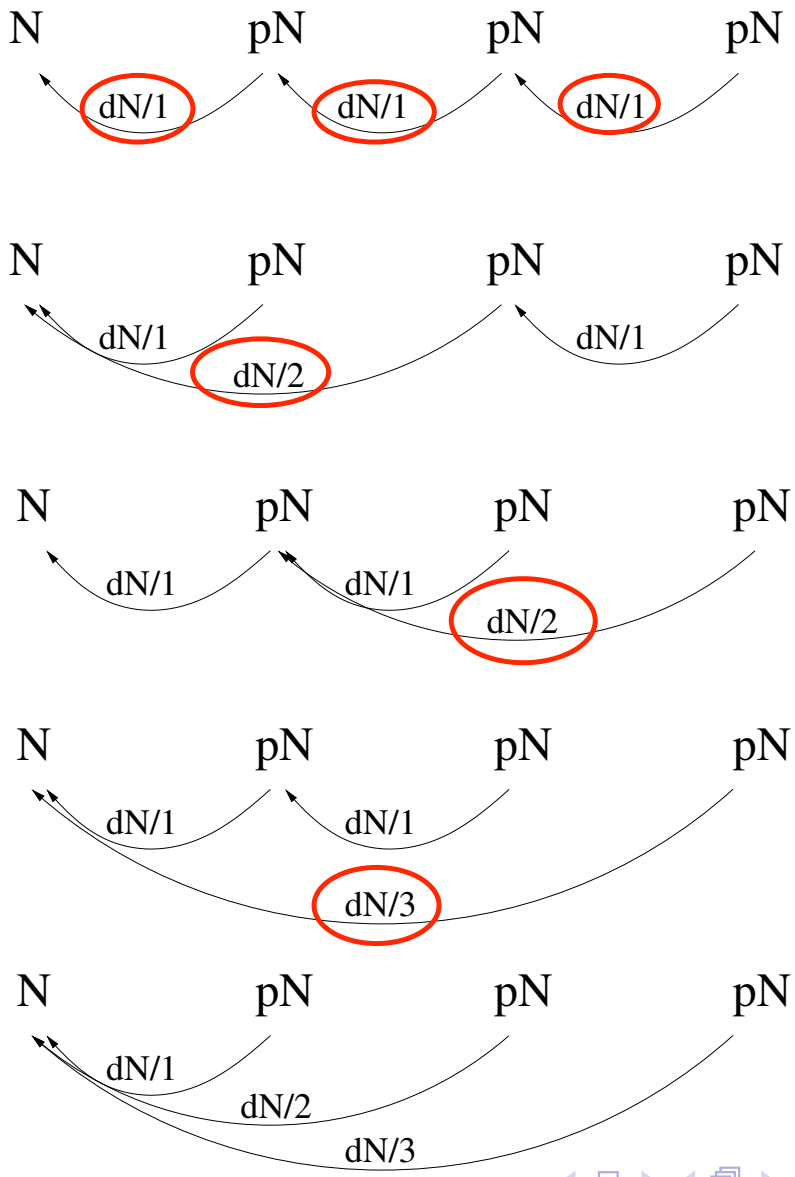
Combinatoire et pondération

1. Combinatoire
   - $\forall$ N pN pN^{+}, calcul de la combinatoire des arcs des arborescences possibles
   - N'est pas constituée de l'ensemble des appariements possibles entre chaque paire de chunks
   - Contrainte par le fait que deux relations de dépendance peuvent se chevaucher mais jamais se croiser [Hays64].
2. Pondération
   - Plus la distance entre deux unités linguistiques est longue, moins il est probable que ces deux unités dépendent l'une de l'autre
     principe de coût intellectuel aussi bien pour l'émeteur (le locuteur, l'auteur de l'écrit)
     que pour le récepteur (l'auditeur, le lecteur).
   - Intégré en pondérant par division les valeurs retournées par le moteur de recherche

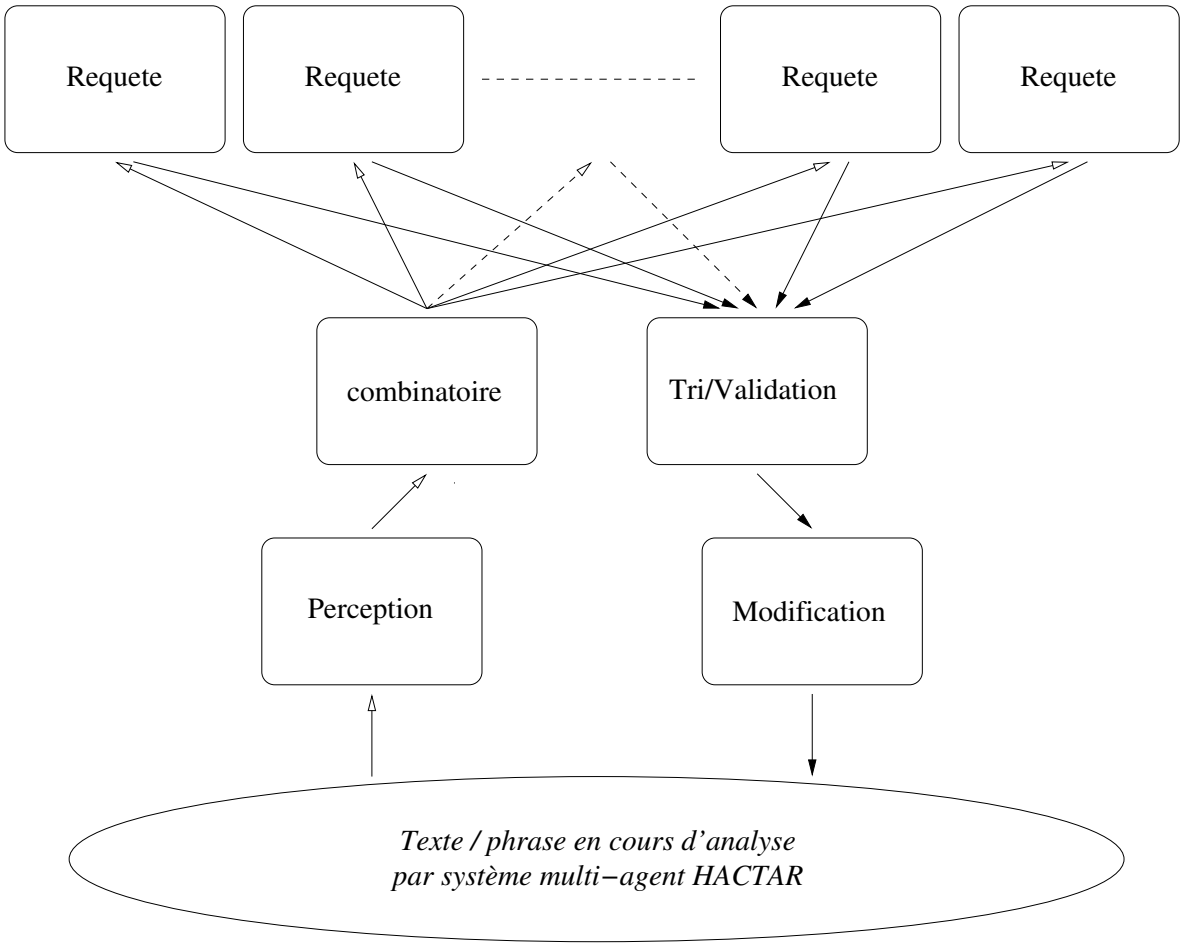
Combinatoire et pondération



Combinatoire et pondération : arcs



Implantation logicielle

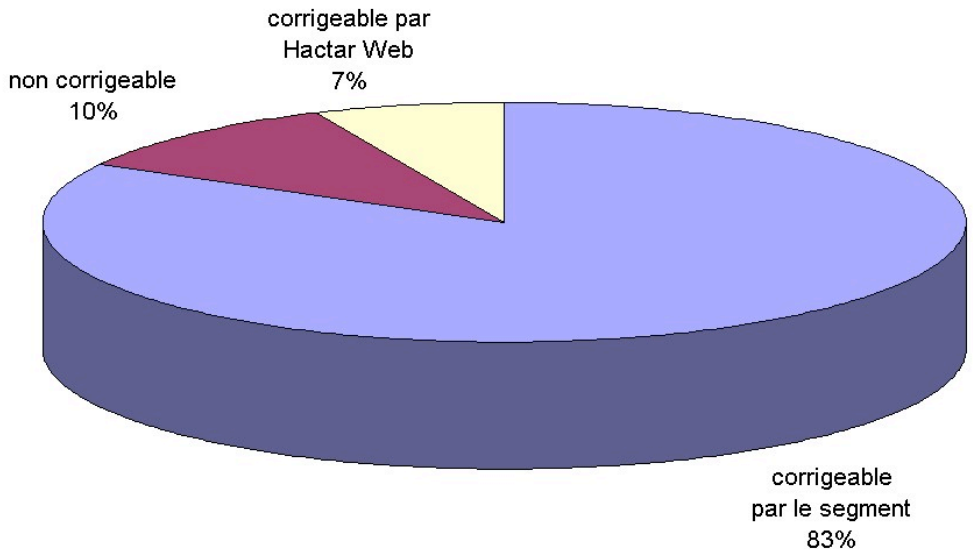


Corpus Journalistique (FR)

- Améliorations apportées à l'analyse de base (système GREYC'98)
- sur 341 erreurs observées

Corrections apportées par le segment	280	82,11%
Corrections apportées par les requêtess	24	7,04%

- soit aussi 39,34% des erreurs non corrigeables par le segment.

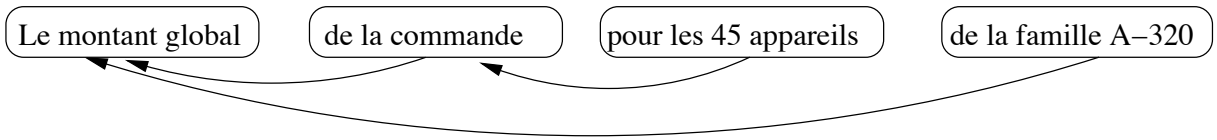


Corpus Newsfeed (FR)

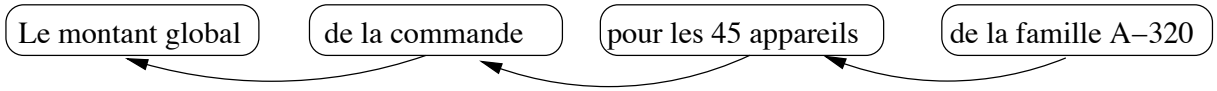
- Exemple :

##
La compagnie privée Spanair et le consortium européen Airbus ont signé un préaccord pour l'achat ferme de 21 Airbus de la famille A-320 et de 24 appareils en option.
==========
un préaccord pour l'achat ferme de 21 Airbus de la famille A-320
---
un préaccord + pour l'achat ferme
pour l'achat ferme + de 21 Airbus
de 21 Airbus + de la famille A-320
---
valider O, refuser O, ne pas tenir compte O
==========

- De l'importance du coût "cognitif"
- sans :



- avec :

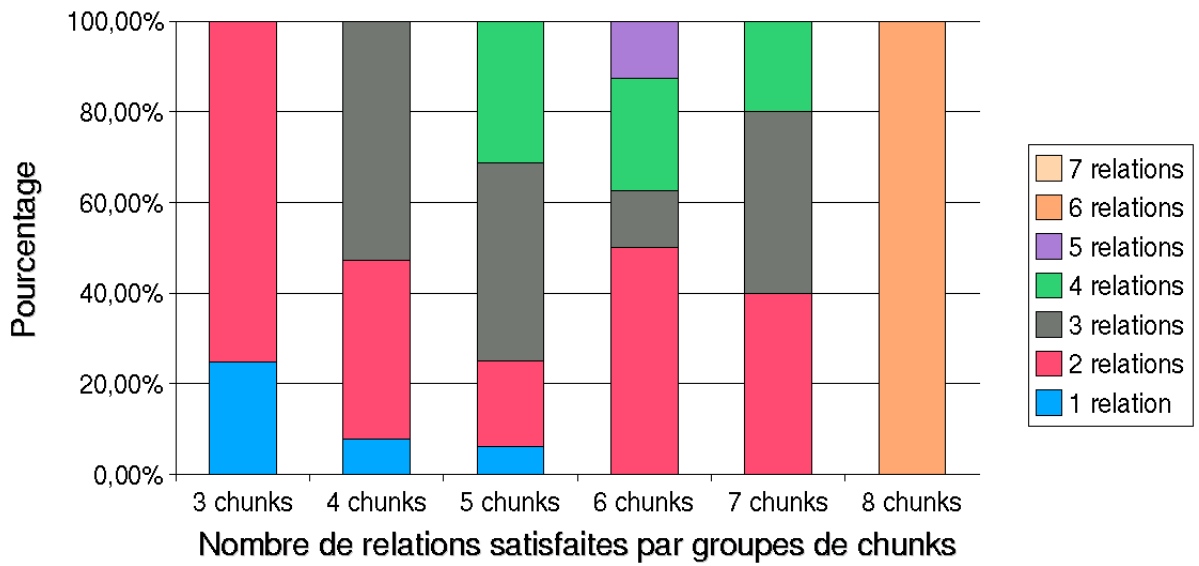


Corpus Newsfeed (FR) : Évaluation

- 300 séquences de chunks

Relations calculées corretement (↓) par séquences de i (→) chunks

sur	3	4	5	6	7	8	Total
1	25%	7%	6%	0%	0%	0%	55
2	75%	39%	18%	50%	40%	0%	185
3	-	54%	43%	12.5%	40%	0%	50
4	-	-	33%	25%	20%	0%	8
5	-	-	-	12.5%	0%	0%	1
6	-	-	-	-	0%	100%	1
7	-	-	-	-	-	0%	0
occurrences	194	76	16	8	5	1	300



Conclusions

- ↗ Méthode permet un certain nombre d'améliorations des résultats de mise en relation des chunks
- ↗ Evite la constitution/acquisition de ressources volumineuses et coûteuses

- ↘ Présume une représentatitivé de l'Internet
- ↘ Inadéquat pour certaines formes d'écrit
- ↘ Peu fiable en temps de calcul (de 50 à 1500ms par séquence $\texttt{N pN pN}^{+}$)
- ↘ Fiabilité des valeurs retournées par les moteurs (?)

Questions ouvertes

- Les valeurs retournées par les moteurs sont-elles fiables ?
- S'agit-il d'une ressource sémantique / pragmatique ?
- Utilité pratique / utilité théorique ?
- Pertinence de l'Internet comme ressource ?

"Internet se joue des obstacles éditoriaux! Les délires les plus dingues étant aussi les plus vifs, le réseau fait depuis sa naissance la part belle aux plus énormes d'entre eux, auxquels des sites sont consacrés, mais qui se propagent également par courrier électronique ou dans les groupes de discussion. C'est ainsi qu'on peut apprendre que Bill Gates, le patron de Microsoft, est l'Antéchrist, qu'Elvis est bien vivant, et que George W. Bush s'emploie à dissimuler les liens du gouvernement américain avec les extraterrestes..."

(G. Dasquié, J. Guisnel, L'effroyable mensonge, thèses et foutaises sur les attentats du 11 septembre, Editions La Découverte, 2002)

question time. . .