Journée Lig-Lidilem

« Que faire face à un corpus (écrit) ? »

Olivier Kraif

Université Grenoble Alpes - 28 mai 2019

Plan de la présentation

> Quelles questions soulève la question ? <

> Quels rapports au corpus ? <

> Corpus et linguistique <

> Corpus-based ou corpus-driven ? <

> Qualitatif ou quantitatif ? <

> Conclusion <

Première série de questions

Que faire face à un corpus (écrit) ?

- par quel bout le prendre ?
- comment le manipuler ?

-> atelier pratiques

Deuxième série de questions

Que faire **face à** un corpus (écrit) ?

- le corpus est-il *donné* ?
- de quel manière *s'impose*-t-il à nous ?
- à quel type d'*évidence* nous invite-t-il ?

Deuxième série de questions

Deuxième série de questions

« (...) un corpus résulte d'un groupement raisonné, conduit par une hypothèse de recherche explicite. Il est donc impossible de définir ce qu'est un "bon" corpus en faisant abstraction de l'objectif visé. »

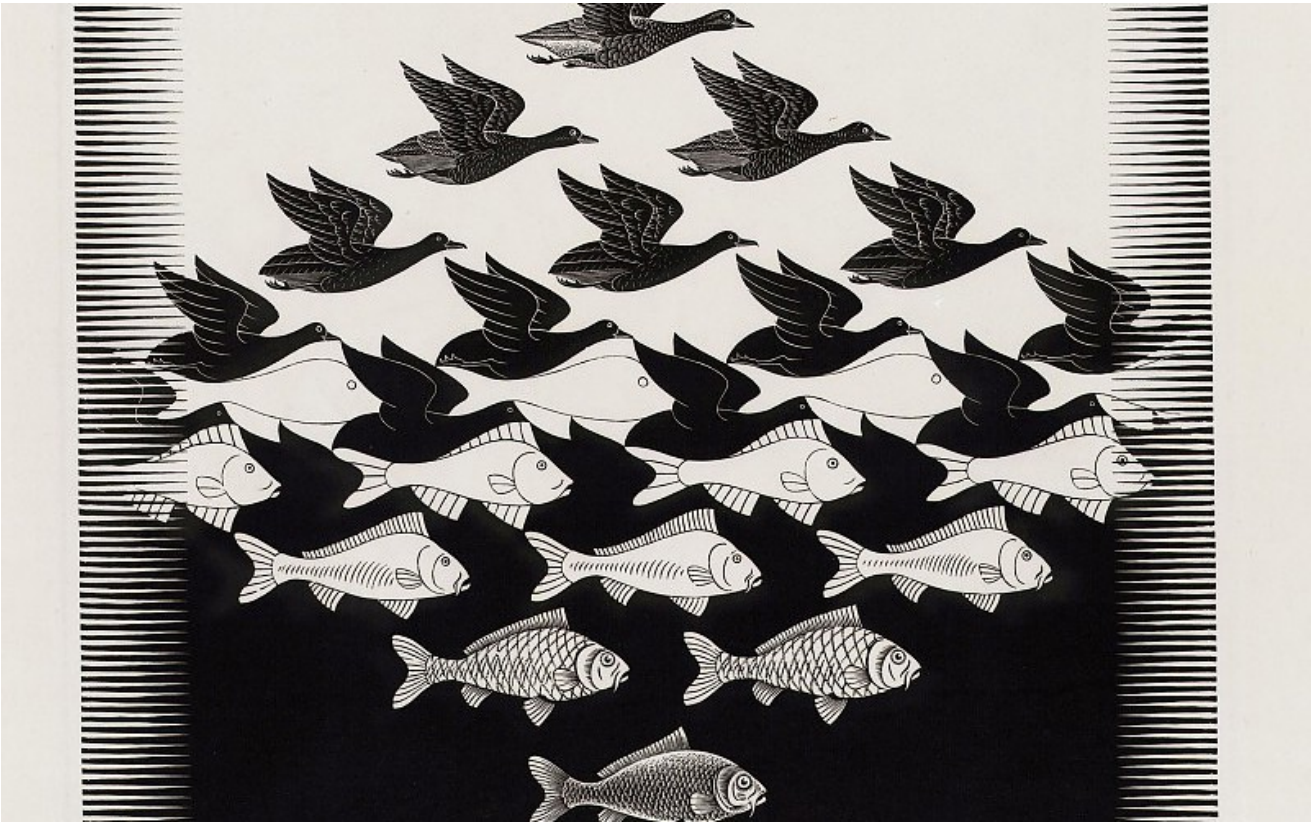
(Habert et al., 1998:35)

Troisième série de questions

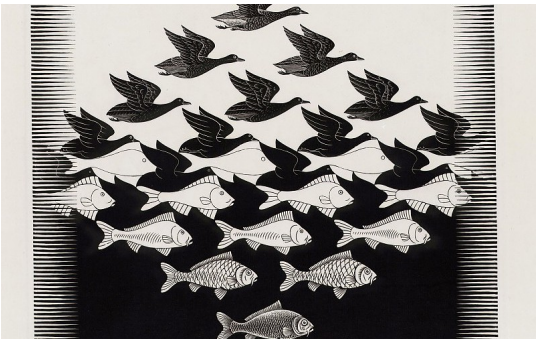
Que faire face à un corpus (**écrit**) ?

- pose la question du *texte*
  - en tant que *discours*
  - en tant qu'unité autonome
- corpus de textes (ou fragment de textes) ?
  corpus de discours (ou fragments de discours) ?

Quel(s) rapport(s) au corpus ?



Deux types de rapports



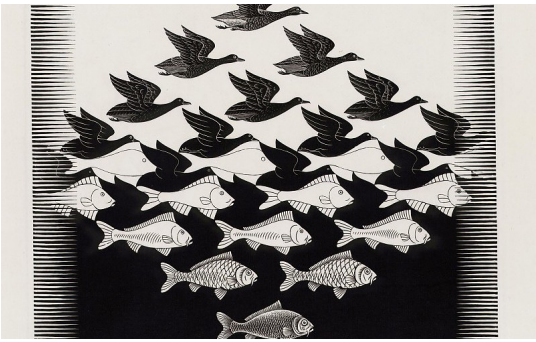
- Corpus *in se*

Le corpus pris en tant que que lui-même

- Corpus-échantillon

Le corpus pris en tant qu'il représente un ensemble plus large

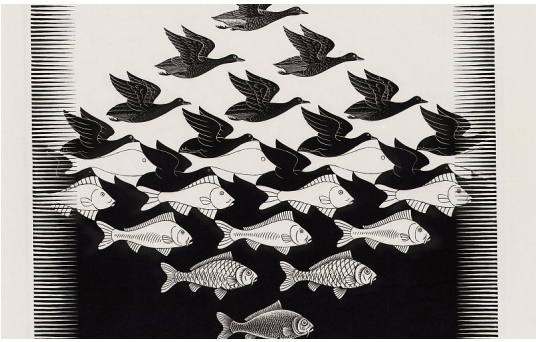
Deux types de rapports



- Corpus *in se*
  - une oeuvre en particulier
  - l'oeuvre écrit d'un auteur
    - Index Thomisticus : Roberto Busa (1913-2011)

→ perspective des études littéraires, de l'herméneutique, de la philologie, de l'analyse de contenu

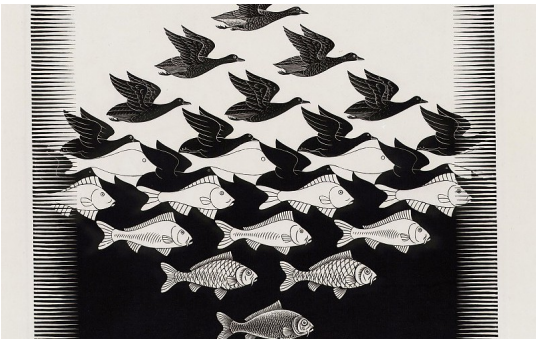
Deux types de rapports



- Corpus-échantillon
  - corpus de référence : BNC, COCA…
  - corpus spécialisés : Scientext, Kiap, CoMeRe, etc.

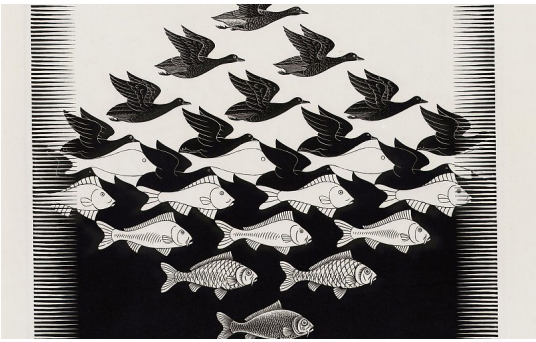
→ perspective de la linguistique de corpus, de la lexicographie, de la terminologie, ...

Deux types de rapports



"A corpus is a collection of pieces of language text in electronic form, **selected** according to external criteria **to represent**, as far as possible, **a language or language variety** as a source of data for linguistic research." (John Sinclair, 2004)

Deux types de rapports

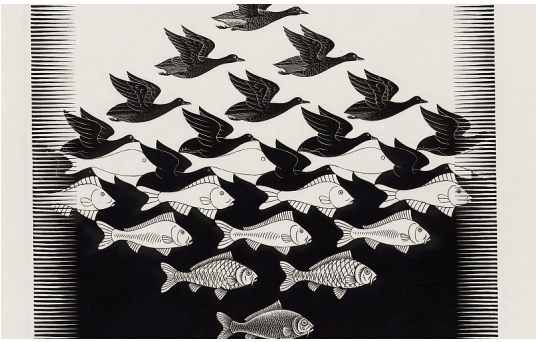


> Ce rapport n’est pas inscrit dans le corpus lui-même

- Corpus PhraseoRom
  - étude d’une œuvre *vs* étude d’un sous-genre
- Poème de Gérard Macé - Chenilles et papillons
  - étude d’une création poétique vs étude traductologique

    cf. http://ouvroir-litt-arts.univ-grenoble-alpes.fr/revues/actalittarts/333-chenilles-amp-papillons

Deux types de rapports



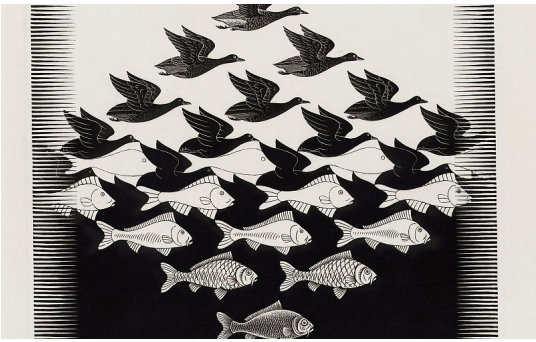
> Sur un plan épistémologique

- Démarche *nomothétique* : recherche de lois générales

vs

- Démarche *idiographique* : description de cas singuliers

Rapport du troisième type



> Corpus-outil

- TAL : pour l’entraînement et/ou l’évaluation
- Notion d’échantillonnage cruciale
- Corpus n’est plus objet d’étude
- Connaissances dérivées : implicites, opératoires

Corpus et linguistique

D’où vient le succès de la linguistique de corpus ?

- Deux explications
  - Big data : milliards de mots accessibles
  - Changement de paradigme

D'où vient le succès de la linguistique de corpus ?

- Avant l'ère du Big Data...
  - XVIIe, Samuel Johnson, Dictionary of English Language
  - 1828, American Dictionary of the English Language
  - 1928, Oxford English Dictionary

  → corpus de citations sous forme de fiches (fichiers, archives)
  - 1959, Randolph Quirk, Survey of English Usage Corpus (enregistré sur bande, transcrit sur papier, indexé par des fiches)

D’où vient le succès de la linguistique de corpus ?

- Avant l’ère du Big Data...
  - Sinclair, 1987 : projet COBUILD premier grand corpus de référence

D'où vient le succès de la linguistique de corpus ?

- Changement de paradigme

« (…) l'écart entre le **sentiment linguistique** des locuteurs, concernant les détails de la langue, et les **faits récoltés objectivement** à partir des textes **est énorme et systématique**. Il nous conduit à émettre l'hypothèse que l'intuition humaine à propos de la langue est spécifique aux individus, et qu'elle ne peut pas du tout constituer un bon **guide** pour décrire ce qui se passe réellement lorsque ces mêmes individus font usage de la langue (Sinclair 1991) »

D’où vient le succès de la linguistique de corpus ?

> Changement de paradigme

- le guide n’est plus dans notre tête (cf. Chomsky, jugement de grammaticalité, position cartésienne)
- le guide est face à nous, hors de nous, dans le corpus

D’où vient le succès de la linguistique de corpus ?

> Révolution copernicienne

- pour la première fois (depuis des siècles), on construit un dictionnaire sans s’appuyer sur ses illustres prédécesseurs : on fait table rase !
- seul le corpus doit guider l’ouvrage du lexicographe

D'où vient le succès de la linguistique de corpus ?

> Mais pourquoi se méfier de l'intuition ?

- parce que saillance (disponibilité) ≠ fréquence
- parce que usage réel largement inconscient
  - p.ex. : « concevoir » dans la presse

    avec le pronom en « il en a conçu le projet »

D’où vient le succès de la linguistique de corpus ?

A la longue, il n'est pas exclu que les autorités puissent	en	concevoir une certaine irritation.
Ils ont trois enfants, Hannah, Joseph et George, et elle	en	conçoit une fierté sans nuance.
" Dangereux pour mon avenir " Il n'	en	conçoit aucune amertume : " Je n'ai pas particulièrement besoin d'argent, donc cela ne m'affecte pas financièrement, mais je suis touché psychologiquement, parce que le monde de l'édition est ainsi fait aujourd'hui que si on ne vend pas, on n'est plus publié.
Il ne pense évidemment pas qu'il n'y aura plus d'événements, de soubresauts, de contestation de la démocratie; il affirme que la démocratie est devenue la norme même pour ceux qui s'y opposent, tels les islamistes qui la prennent pour référence, quelque fureur qu'ils	en	conçoivent.
Il	en	a conçu - " comme bien d'autres " - un remords, un malaise, lequel s'est traduit par une " fièvre narcissique morbide " , une ironie qu'il a dirigée en premier lieu contre sa famille, puis globalement contre tout, lui inclus.
A Lille toujours, dans le bureau populaire de Moulins, Marion avoue qu'elle était à Paris, il y a une semaine à la même heure, et	en	conçoit du remords : " Je me suis sentie coupable.
Mario Luni	en	conçoit aujourd'hui bien des regrets.
Darina Al-Joundi	en	a aussi conçu un livre, traduit dans six langues.
Il ne faut	en	concevoir ni jalousie ni fascination.
Elle n'	en	conçoit aucun complexe.
L'écrivain refusé partout	en	conçoit inévitablement un peu d'aigreur, légitime parfois, car de bons livres aussi sont ignorés.

D’où vient le succès de la linguistique de corpus ?

- 26/43 occurrences :

concevoir = éprouver un sentiment

(21 négatifs, 5 positifs ou neutre)

D'où vient le succès de la linguistique de corpus ?

- Double avantage du corpus
  - indication de la fréquence
  - dimension contextuelle : unités de sens étendues (extended units of meaning)

D'où vient le succès de la linguistique de corpus ?

> Exemple d'approche guidée par le corpus (corpus-driven)

- Hanks, 2004, Corpus Pattern Analysis.
  - description d'une entrée à partir d'une concordance (200-400 contextes)
  - classement des contextes en fonction des analogies lexico-syntaxiques
  - identification des principaux patterns (norms) et de leurs interprétations

D'où vient le succès de la linguistique de corpus ?

PDEV: announce

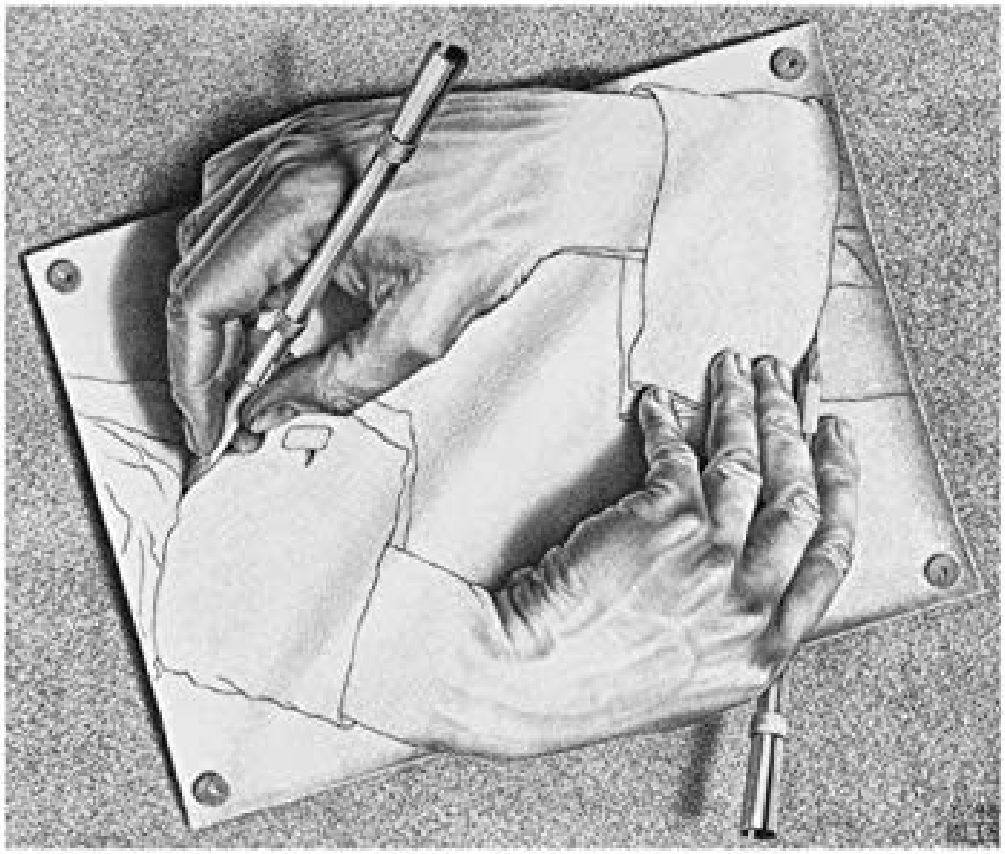
Access full data

Displayed here are **All patterns** .
Other options:

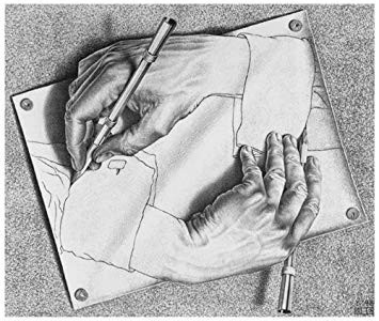
samplesize: **250**
patterns: **4**

1	*Pattern:* **Human** *or* **Institution** **announces** THAT-CLAUSE *Implicature:* FORMAL. Human *or* Institution makes a formal statement to a public audience that [CLAUSE] is or will be the case + *Example:* *the government suddenly* ***announced*** *that a bridge that is a big part of the project would be paid for entirely with public money*	**40.8%** **...More data** FrameNet
2	*Pattern:* **Human** *or* **Institution** **announces** **Event** *or* **Plan** *Implicature:* FORMAL. Human *or* Institution makes a formal statement that Event has happened or Plan will happen + *Example:* *The government* ***announced*** *further trade liberalization measures*	**49.6%** **...More data** FrameNet
3	*Pattern:* **Human** **announces** QUOTE *Implicature:* Human says [QUOTE] in the manner of a formal public statement *Example:* *a fellow train man* ***announced*** *: `This train is bound for London Bridge, fingers crossed.'*	3.2% **...More data** FrameNet
4	*Pattern:* **Human** *or* **Institution** **announces** **Entity** *Implicature:* Human *or* Institution makes a formal public statement that Entity will be made available *Example:* *Both firms have* ***announced*** *small computers and plan big sales campaigns.*	5.6% **...More data**

Corpus-driven
vs
Corpus-based



Êtes-vous plutôt Corpus-based ou Corpus-driven ?

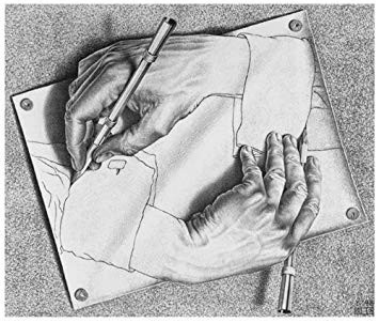


Corpus-based versus *corpus-driven* linguistics

« The distinction between corpus-based and corpus-driven language study was introduced by Tognini-Bonelli (2001). Corpus-based studies typically **use corpus data in order to explore a theory or hypothesis, aiming to validate it, refute it or refine it**. The definition of **corpus linguistics as a method** underpins this approach.

Corpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that **the corpus itself should be the sole source of our hypotheses about language**. It is thus claimed that **the corpus itself embodies a theory of language** (Tognini-Bonelli 2001: 84-5). » (http://corpora.lancs.ac.uk/clmtp/1-cb-cd.php)

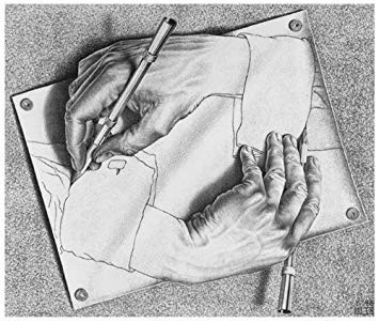
Êtes-vous plutôt Corpus-based ou Corpus-driven ?



- Approche **déductive** vs approche **inductive**
- Approche inductive pure :
  - empirisme **radical**

    Isaac Newton : « Hypotheses non fingo » les lois générales dérivent de l'expérience. Pas la peine d'émettre d'hypothèse explicative *a priori*.
  - position problématique : peut-on faire des observations sans partir d'hypothèses ?

Êtes-vous plutôt Corpus-based ou Corpus-driven ?



- Composition du BNC

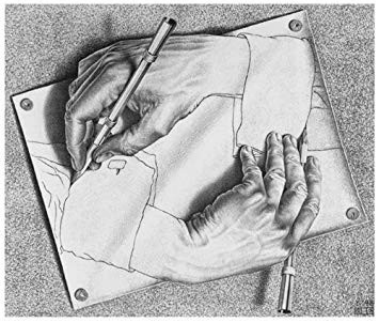
Table 1. Text type

	texts	w-units	%	s-units	%
Spoken demographic	153	4233955	4.30	610557	10.13
Spoken context-governed	755	6175896	6.27	427523	7.09
Written books and periodicals	2685	79238146	80.55	4395581	72.94
Written-to-be-spoken	35	1278618	1.29	104665	1.73
Written miscellaneous	421	7437168	7.56	487958	8.09

Table 3. Written Domain

	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

Êtes-vous plutôt Corpus-based ou Corpus-driven ?



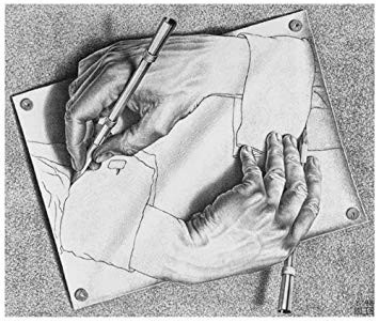
> Le corpus dérive d'une série de choix raisonnés

- équilibre oral / écrit
- équilibre entre les types de discours et genre de textes
- équilibre entre domaines, entre aires géographiques

> Ces choix n'empêchent pas une part d'arbitraire

- oral → 17 % de l'ensemble ?
- écrits de fiction → 27 % de l'écrit ?

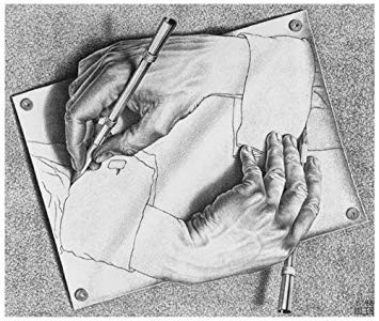
Êtes-vous plutôt Corpus-based ou Corpus-driven ?

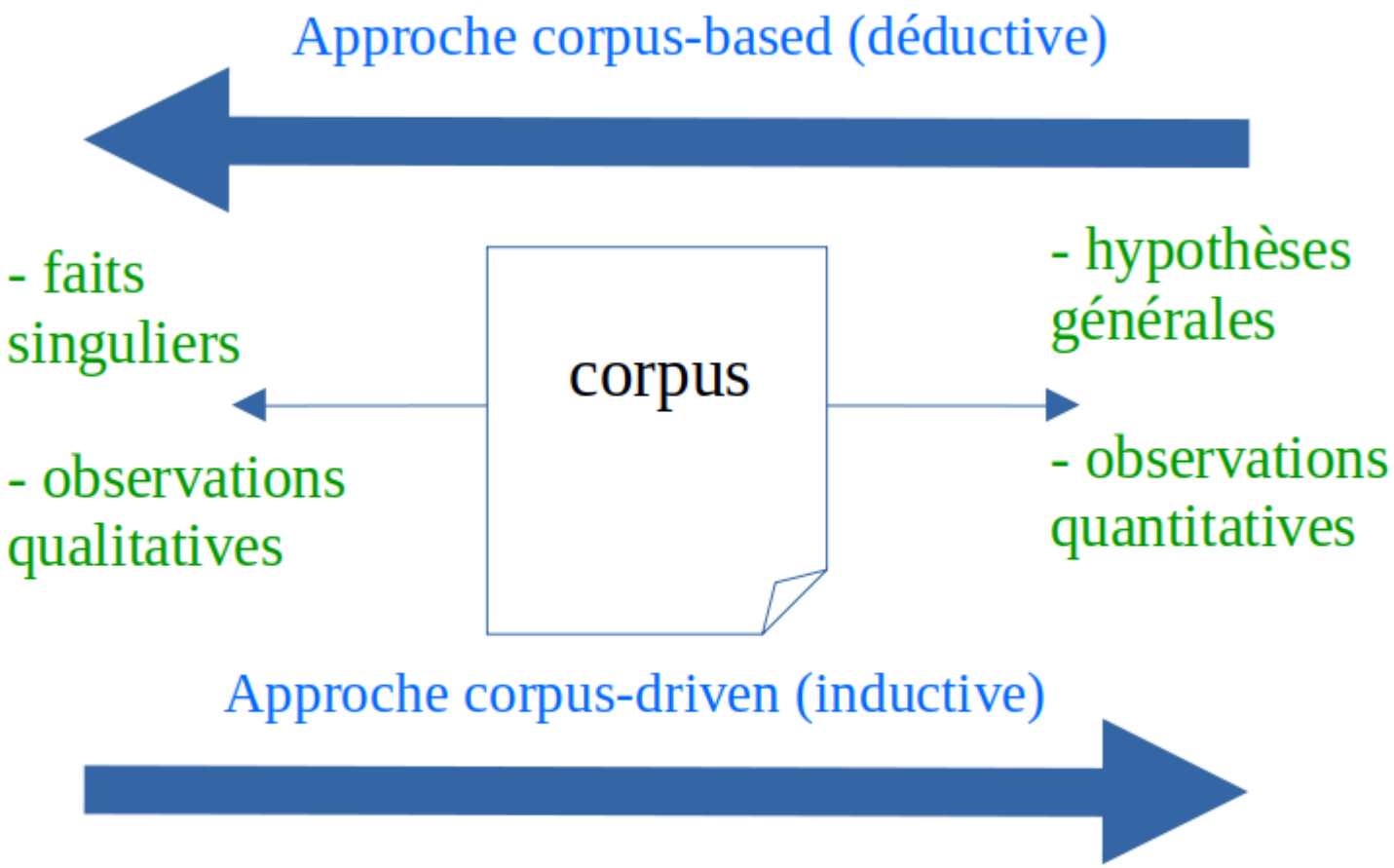


> Pourrait-on constituer un corpus de référence sans procéder à une telle reconstruction ? sans faire d'hypothèse sur ce qu'est le périmètre de la langue que l'on vise ?

> Autre problème : la science est cumulative par essence. Doit-on oublier toutes les observations générales (résultat du travail de ses prédécesseurs) à chaque fois qu'on revient au terrain ? Le scientifique doit-il se contraindre à l'amnésie ?

Êtes-vous plutôt Corpus-based ou Corpus-driven ?





Observations qualitatives
vs
Observations quantitatives

Qualitatif ou quantitatif ?

Différents types d'observation :

- fréquence des mots et expressions, accroissement du vocabulaire, nombre moyen de mots par phrases, de syllabes par mots, analyses factorielles, etc.

- concordances, exemples en contexte

Qualitatif ou quantitatif ?

Différents types d'observation :

- fréquence des mots et expressions, accroissement du vocabulaire, nombre moyen de mots par phrases, de syllabes par mots, analyses factorielles, etc.
  → quantitatives
- concordances, exemples en contexte
  → qualitatives

Qualitatif ou quantitatif ?

Et pourtant

- le comptage des mots est-il seulement une opération quantitative ?
- mais qu’est-ce qu’un mot ?
- qu’est-ce qu’une phrase ? une syllabe ?
- le vocabulaire doit-il se compter sur la base des formes ou des lemmes ?
- La lemmatisation est-elle une opération univoque ?

  → norme de dépouillement (Née, 2018)

Qualitatif ou quantitatif ?

Toutes les données quantitatives reposent, *in fine*, sur des objets linguistiques définis en fonction de critères qualitatifs

- Quand on élabore une référence pour le TAL (gold ou silver standard), on mesure l'accord intersubjectif
- Même pour des annotations simples
  - segmentation en mot
  - étiquetage / lemmatisation
  - alignement de phrases
  - entités nommées
  - etc.

Qualitatif ou quantitatif ?

Les métriques quantitatives reposent sur des catégories intersubjectives :

- précision
- rappel
- kappa
- etc.

Qualitatif ou quantitatif ?

Les observations qualitatives peuvent-elles se passer de validation quantitative ?

Quel est le statut de l'exemple (et des exempliers ?)

Qualitatif ou quantitatif ?

Le statut de l'exemple

- illustrer, montrer un phénomène ?

  → oui

- démontrer un phénomène ?

  → en LDC, non

- quid de l'attestation isolée ?

Prononc. : [ɛ ̃fɔ ʀmatik]. **Étymol. et Hist.** 1962 (Terme inventé par Ph. Dreyfus d'apr. Gilb. 1971); 1966, 16 nov. *(Le Monde, ibid.)*. Dér. de *informat(ion)*; suff. -ique**. **Bbg.** Bertini (M.-T.), Tallineau (Y.). Pt vocab. *L'Informat. nouv.* 1977, nº80, p. 23. - Dossiers de mots. *Néol. Marche*. 1977, nº3, p. 37. - Encyclop. (XII)... *L'Informat.* 1973, nº45, p. 70; (XIII)... 1973, nº46, pp. 61-64.

Qualitatif ou quantitatif ?

La LDC s'intéresse à des faits reproductibles

- très fréquents ou rares
- mais dont l'occurrence n'est pas imputable au hasard
- au-delà de la parole c'est toujours la langue qui est visée

Qualitatif ou quantitatif ?

Un exemple tiré de Scientext

1	les manuels comportent une dimension heuristique , encore peut -on	douter	s' ils sont effectivement donnés à faire aux élèves .	#1 - HDR - Sciences de l'éducation - Développement
2	perturbations dans les salles de permanence . Outre qu' on peut	douter	de l' efficacité d' une telle prescription , on voit	#6 - HDR - Linguistique - Développement
3	systèmes de suppléance réside , comme on peut s' en	douter	, dans la lenteur de la composition des messages .	#218 - Article - TAL - Introduction
4	nettement sémantique et visuographique , proche à n' en pas	douter	de ce que l' on peut considérer comme un cas	#423 - Article - Linguistique - Développement
5	ont permis de souligner à quel point le conflit faisait	douter	C' est par exemple le cas de Gyoungho et Kwon (	#591 - Thèse - Psychologie - Développement
6	se comprendre grâce à la capacité pour la source de faire	douter	Or , le conflit , d'après cette approche , peut	#591 - Thèse - Psychologie - Développement
7	doute de sa propre réponse , mais il va également	douter	de la réponse de l' autre , celle -ci n' ayant	#591 - Thèse - Psychologie - Développement
8	apparaît que c' est peut -être justement parce qu' il fait	douter	que le conflit prend un sens tout particulier dans une	#591 - Thèse - Psychologie - Développement
9	perception des compétences propres indiquent que les désaccords semblent faire	douter	les individus et , de fait , rendre la tâche plus	#591 - Thèse - Psychologie - Développement
10	par exemple Nicholls , 1984a) , ou amenés à en	douter	par un feed-back négatif (Elliott et Dweck , 1988	#591 - Thèse - Psychologie - Développement
11	la peur de l' échec montrent que les désaccords font	douter	l' individu de ses connaissances et de ses compétences .	#591 - Thèse - Psychologie - Développement
12	des raisons personnelles ou pour des raisons situationnelles , fait	douter	les individus de leurs connaissances , alors ce but risque	#591 - Thèse - Psychologie - Conclusion
13	est pas explicitement décrite et porte même le lecteur à	douter	En effet , dans la deuxième phrase de la deuxième partie	#593 - Thèse - Linguistique - Développement
14	type processus produit ont permis , à n' en pas	douter	, d' identifier un corpus important de variables pouvant interférer	#594 - Thèse - Sciences de l'éducation - Développement
15	des pratiques d' enseignement ont , à n' en pas	douter	, suscité un intérêt socialement reconnu ; " l' importance	#594 - Thèse - Sciences de l'éducation - Développement
16	) . Ces recherches ont , à n' en pas	douter	, apporter une intelligibilité sur le fonctionnement dynamique d' une	#594 - Thèse - Sciences de l'éducation - Développement
17	actions restantes . Celles -ci , à n' en pas	douter	, ne se réduisent nullement à des réflexes , stimulés	#594 - Thèse - Sciences de l'éducation - Développement
18	la reproduction , ont permis , à n' en pas	douter	, d' apporter une contribution explicative concernant l' inadaptation de	#594 - Thèse - Sciences de l'éducation - Développement
19	ce cadre théorique a permis , à n' en pas	douter	, de mieux comprendre et d' expliquer les phénomènes observés	#594 - Thèse - Sciences de l'éducation - Développement
20	en anglais ? Le logiciel est trop susceptible de faire	douter	un élève (trop de choix et de questions piège)	#595 - Thèse - Sciences de l'éducation - Développement
21	ça c' était un peu de ma faute aussi j' aurais dû me	douter	/ j' étais un peu coincé par le temps / là	#595 - Thèse - Sciences de l'éducation - Développement
22	particularités énumérées plus haut . Diverses études conduisent notamment à	douter	de la validité des principes de monosémie et de monoréférentialité	#599 - Thèse - Linguistique - Développement
23	Merci de m' avoir montré qu' il était important de	douter	constamment pour assurer un diagnostic aussi fiable que possible .	#603 - Thèse - Médecine - Remerciements
24	cuisses de cette fraischeur (...) On ne saurait	douter	que cette mode des arrosages clandestins ne délimite en propre ,	#608 - Thèse - Linguistique - Développement

Qualitatif ou quantitatif ?

Un exemple tiré de Scientext

- sur 27 occurrences de *douter*
- 7 occurrences de *à n'en pas douter*
- 6 en position d'incise

A partir de ces statistiques on peut effectuer des tests (significativité, corrélation, etc.) pour déterminer si l'hypothèse nulle (hasard) est à rejeter.

Qualitatif ou quantitatif ?

Un exemple tiré de Scientext

1	les manuels comportent une dimension heuristique , encore peut -on	douter	s' ils sont effectivement donnés à faire aux élèves .	#1 - HDR - Sciences de l'éducation - Développement
2	perturbations dans les salles de permanence . Outre qu' on peut	douter	de l' efficacité d' une telle prescription , on voit	#6 - HDR - Linguistique - Développement
3	systèmes de suppléance réside , comme on peut s' en	douter	, dans la lenteur de la composition des messages .	#218 - Article - TAL - Introduction
4	nettement sémantique et visuographique , proche à n' en pas	douter	de ce que l' on peut considérer comme un cas	#423 - Article - Linguistique - Développement
5	ont permis de souligner à quel point le conflit faisait	douter	C' est par exemple le cas de Gyoungho et Kwon (	#591 - Thèse - Psychologie - Développement
6	se comprendre grâce à la capacité pour la source de faire	douter	Or , le conflit , d'après cette approche , peut	#591 - Thèse - Psychologie - Développement
7	doute de sa propre réponse , mais il va également	douter	de la réponse de l' autre , celle -ci n' ayant	#591 - Thèse - Psychologie - Développement
8	apparaît que c' est peut -être justement parce qu' il fait	douter	que le conflit prend un sens tout particulier dans une	#591 - Thèse - Psychologie - Développement
9	perception des compétences propres indiquent que les désaccords semblent faire	douter	les individus et , de fait , rendre la tâche plus	#591 - Thèse - Psychologie - Développement
10	par exemple Nicholls , 1984a) , ou amenés à en	douter	par un feed-back négatif (Elliott et Dweck , 1988	#591 - Thèse - Psychologie - Développement
11	la peur de l' échec montrent que les désaccords font	douter	l' individu de ses connaissances et de ses compétences .	#591 - Thèse - Psychologie - Développement
12	des raisons personnelles ou pour des raisons situationnelles , fait	douter	les individus de leurs connaissances , alors ce but risque	#591 - Thèse - Psychologie - Conclusion
13	est pas explicitement décrite et porte même le lecteur à	douter	En effet , dans la deuxième phrase de la deuxième partie	#593 - Thèse - Linguistique - Développement
14	type processus produit ont permis , à n' en pas	douter	, d' identifier un corpus important de variables pouvant interférer	#594 - Thèse - Sciences de l'éducation - Développement
15	des pratiques d' enseignement ont , à n' en pas	douter	, suscité un intérêt socialement reconnu ; " l' importance	#594 - Thèse - Sciences de l'éducation - Développement
16	) . Ces recherches ont , à n' en pas	douter	, apporter une intelligibilité sur le fonctionnement dynamique d' une	#594 - Thèse - Sciences de l'éducation - Développement
17	actions restantes . Celles -ci , à n' en pas	douter	, ne se réduisent nullement à des réflexes , stimulés	#594 - Thèse - Sciences de l'éducation - Développement
18	la reproduction , ont permis , à n' en pas	douter	, d' apporter une contribution explicative concernant l' inadaptation de	#594 - Thèse - Sciences de l'éducation - Développement
19	ce cadre théorique a permis , à n' en pas	douter	, de mieux comprendre et d' expliquer les phénomènes observés	#594 - Thèse - Sciences de l'éducation - Développement
20	en anglais ? Le logiciel est trop susceptible de faire	douter	un élève (trop de choix et de questions piège)	#595 - Thèse - Sciences de l'éducation - Développement
21	ça c' était un peu de ma faute aussi j' aurais dû me	douter	/ j' étais un peu coincé par le temps / là	#595 - Thèse - Sciences de l'éducation - Développement
22	particularités énumérées plus haut . Diverses études conduisent notamment à	douter	de la validité des principes de monosémie et de monoréférentialité	#599 - Thèse - Linguistique - Développement
23	Merci de m' avoir montré qu' il était important de	douter	constamment pour assurer un diagnostic aussi fiable que possible .	#603 - Thèse - Médecine - Remerciements
24	cuisses de cette fraischeur (...) On ne saurait	douter	que cette mode des arrosages clandestins ne délimite en propre ,	#608 - Thèse - Linguistique - Développement

Qualitatif ou quantitatif ?

La possibilité de généraliser (induction)

- dépend d'aspects quantitatifs (statistiques)
- dépend du critère de représentativité (dispersion, échantillonnage, etc.)

La portée des observations qualitatives dépend donc bien de critères quantitatifs !

Conclusion

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abella , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
ltats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

Que faire face à un corpus textuels ?

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abeilla , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
llats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

- se laisser guider par l'accumulation des exemples analogues
- classer, catégoriser (cœur de l'activité linguistique)
- observer les corrélations, identifier des récurrences
- émettre des hypothèses et les tester
- revenir au texte : seul le co-texte et la connaissance du contexte permet l'interprétation (importance des métadonnées)

Et le TAL dans tout ça ?

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abeilla , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
ltats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

Pourvoyeur de nouveaux instruments

- Annotations
- Visualisations
- Outils de requête

→ nouvelles méthodologies à inventer

→ nécessité de contrôler les biais et artefacts

Et le TAL dans tout ça ?

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abeilla , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
ltats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

Grand consommateur de corpus

- entraînement des systèmes (statistiques, neuronaux, …)
- évaluation des systèmes (corpus de référence)

Pour le moment, le TAL profite surtout des connaissances implicites qu'il tire des corpus bruts.

Merci !

Références

Habert B., Fabre C., Issac F. (1998). *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, InterEditions.

Hanks, P. (2004). Corpus Pattern Analysis, in G. Williams & S. Vessier (eds) *Proceedings of the 11th Euralex International Congress*, Université de Bretagne Sud, Lorient, pp. 87-98.

Néé, E. (dir.)(2018) *Méthodes et outils pour l'analyse des discours*, PUR, coll. Didact Méthodes.

Sinclair, J. (1991). *Corpus, concordance, collocation*, Oxford University Press.

Sinclair, J. (2004). *Trust The Text , Language, Corpus and Discourse*, Routledge (Taylor and Francis).

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work,* Amsterdam and Philadelphia: John Benjamins.

Et le TAL dans tout ça ?

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abeilla , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
ltats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

Grand consommateur de corpus

- entraînement des systèmes (statistiques, neuronaux, …)
- évaluation des systèmes (corpus de référence)

Pour le moment, le TAL profite surtout des connaissances implicites qu'il tire des corpus bruts.

Et le TAL dans tout ça ?

analyse avant_de poser , dans une dernière partie , les	conclusions	de notre recherche . L' anglais dispose aujourd'hui de trois
s de_par la méthode utilisée mais également à_travers les	conclusions	qu' elle apporte . Ainsi , selon Hernandez (2004
Cela rejoint les	conclusions	de Hernandez (2004 : 194) sur la relation
Finalement , un ensemble de	conclusions	et perspectives clôturera l' article . Natural language generatior
Finally , a set of	conclusions	and pointers to further work end up the paper .
En	conclusion	, on observe que l' architecture , bien_qu' elle s'
résultats obtenus sur le corpus Clef 2004 -2005 et nos	conclusions	sur les voies d' amélioration possibles pour notre système .
, nous proposons une analyse de notre méthode et les	conclusions	que nous en avons tiré . Le projet Ritel du
Uruguay , Fernando Pérez Abeilla , a dit que les	conclusions	sont très encourageantes . Tableau 1 : Exemples extraits du
Cette	conclusion	concorde avec nos résultats précédents . En_termes_d' efficac
aquelle la compréhension d' un texte est assimilée aux	conclusions	qu' un lecteur humain peut tirer de sa lecture par
d' un argument qui justifie l' autre élément comme une	conclusion	. L' analyse des mots de liaison (connecteurs)
de ne présenter que les faits en faveur de notre	conclusion	(faits bons à dire) et de ne rien
évoque immédiatement l' idée de priorité , qui est une	conclusion	que l'on souhaite favoriser dans l' esprit du lecteur ,
ltats présentés ci-dessus permettent de tirer un certain nombre de	conclusions	. En ce qui concerne le rappel , un des

Grand consommateur de corpus

- entraînement des systèmes (statistiques, neuronaux, …)
- évaluation des systèmes (corpus de référence)

Pour le moment, le TAL profite surtout des connaissances implicites qu'il tire des corpus bruts.