Analyse automatique des citations dans un corpus d'articles de SHS

Le projet RHECITAS

Ludovic TANGUY

Objectifs et démarche

- Analyse des citations dans les publications de SHS en français
  - Enrichissement des publications en ligne
  - Etude des fonctions rhétorique et discursives
  - Observation des habitudes des différentes disciplines
- Réalisation d'une chaîne d'analyse automatique
  - Extraction des références bibliographiques et des appels de citation
  - Caractérisation de ces références sur la base d'une étude linguistique des contextes de citation
- Premiers travaux de ce type en français et en SHS

Financement et Partenaires

- **Financement TGE-Adonis (Très Grand Equipement pour l'Accès unifié aux DOnnées NumérIques des SHS) - CNRS**
- **CLLE-ERSS** (Cognition Langues Langage Ergonomie – Equipe de Recherche en Syntaxe et Sémantique – UMR 5263) : CNRS & Université de Toulouse
  - C. Fabre, LM. Ho-Dac, MP Péry-Woodley, J. Rebeyrolle, F. Sajous, F. Lalleman
- **INIST** (Institut de l'Information Scientifique et Technique – UPS 76) : CNRS
  - C. François, D. Besagni
- **IRIT** (Institut de Recherche en Informatique de Toulouse – UMR 5505) : CNRS & Université de Toulouse
  - F. Benamara, J. Mothe, P. Muller
- **Synapse Développement** (SARL,Toulouse)
  - P. Séguéla

PLAN

- Etat de l'art en analyse des citations
  - Différentes approches de la question
    - Bibliométrie
    - Classification des citations
    - Extraction de termes
  - TAL et Analyse des citations
- Aspects techniques
  - Ressources disponibles et contraintes
  - Corpus et machinerie
- Premier objectif : classification des citations
- Second objectif : extraction des mots-clés
- Bilan et autres considérations

ANALYSE DES CITATIONS

Différentes approches

- 1/ Bibliométrie
  - « *Qui* cite X ? » (« *Combien* d'articles citent X ? »)
  - Méthode : analyse quantitative des index de citations
  - Objectifs: évaluation de la recherche, accès aux publications, veille scientifique
- 2/ Classification des citations
  - « *Pourquoi* X cite-t-il Y? »
  - Méthodes : enquêtes, analyse linguistique des contextes de citation
  - Objectifs : bibliométrie qualitative
- 3/ Analyse de contenu
  - « *Que* retient X de Y quand il le cite »
  - Méthode : extraction de mots clés des contextes de citation
  - Objectifs: indexation, résumé automatique (des articles cités)

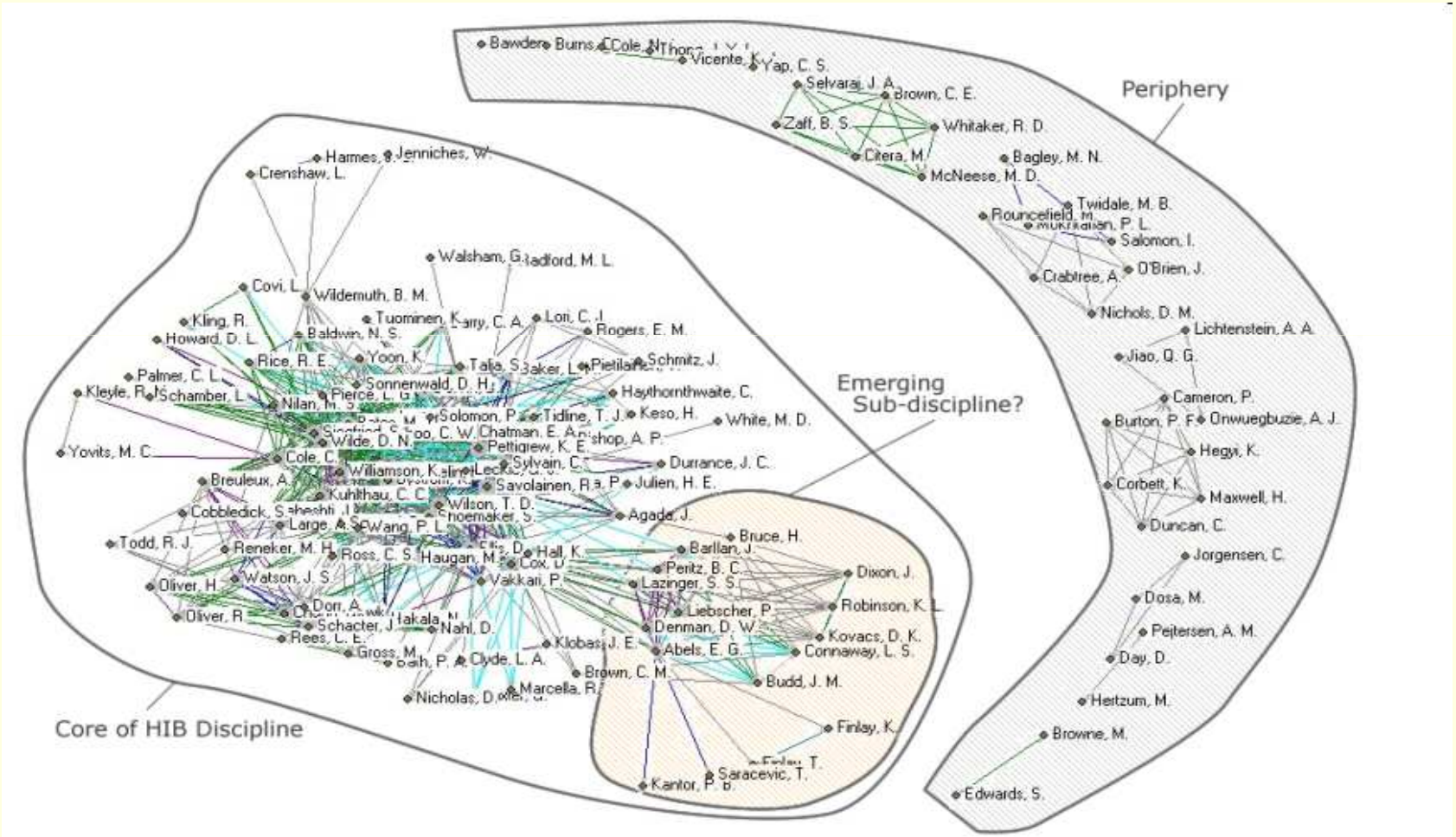
Première approche : Bibliométrie (scientométrie)

- Analyse quantitative des publications scientifiques
  - Sur les seules relations de citation entre publications (l'article X cite l'article Y)
  - Calcul de la notoriété d'un article, d'une revue d'un chercheur, via le nombre de citations dont il est la cible (facteur d'impact, h-index, etc.)
  - Etude des réseaux de co-citation
- Utilisations
  - Evaluation de la recherche (chercheur / laboratoire / revue)
  - Identification des « fronts de recherche » (analyse des co-citations)
- Techniques
  - Recueil de publications intégrales ou de notices
  - Analyse et normalisation de la bibliographie d'un article
  - Construction et analyse d'un graphe de citations (citation map)

Applications

- Pionnier : Eugene Garfield (dès 1952)
  - Fondateur de ISI (Institute for Scientific Information) en 1960
  - Fournisseur de services bibliométriques pour la communauté scientifique (index de citations)
  - Services actuels : Web of Science & Web of Knowledge
- Autres bases de données disponibles :
  - SCI (Thomson)
  - Scopus (Elsevier)
  - Google Scholar (Google)
  - CiteSeer (IST)

Exemple : réseau de co-citations entre auteurs (McKechnie 2005)



Enjeux

- Développement croissant des services et bases de données avec la disponibilité des supports en ligne
  - Revues électroniques, diffusion directe par les auteurs, archives spécialisées, dépôts centralisés
- Enjeux politiques et économiques de tels outils
  - Services généralement payants (abonnement des centres de ressources)
  - Implications des éditeurs (indexation de leurs propres revues)

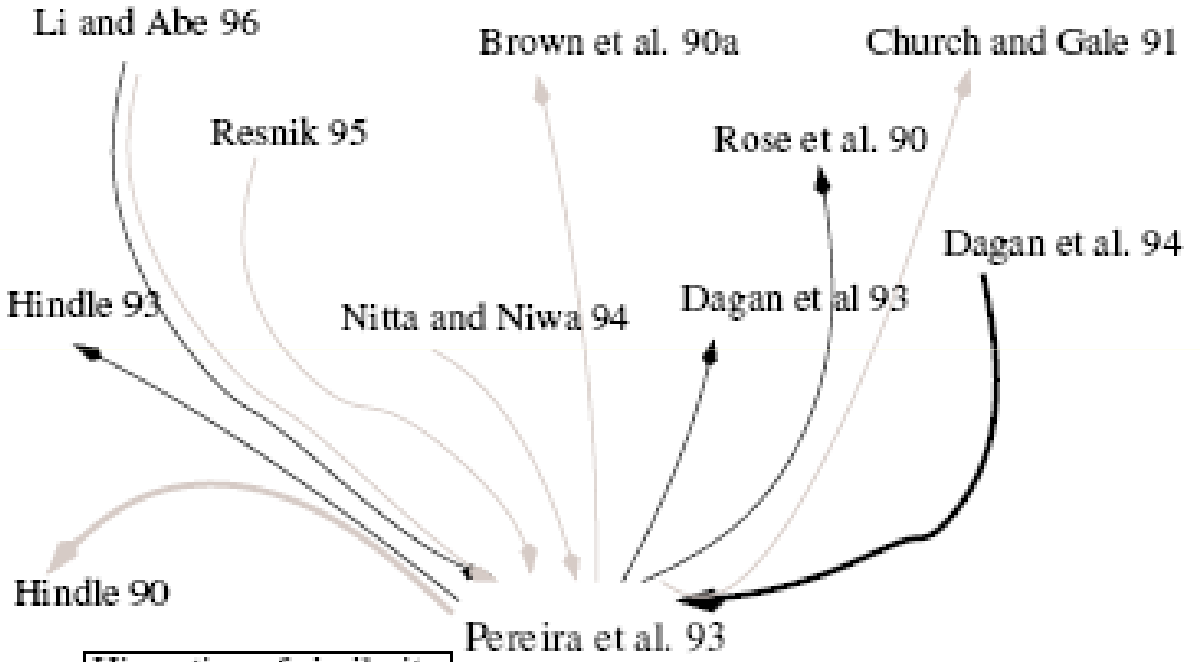
Critiques

- Couverture des bases de données
  - Couvertures très variables entre les bases de données (~40 % de recouvrement)
  - Variations importantes entre les disciplines (médecine > sciences dures > sciences humaines)
- Fiabilité
  - Les références sont extraites automatiquement
  - Silence, et difficultés à normaliser
    - Efforts en ce sens : DOI et Crossref
- Pas de prise en compte des « types » de citation
  - Une citation négative (critique) compte autant qu'une positive

Deuxième approche : Classification des citations

- Identification des motivations/fonctions des citations
  - Positif/négatif, important ou non, etc.
- Objectifs
  - Enrichissement des index de citation
  - Etudes sur les mécanismes de la citation
    - Linguistique, didactique, sociologie des sciences
- Travaux toujours au stade expérimental
  - Etudes en infodoc
  - Travaux en TAL

Carte de citations enrichie (Teufel et al 2006)



Fonctions rhétoriques des citations

- Typologie des fonctions : exemple initial (Garfield, 1962)

- Paying homage to pioneers.
- Giving credit for related work (homage to peers).
- Identifying methodology, equipment, etc.
- Providing background reading.
- Correcting one's own work.
- Correcting the work of others.
- Criticizing previous work.
- Substantiating claims.
- Alerting to forthcoming work.
- Providing leads to poorly disseminated, poorly indexed, or uncited work.
- Authenticating data and classes of fact (physical constants, etc.).
- Identifying original publications in which an idea or concept was discussed.
- Identifying original publication or other work describing an eponymic concept or term (. . .).
- Disclaiming work or ideas of others (negative claims).
- Disputing priority claims of others (negative homage)

Autres typologies (1)

- **Krampen and Montada 2002 (psycho clinique)**

Citation category	**%**
Direct reference to an empirical finding in the cited document	30
Simple mention (of the type "compare here also," "see also," "see, for example") without any further more specific reference to the cited document	25
Direct reference to a theory or concept in the cited document	20
Direct reference to a method in the cited document	9
Overview citation (of the type "for an overview, see here," "see summary in") without any further reference to the cited document	5
Use of a data collection method (such as a test) taken from the cited document	3
Word-for-word quotation of text in the cited document	3
Use of a statistical method taken from the cited document	2
Substantial, theoretical, or methodological critique of the cited document	1
Use of a table, figure, or list taken from the cited document	0
Other citation type (for unclear citations)	2

Autres typologies (2)

- S. Teufel 2006 (TAL)

Category	Explanation	Distribution (%)
Weak	Weakness of cited approach	3.1
CoCoGM	Contrast/Comparison in Goals or Methods(neutral)	3.9
CoCo-	Author's work is stated to be superior to cited work	1.0
CoCoR0	Contrast/Comparison in Results (neutral)	0.8
CoCoXY	Contrast between 2 cited methods	2.9
PBas	Author uses cited work as basis or starting point	1.5
PUse	Author uses tools/algorithms/data/definitions	15.8
PModi	Author adapts or modifies tools/algorithms/data	1.6
PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper	1.6
PSim	Author's work and cited work are similar	3.8
PSup	Author's work and cited work are compatible/provide support for each other	1.1
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted/unknown citation function	62.7

Critique des typologies

- Points de vues différents
  - Fonctionnel, rhétorique, formel
- Phénomènes très variés
  - Parfois plusieurs fonctions à une même citation
- Difficulté à appliquer
  - Nécessité d'une expertise du domaine précis
  - Accord inter-juge faible
- Biais vers des publications d'un type précis
  - Généralement de sciences dures
  - emprunt de méthodes, de données, d'algos, comparaison de résultats, etc.
- Généralement peu adaptées à des publications en sciences humaines

Phénomènes transversaux

- Dans l'ensemble, très peu de citations «négatives»
  - Moins de 10% dans les études locales
  - Résultat utilisé pour confirmer la légitimité des analyses quantitatives
- Certaines fonctions sont couvertes par une notion floue de « superficialité » (*perfunctory*)
  - Positionnement des travaux, hommages, périphéries, etc.
  - Citations de « background »

Travaux de Teufel et al. (2006)

- Initialement, travaux sur le zonage rhétorique (argumentative zoning)
  - Identification automatique (par marqueurs) du rôle des différentes parties d'un texte
- Extension aux citations dans les articles scientifiques
- Méthode
  - Etiquetage manuel
  - Définition de marqueurs (patrons lexico-syntaxiques, critères dispositionnels, etc.)
  - Apprentissage automatique (classificateur bayésien)
- Corpus : 360 articles de TAL en anglais, au format XML

Teufel : Marqueurs utilisés

- 1700 « cue phrases » sur texte étiqueté
- agents (l'auteur du papier / quelqu'un d'autre)
- actions : 20 classes, certaines directement orientées vers les fonctions
- temps verbaux, modaux
- position relative dans le texte
- autocitation (au moins un auteur commun entre papier analysé et papier cité)

Teufel : Résultats

- Pour les 12 catégories :
  - Précision par catégorie entre 56 et 80%
  - Efficacité globale (accuracy) : 77%
- Modèle final à quatre catégories :
  - Positif / Négatif / Contraste / Neutre
  - 83% de précision globale
- Résultats très encourageants pour un tâche apparemment très difficile
- Mais : corpus assez « facile » (grandes revues/conférences de TAL très normalisées)

Troisième approche : Analyse de contenu des citations

- Principe : associer des termes/concepts à une citation
- Objectifs :
  - Organiser les domaines scientifiques et les réseaux de citation
  - Identifier et qualifier des fronts de recherche et des sous-disciplines émergentes
  - Associer des mots-clés aux travaux cités (RI)
  - Résumé automatique des articles cités

Méthode d’extraction

- Méthodes :
  - Analyse superficielle des contextes de citations pour un article cité donné
  - Sélection des termes pertinents par recoupement des contextes de différents articles citants.

- Exemples ciblés:
  - *Such estimation is simplified from **HITS algorithm** (Kleinberg, 1998).*
  - *For a **comparison to other taggers**, the reader is referred to (Zavrel and Daelemans, 1999).*

Travaux de Schneider (2004)

- Approche hybride combinant une méthode bibliométrique et une analyse des contextes
- Corpus : 2517 articles médicaux de Medline et SCI
  - Entièrement analysés au niveau bibliométrique
- Extraction de « clusters » d'articles par analyse des cocitations
- Analyse automatique des contextes de citations (extraction des SN)
- Sélection des SN les plus souvent associés à un cluster
  - Validation par ressources externes (MeSH) et manuellement

Exemples (Schneider 2004)

- Contextes de citation d'un même article :
  - "*Plaque Index* (28) and *Gingival Index* (18) were recorded*.*"
  - "Data recorded during each examination included age, self-reported smoking (current smoker or non-smoker), and betel nut chewing status (current user or non-user), bacterial *plaque* (*Plaque Index*),51 and calculus accumulation (CI),52 *gingival* inflammation, (GI),53 ..."
  - "All subjects underwent clinical periodontal examination including the measurement of probing depth (PD), attachment level (AL), *gingival index* (GI),20 *plaque index* (PI),21 ..."
  - "*Gingival index* (GI): GI was used to assess the severity of *gingival* inflammation.23"
  - "*Plaque index* (PI)27 and *gingival index* (GI)28 scores ranged from 1 to 2 and from 2 to 3 for all teeth, respectively."
- Au final, sélection de « Plaque index » et « Gingival index »

Le projet Rhecitas : Vue d’ensemble

Contexte général

- Double difficulté :
  - Le français
    - Pratiquement absent des index de citations
    - Certaines disciplines ne publient plus en français
  - Les Sciences Humaines et Sociales
    - Mal organisées en terme de supports de publication
    - Très peu représentées dans ces index (psycho, droit, économie)
    - Peu ou pas étudiées dans les études sur la citation
    - Certaines disciplines n'utilisent pas couramment les bibliographies (littérature, arts)
- Travaux les plus proches :
  - Projet Scientext (A. Tutin, F. Grossman, F. Rinck)
  - Objectifs descriptifs et didactiques

Conséquences

- Pas de données prétraitées
  - Uniquement des articles « bruts »
- Pas de réseaux de citations construits
  - Pas d’approches quantitatives possibles
- Pas de typologie adaptée
  - Pas d’étude fine des fonctions
- Pas de modèles d’analyse disponibles
  - Pas de marqueurs prédéfinis

Objectifs généraux

- Sur des corpus français de SHS
  - Etudier les caractéristiques des contextes de citation
  - Proposer une typologie de haut niveau
  - Extraire toute information pertinente
  - Le tout automatiquement
- Objectifs :
  - Etude de faisabilité d'une application à large échelle
  - Proposer un ajout d'information sur les publications en ligne
  - Etudier les phénomènes discursifs de la citation
  - Approche comparative entre les disciplines

Objectifs plus précis

- 1 : Approcher l’ « importance » d’une citation
  - Par des marqueurs linguistiques ad hoc
- 2 : Extraire des informations associées à une citation
  - Termes et concepts empruntés au travail cité
  - Termes entre guillemets et énoncés définitoires impliquant une référence

Le corpus

Définition du corpus

- Choix initial :
  - Publications intégrales en français
  - Dans les domaines des SHS
- Critères :
  - Disponibilité des sources
  - Formats facilement exploitables
  - Disciplines variées
  - Considérations politiques (projet financé par le TGE Adonis du CNRS)

Panorama des publications en ligne

- Revues.org
  - Développé par le Centre pour l'édition électronique ouverte (CLEO, CNRS)
  - XHTML
- Cairn.info
  - À l'origine quatre maisons d'édition (Belin, De Boeck, La Découverte et Erès
  - XHTML + PDF
  - Accès payant, mais contrat CNRS depuis juin 2008
- Erudit.org
  - Consortium interuniversitaire composé de l'Université de Montréal, de l'Université Laval et de l'Université du Québec à Montréal
  - PDF / XHTML, documents moins récents : PDF
- Persée
  - Site de numérisation rétrospective de revues françaises en sciences humaines et sociales
  - Université Lumière Lyon 2, et le Centre Informatique National pour l'Enseignement Supérieur (CINES),
  - documents en XHTML, originaux TEI (XML) non accessibles en ligne
  - Accès libre sur toutes les collections
- HAL-SHS
  - archive institutionnelle des EPST français pour le dépôt par les chercheurs
  - Documents principalement en pdf
  - Pas de modèle éditorial, grande disparité

Corpus actuel

- 612 articles issus du portail *revues.org* :
  - 8.105 références, 10.482 appels de citation identifiés et analysés, 4,5 millions de mots
  - 11 journaux (dont interdisciplinaires)
    - Psychologie, didactique, linguistique, ethnologie, géographie, sociologie, histoire, droit, anthropologie, philosophie
- Documents au format XHTML normalisé
- Moissonnés par le protocole OAI-PMH
  - Norme internationale des Archives Ouvertes (Open Archive Initiative – Protocol for Metadata Harvesting)

La machinerie

Chaîne de traitement

- Extraction automatique des références et des contextes de citation
  - Programmes Perl spécifiques (INIST)
  - Inspirés des modules Paracite (utilisés dans CiteSeer)
- Analyse syntaxique
  - Cordial Analyseur (Synapse Développement)
  - Etiquetage, lemmatisation, identification des fonctions syntaxiques
- Annotation des citations sur la base de marqueurs linguistiques (CLLE-ERSS & IRIT)
  - Grammaires locales via la plateforme logicielle GATE (General Architecture for Text Engineering)

Extraction des références et repérage des appels de citation

- Analyse des références
  - Données non explicitement marquées
    - Uniquement « Bibliographie » et liste d'items
  - Extraction par patrons :
    - Des auteurs (uniquement le nom)
    - De l'année de publication
- Marquage des appels de citation
  - Repérage des auteurs et des années
  - Marquage XML spécifique ajouté aux marquages XHTML initiaux

Exemple de données marquées

- À l'instar de ce que nous enseignent **<cit idref="2">Howard Gardner (1993)</cit>** , **<cit idref="8">Francisco Varela, Eleanor Rosch et Evan Thompson (1994)</cit>**, **<cit idref="1">Didier Demazière et Claude Dubar (1997)</cit>** ou encore **<cit idref="13">Ludwig Wittgenstein (1958)</cit>**, les résultats de la présente étude nous mettent en garde contre la tentation de définir une catégorie par un ensemble d'attributs normatifs.

- Le rapport de stage relève de ce que Yves **<auteur>Reuter</auteur>** définit comme « l'écrit de recherche en formation ».

La plateforme GATE

- General Architecture for Text Engineering
  - Université de Sheffield, 1995
- Plateforme développée pour le marquage automatique de textes
  - Application centrales : extraction d'information et fouille de textes
- Modulaire, avec définition de chaînes de traitements en cascade
  - Segmentation, étiquetage, projection de lexiques, etc.
- Avantages :
  - Gestion des annotations XML multi-niveaux
  - Fonctionnalités d'affichage et de recherche
  - Langage de définition de patrons (JAPE)
  - Communauté bien établie et standard répandu
  - Logiciel libre et multi-plateforme

Développement dans GATE

- Module pour l'analyse syntaxique
  - Intégration de Cordial Analyseur
  - Alternative à TreeTagger : meilleure qualité et analyse syntaxique
- Transducteurs en cascade pour la recherche de patrons
  - Langage JAPE
  - Prend appui sur le marquage initial (HTML + citations) et les informations (morpho) syntaxiques

Premier Objectif : repérer les citations « importantes »

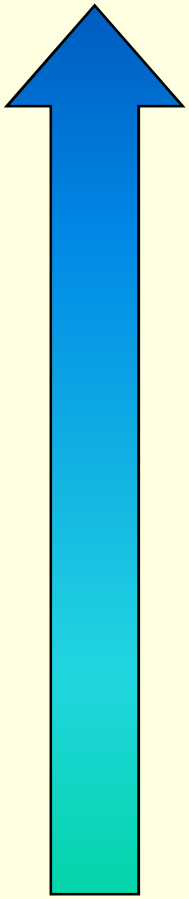
Premières intuitions

- *Citations de « background » :*
  - Au sein d'énumerations
  - Faiblement intégrées à la phrase : parenthèses, notes de bas de page, etc. (Swales 1990)
  - Situées en début d'article
  - Associées à des marqueurs (« voir aussi », « par exemple », etc.)
- Exemples :
  - ***Plusieurs études*** *ont montré que les mères favorisent les phrases courtes (Brown et Bellugi, 1964 ; Drach, 1969 ; Lord, 1975 ; Moerk, 1975 ; Nelson, 1973 ; Newport, 1975 ; Phillips, 1973 ; Sachs, Brown et Salerno, 1972 ; Shatz et Gelman, 1973 ; Snow, 1972, 1977).*
  - *Dans les sociétés occidentales,* ***les auteurs*** *(Caroll, 1974 ; Montandon et Crettaz, 1981 ; Lombardo, 1989 ; Chauvenet et al., 1994 ; Crouch, 1995 ; Crawley, 2004 ; Chauvenet, 2006) ont montré que ce groupe [...]*
  - *On sait maintenant (****voir par exemple*** *Perfetti, 1985) que la différence entre bons et mauvais lecteurs porte surtout [….]*

Premières intuitions (suite)

- *Citations « Importantes »*:
  - Isolées
  - Bien integrées dans la phrase (sujet, SN possessifs, etc.)
  - En collocation avec des marques de première personne, des verbes spécifiques
- Exemples:
  - *En cela, **nous** allons dans le sens des **considérations de** J.-P. Bronckart (1994) sur la double nature du genre.*
  - *C'est ainsi que **nous** avons établi, **à la suite de** Bogaards (1988) , une distinction théorique entre 'processus' et 'stratégie'.*
  - *[…] hypothèse **formulée par** Lenton (1989) et correctement **rejetée par** Hagan (1991) .*

Proposition d’échelle d’intégration



- Sujet (+ début de phrase)
- Sujet (+ marques première personne)
- SN possessif (+ marques première personne)
- « *Selon X* »
- « *Par exemple X* »
- Dans la phrase, isolée
- Entre parenthèses, isolée
- Dans la phrase, dans un
- Entre parenthèses, isolée, en fin de phrase
- Entre parenthèses, dans une énumération, en fin de phrase

Traits repérés automatiquement
(Fanny Lalleman et Marjorie Raufast)

- La citation est seule ou dans une énumération.
  - Seule (75%) :
    - *Hampson et Nelson (1993) ont examiné le discours maternel dans deux contextes : repas et jeux libres*.
  - Enumération (25%) :
    - *Plusieurs études ont montré que les mères favorisent les phrases courtes (Brown et Bellugi, 1964 ; Drach, 1969 ; Lord, 1975 ; Moerk, 1975 ; Nelson, 1973 ; Newport, 1975 ; Phillips, 1973 ; Sachs, Brown et Salerno, 1972 ; Shatz et Gelman, 1973 ; Snow, 1972, 1977).*
- La citation est ou n'est pas entre parenthèses :
  - Parenthèses (80%) :
    - *L'importance de l'automatisation des processus de base de la lecture a été prise en compte par tout un secteur de recherche particulièrement prolifique, y compris en langue seconde (McLaughlin, 1990).*
  - Phrase (20%) :
    - *Une définition très succincte de la notion est fournie par S. Moirand (1979 : 19) : une stratégie de lecture correspond à « comment le lecteur lit ce qu'il lit. »*

Traits (2)

- La position de la citation dans la phrase :
  - Initiale (8%) :
    - *C. Dévelotte (1989) se propose de mettre en évidence les stratégies individuelles de lecture mises en œuvre par des apprenants-lecteurs en F.L.E pour reconstruire le sens d'un article de presse.*
  - Centrale (42%) :
    - *C'est ainsi que nous avons établi, à la suite de Bogaards (1988) , une distinction théorique entre 'processus' et ''stratégie.*
  - Finale (50%) :
    - *Il ne peut pas non plus rendre compte du rôle global – et non pas ponctuel – que joue autrui dans le développement des compétences langagières (Pekarek, 1999b)*
- La citation est sujet (6%) :
  - *Dans un souci méthodologique Matthey (1996) **tente** de distinguer dans l'interaction les différents niveaux d'analyse qui sont pertinents pour l'étude de l'acquisition.*

Traits (3)

- La présence dans une structure du type « selon X » (1,5%)
  - « Selon X» :
    - ***Selon*** *Long (1996) , celle-ci a pour but de parvenir à une certaine transparence sémantique.*
  - « Pour X » :
    - *De même, **pour** Gass et Selinker (1994 : 333) , ce terme désigne « the language that is available to learners, that is, exposure. »*
  - « D'après X » :
    - *Notre deuxième mesure de la flexibilité du taux de change (**d'après** Hausmann, Panizza et Stein, 2001) est donnée par les réserves monétaires internationales sur M2 […]*

Traits (4)

- Les citations à proximité d'un pronom de première personne (2%) :
  - ***J'****ai ainsi montré que les oppositions syntaxiques ergatif/accusatif, actif/passif, de même que la notion de sujet (ou d'objet) syntaxique, n'avaient pas de pertinence à être posées en LSF (Cuxac, 2000).*
  - *En cela, **nous** allons dans le sens des considérations de J.-P. Bronckart (1994) sur la double nature du genre, lorsqu'il postule que chaque texte bien que relevant fondamentalement d'un genre, constitue une unité autosuffisante.*
- Les citations à proximité d'une structure du type « par exemple X » (2%)
  - *On sait maintenant (**Voir par exemple** Perfetti, 1985) que la différence entre bons et mauvais lecteurs porte surtout sur l'efficacité des mécanismes «de bas niveau»[…]*
  - *[…] l'opération même de ce pouvoir transforme le sujet transgresseur en sujet délinquant, 'matérialisant' le savoir qui naît de l'examen minutieux de son corps et de sa psyché (**voir aussi** Rose, 1999).*

- ***Evaluation** : 95% de précision et de rappel pour le repérage de chacun de ces traits*

Expérimentation et résultats

Méthode

- Définition d'une base pour la comparaison
  - Annotation manuelle
  - Prenant en compte des variations
    - Types d'articles
    - Types d'annotateurs
- Comparaison avec une méthode automatique
  - Basée sur l'échelle proposée
- Unités prises en compte :
  - Annotation manuelle des références
  - Classification automatique des citations

Annotation manuelle

- (F. Lalleman, 2009)
- Divers problèmes dus au contexte politico-universitaire…
- Au final :
  - 1 discipline, 2 articles, 3 annotateurs
  - 107 citations, 69 références
- Questionnaire :
  - Pour chaque référence dans cet article, la jugez-vous importante ou liée à des informations d'arrière-plan ?
  - Avec des exemples de fonctions issus de (Garfield 62)

Annotation manuelle : détails

- Discipline : Linguistique et didactique du FLE (Revue : AILE)
- 2 articles de types différents :
  - Article 1 : état de l'art
  - Article 2 : expérimentation
- 3 annotateurs de compétences différentes
  - Tous étudiants de Master 2 SDL
    - 1 expert des thématiques traitées
    - 1 spécialiste du FLE (mais autre thématique)
    - 1 spécialiste d'une autre discipline (TAL)

Accord inter-annotateur

- Calcul du coefficient Kappa :

Kappa	**Juge 1 vs 2**	**Juge 1 vs 3**	**Juge 2 vs 3**
Article 1	0.84	- 0.12	- 0.05
Article 2	0.59	0.20	0.26

- Accord élevé entre experts, faible avec le non-spécialiste
  - Cas de désaccord marqué sur le premier article
  - Situation moindre pour le second article
- Clairement, double influence de la compétence et du type d'article

Classification automatique

- Règle de décision :
  - Citation importante = dans la phrase + isolée
    - L'importance prévaut dans le cas de citations multiples
- Evaluation :

Exactitude (*accuracy*) (%)	**Juge 1**	**Juge 2**	**Juge 3**
Article 1	38	31	61
Article 2	63	63	74

- Meilleure pour le juge non-spécialiste
- Meilleure pour l'article « classique »
- Mais au final, résultats assez faibles

Conclusions sur cette tâche

- Une tâche difficile
  - Confirme les résultats de Hanney et al. (2005) pour l'accord inter-annotateur
- Une tâche qui dépend
  - Des compétences du lecteur
  - Du type d'article
- Une méthode de classification proche du comportement d'un non-spécialiste
  - Les marques d'insertion dans le texte comme stratégie par défaut en l'absence de connaissances du domaine

Second objectif : extraction de mots-clés

Principes

- Pas de croisement des références
  - Pas de possibilité de regrouper les contextes de citation d'une même référence
  - Pas d'approche par fréquence
- Solution de repli : analyse plus fine des contextes
  - Marques explicites permettant d'identifier les termes/concepts associés par l'auteur à une citation
- Deux marqueurs :
  - Guillemets
  - Enoncés définitoires

Passages entre guillemets

- Extraction de SNs entre guillemets :
  - *Celles-ci semblent englober plusieurs aspects d'un modèle langagier représentant approximativement un des stades du* ***« processus de complexification »*** *(Corder, 1978) d'une langue institutionnalisée*
  - *C'est-à-dire que l'on interprète souvent ce passage de la* ***« revanche à la contrainte »*** *(Cohen, 1985 , 76) en y voyant un développement profondément positif.*
- *Limites :*
  - *La libération conditionnelle devient avant tout une mesure de gestion peu coûteuse de condamnés objectivés comme des « **déchets** » […] (Simon, 1993, 142 et 259; Simon et Feeley, 2003, 99; 1994, 193 ; Lynch, 1998).*
  - *Le* ***« hourrah football »*** *(Sansot, 1991, p. 142),* ***le football des trottoirs*** *(Therme, 1995), investit la rue […]*

Enoncés définitoires

- Basé sur les travaux de J. Rebeyrolle (2000)
  - Patrons déjà décrits et formalisés
  - Essentiellement basés sur des verbes : *appeller*, *nommer*, *dénommer, utiliser/proposer/employer le terme/mot/expression*
  - Adaptés au contextes de citation (citation proche)
- Exemples
  - *[…] les relations causales entre les événements enchaînés seront manifestées à travers une organisation logique des paramètres opposés et complémentaires, ce que Yau (1992 :146)* ***appelle*** *le paradigme de coordination.*
  - *L'individualisation des trois formations discursives de la déviance criminalisée que je propose se limitera, pour l'essentiel, à ce que Foucault* ***nomme*** *la différenciation primaire des objets (97)*
- Plus de 800 termes extraits (650 guillemets + 150 définitions), précision de 80%

Remarques

- Phénomène très fréquent dans notre corpus
- La fonction de citation associée est à rapprocher de
  - *Identifying original publication or other work describing an eponymic concept or term (Garfield 62)*
  - *Direct reference to a theory or concept in the cited document (Krampen & Montada 2002)*
  - *Author uses tools/algorithms/data/definitions (Teufel 2006)*
- Plusieurs résultats très stables
  - Confirmés par des recherches Web associant le terme et l'auteur

Mots de la fin

Fiches synthétiques

- Génération automatique d'une bibliographie filtrée et annotée :

TITLE: Quelques sources de variation chez les enfants

AUTHOR: Jisa, Harriet et Richaud, Frédérique

REF: AILE, vol 4, 1994

Reference	Context and relevant features
Ref 7: BERNSTEIN, B. (1971). Class, Codes and Control, Volume 1. London : Routledge and Kegan Paul.	A propos de l'utilisation différentielle de noms et de pronoms, il nous semble important d'examiner le travail de **Bernstein (1971)**.
Ref 10: BLOOM, L., P. LIGHTBROWN & L. HOOD (1974). « Imitation in language development: if, when and why ? », Cognitive Psychology 6 : 80-420.	**Bloom, Lightbrown and Hood (1974/1991)** ont examiné l'utilisation de l'imitation dans le discours spontané de six enfants.
Ref 18: BRUNER, J. (1983). Le développement de l'enfant : Savoir faire, savoir dire. Paris: Presses Universitaires de France.	Il se peut toutefois que l'étude des corrélations, qui porte sur ce que **Bruner (1983)** appelle **les procédures**, c'est-à-dire les adaptations faites par le partenaire plus compétent, masque la véritable importance du travail sur le code dans la co-participation.

[...]

Conclusions

- Travail encore très embryonnaire
  - Malgré l'importance des traitements mis en place
- Difficultés
  - Accessibilité et richesse des données
  - Phénomènes visés très complexes
    - Grande variabilité (types d'articles / disciplines / points de vue)
    - Multiplicité des niveaux d'analyse et complexité des regroupements
  - Définition du contexte de citation
    - Besoin de quitter la phrase (coréférence…)

Autres pistes

- Comparaison de disciplines
  - Différences significatives des fréquences de traits entre les revues
    - Linguistique : plus d'énumérations, moins de parenthèses, plus de débuts de phrase…
- Etudier les articles globalement
  - En fonction de leur profil en terme de citations
  - Intuitions : différents types d'articles entraînent des interprétations différentes des indices
- Premières expériences sur les profils de citation
  - Etudier la répartition des appels
    - Dans le texte (vecteur de positions)
    - Entre les références (table de fréquence des appels par référence)

Exemples de profils de documents

- Répartition des citations dans le texte