RHECITAS : RHEtorique des relations de CITation dans les Articles de SHS

Ludovic TANGUY

Objectifs et démarche

- Analyse des citations dans les publications de SHS en français
  - Etude des fonctions rhétorique et discursives
  - Observation des habitudes des différentes disciplines
  - Enrichissement des publications en ligne
- 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

Financemenet et Partenaires

- **Financement TGE-Adonis (Très Grand Equipement pour l'Accès unifié aux DOnnées Numérlques 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

- Présentation du projet
- Contexte : l'analyse des citations
  - Analyse quantitatives et bibliométrie
  - Analyse qualitatives
    - Fonctions des citations
    - Analyses thématiques des contextes de citation
- Corpus
  - Les portails de publications
  - Prétraitements : repérage des citations et des appels
  - Corpus de travail actuel
- Traitements
  - Plateforme GATE
  - Phénomènes repérés, grammaires locales
- Résultats intermédiaires
- Première analyses
- Pistes et exploitations

ANALYSE DES CITATIONS : Analyses quantitatives

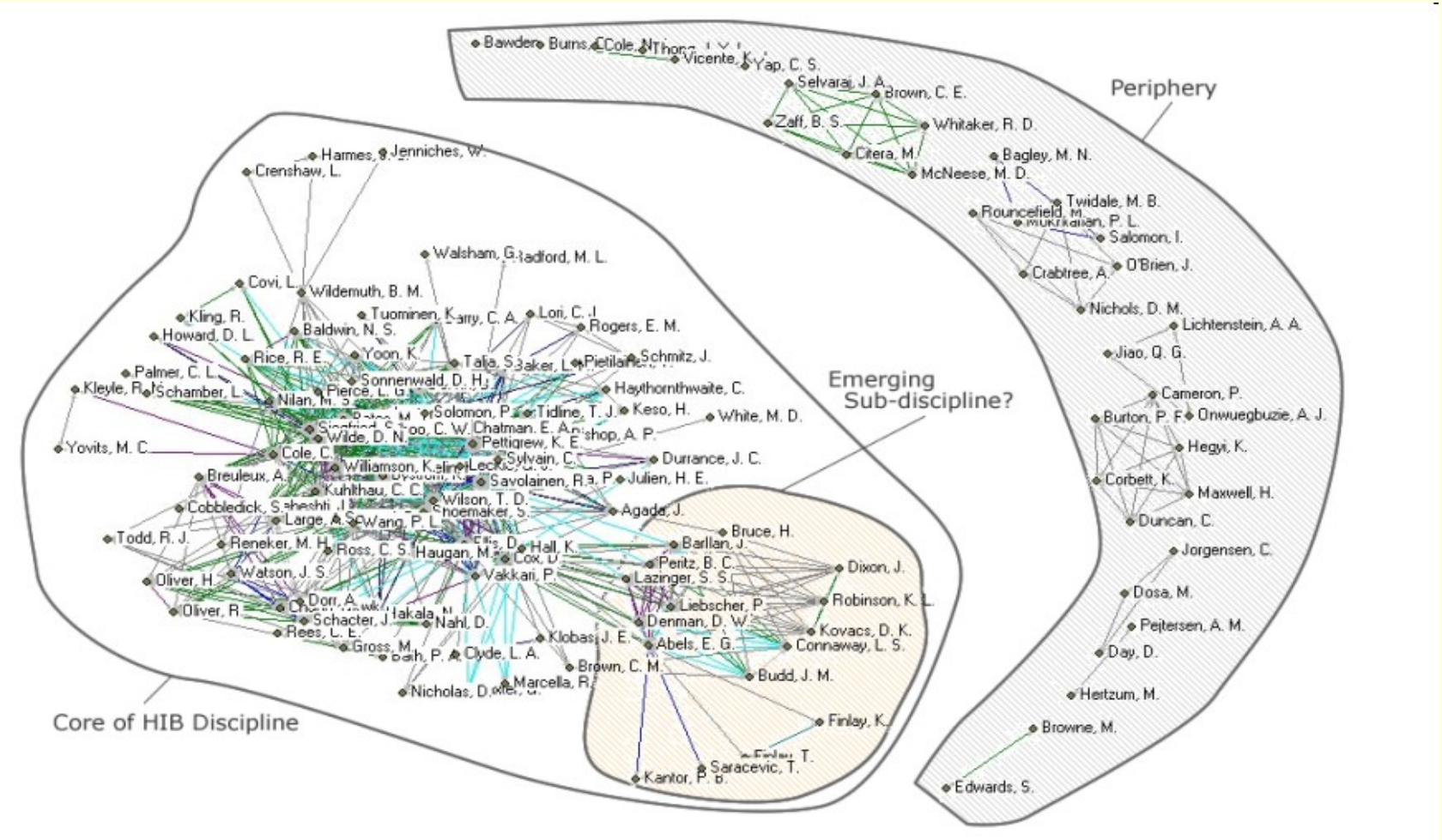
Analyses quantitatives

- Bibliométrie :
  - Analyse quantitative des publications scientifiques
  - Généralement, mesure des relations de citation entre les publications (l'article X cite l'article Y)
  - Calcul de la notoriété d'un article ou d'une revue, i.e. du nombre de citations dont il est la cible (facteur d'impact)
  - Etude des réseaux de co-citation
- Utilisations
  - Evaluation de la recherche
  - 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
  - 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 en ligne, diffusion 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

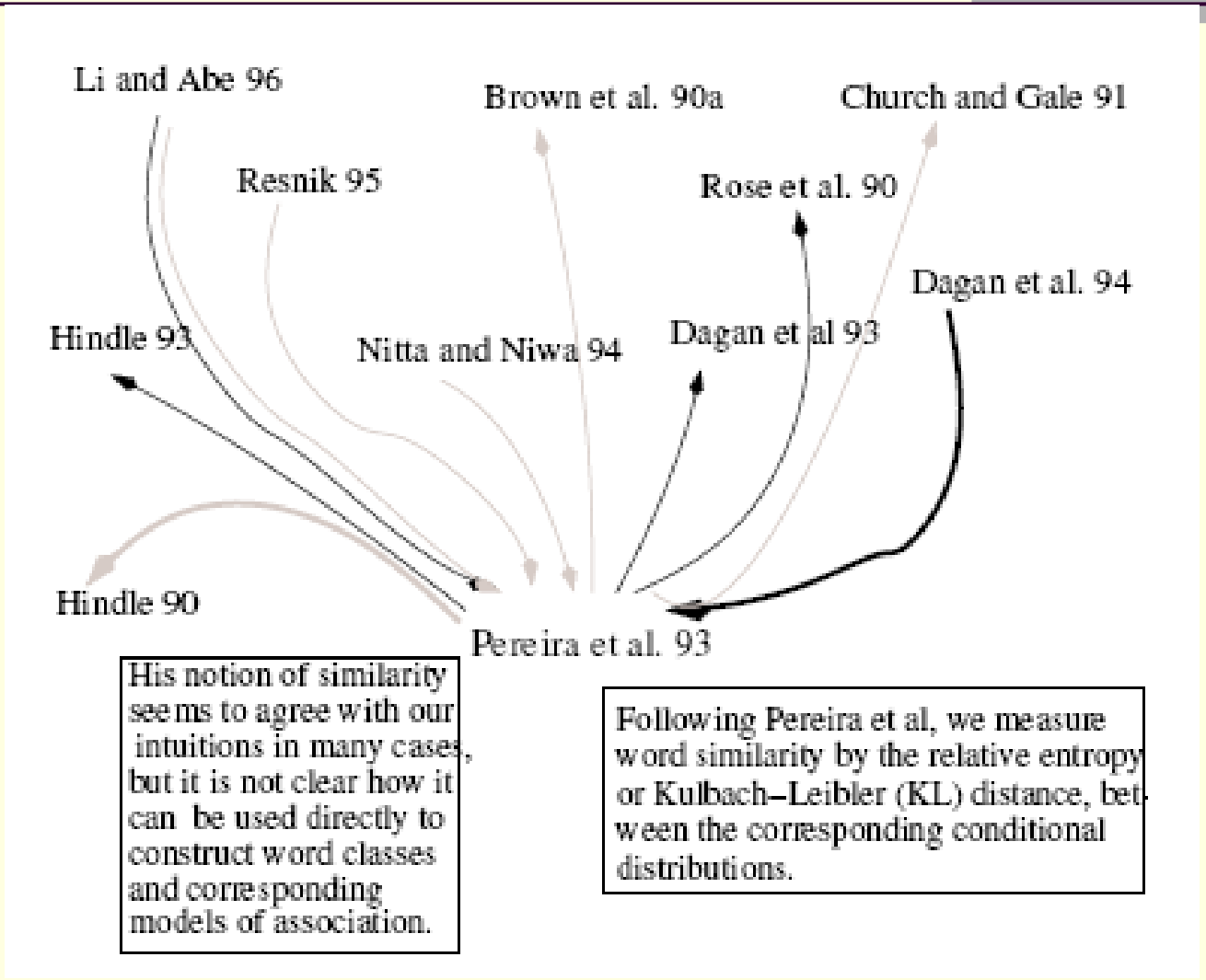
- Couverture des bases de données
  - Couvertures très variables entre les bases de données
  - Variations importantes entre les disciplines (Médecine > sciences > humanités)
- Fiabilité des analyses
  - Les références sont extraites automatiquement
  - Silence, et difficulté à normaliser
- Pas de prise en compte des « types » de citation
  - Une citation négative (critique) compte autant qu'une positive

ANALYSE DES CITATIONS : Analyses qualitatives

Analyses qualitatives

- Analyse des contextes de citations et plus seulement des listes de références
- Différentes approches
  - Identification des motivations des citations
    - Positif/négatif, important ou non, etc.
  - Identification de thématiques
    - Concepts liés à un acte de citation
- Objectifs
  - Etudes sur les mécanismes de la citation
    - Linguistique, didactique, sociologie des sciences
  - Enrichissement des index de citation
  - Indexation des documents
    - Repérage de mots-clés dans les documents citants

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 (psychologie)**

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
  - E.g. emprunt de méthodes, de données, comparaison de résultats
- 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 de « superficialité » (*perfunctory*)
  - Rendre hommage aux pionniers et aux collègues
    - L'auteur se donne une autorité par la référence aux travaux reconnus
  - 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)

Analyse des thématiques

- Principe : associer des concepts à une citation
- Objectifs :
  - Organiser les domaines et les réseaux de citation
  - Identifier et qualifier des fronts de recherche et des sous-disciplines émergentes
  - Associer des mots-clés aux travaux cités
- Méthodes :
  - Analyse superficielle des contextes de citations pour un article cité donné
  - Sélection des termes pertinents par recoupement des contextes de différents articles citants.

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 »

En résumé

- Etudes systématiques à moyenne échelle
  - Au sein d'une discipline
  - Bénéficiant de la disponibilité de corpus homogènes pré-traités
- Enjeux applicatifs clairs
- Et nous là-dedans ?
  - Textes en français (très très mal indexés, voire ignorés)
  - Sciences humaines et sociales (idem)

Le projet Rhecitas : Vue d'ensemble

Une exploration des phénomènes

- 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

Des restrictions

- Pas de données pré-traitées disponibles
  - Pour le français
  - Pour les SHS
- Pas de typologie satisfaisante
  - Développée pour les disciplines visées
  - Suffisamment générique pour une étude transversale
- Pas de réseaux de citation
  - Nécessite une masse de données et une communauté structurée

Premiers objectifs visés

- Approcher l' « importance » d'une citation
  - Intuition : échelle d'intégration de la citation dans le texte
  - Basé sur les remarques de Swales (1990)
  - Intégration forte :
    - Sujet de la phrase, en début de paragraphe, cooccurrence avec des marques de 1ère personne
  - Intégration faible :
    - Dans une énumération, entre parenthèses, en fin de phrase, dans l'introduction
- Extraction d'informations associées
  - Termes et concepts empruntés à un auteur
  - Passages courts entre guillemets et énoncés définitoires impliquant une référence

Les corpus

Les corpus

- Choix initial :
  - Publications en français
  - Dans le domaine 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

- 274 articles issus du portail *revues.org* :
  - *AILE* (apprentissage des langues)
  - *Champ Pénal* (criminologie)
  - *CyberGéo* (géographie)
  - *LIDIL* (linguistique)
  - *Terrain* (ethnologie)
  - 7000 références, 6000 appels de citation identifiés et analysés
- 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)
- Analyse syntaxique
  - Cordial Analyseur (Synapse Développement)
  - Etiquetage, 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 » (2001, 2004).

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
- (Petite démo rapide)

Les détails des phénomènes visés

Patrons développés (Travail effectué par Fanny Lalleman et Marjorie Raufaste)

- La citation est seule ou dans une énumération.
  - Seule :
    - *Hampson et Nelson (1993) ont examiné le discours maternel dans deux contextes : repas et jeux libres.*
  - Enumération :
    - *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 :
  - Parenthèse :
    - *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 :
    - *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. »*

Patrons (2)

- La position de la citation dans la phrase :
  - Initiale :
    - *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
    - *C'est ainsi que nous avons établi, à la suite de Bogaards (1988) , une distinction théorique entre 'processus' et ''stratégie.*
  - Finale
    - *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 position de la citation dans le paragraphe :
  - Initiale :
    - *I.1. Le lexique*

      Nelson (1973) a suivi dix-huit enfants anglophones entre l'âge de un an (1.0) et deux ans (2.0) afin d'étudier le passage des énoncés à un mot aux énoncés à plusieurs mots. […]

Patrons (3)

- La citation est sujet (récupération du verbe) :
  - *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.*
- La citation est dans une structure du type « selon X »
  - « 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 […]*

Patrons (4)

- Les citations à proximité d'un pronom personnel de première personne :
  - Pronom singulier :
    - ***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).*
  - Pronom pluriel :
    - *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 »
  - *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).*

Détection de concepts et de thématiques

- Les citations à proximité d'une structure entre guillemet :
  - *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.*
- Enoncés définitoires : travail basé sur les travaux de J. Rebeyrolle (2000)
  - *[…] 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).*

Autres traits calculés (traitement ultérieur, hors GATE)

- Présence de la citation dans une zone dense (au moins 4 autres citations dans une fenêtre de 200 caractères)
- Position relative dans le texte de la citation
- Etalement des appels de la même citation
- Nombre d'appels de la même citation
- Nombre total de citations dans l'article

Répartition des traits dans les sous-corpus

%	Densité	Enum	Parenth	Début Phr	Sujet	Selon	Pron	Début de par.	Définit	Guillem
AILE	9,09	39,06	69,70	15,32	15,49	2,36	3,03	5,56	0,84	2,02
CHAMP	2,57	26,56	83,74	6,07	7,41	1,49	2,99	1,39	0,15	3,96
CYBER	6,82	25,76	65,15	6,82	7,58	1,52	3,79	1,52	0,00	3,79
LIDIL	5,81	33,10	70,25	8,28	7,04	1,76	6,34	1,06	0,70	3,35
TERRAIN	1,97	19,00	83,22	5,32	7,43	1,13	3,79	0,94	0,44	7,33
TOTAL	3,53	25,74	80,03	7,08	8,29	1,48	3,68	1,65	0,40	4,97

Les (toutes premières) analyses

Niveaux d'analyse

- Trois types d'entités à analyser :
  - Les appels de citation (ou auteurs)
  - Les références (ensemble des appels associés)
  - Les articles
- Premier niveau envisagé : la référence
  - Calcul de ses caractéristiques en regroupant les caractéristiques de ces appels
  - E.g taux de citation à l'initial, taux de citation dans une zone dense, etc.
  - Calculs : étalement, médiane des positions, etc.
- Recollage des auteurs isolés à une référence
  - Problème pour les références multiples ayant le même auteur
  - Simple calcul du nombre d'occurrences de l'auteur dans le texte

Première analyse : classification

- Apprentissage automatique non supervisé
- Objectif : construire des classes « naturelles » en fonction des variables
  - Définition d'une distance entre deux références
- Algorithme : Expected Maximization
- Principes :
  - Partir d'une répartition aléatoire en classes
  - Par approches successives, déplacer les éléments d'une classe à l'autre en cherchant à
    - Minimiser la distance intra-classe
    - Maximiser la distance inter-classe

Premiers résultats

- Deux classes naturelles
- Première classe (73% des références)
  - Dans des parenthèses
  - Plutôt dans des énumérations
  - Non sujet, sans cooccurrences avec des pronoms, pas d'attaques, peu d'occurrences des auteurs
- Deuxième classe (27% des références)
  - Dans la phrase, quelques fois sujet, auteurs répétés, etc.
- Intuition : « background » vs. « importantes » ?

Précautions

- Pas d'indépendance a priori des variables
  - Entre parenthèses -> pas sujet
- Certains traits sont trop sporadiques pour être bien pris en compte
  - Pronoms / sujet / présence de guillemets, etc.
- La notion d'importance vs background n'a pas été bien balisée
  - Petite expérience avec Fanny et Marjorie (pas les mieux placées…)
  - Kappa = 0.6 (« acceptable » mais faible)

Pistes pour cette distinction

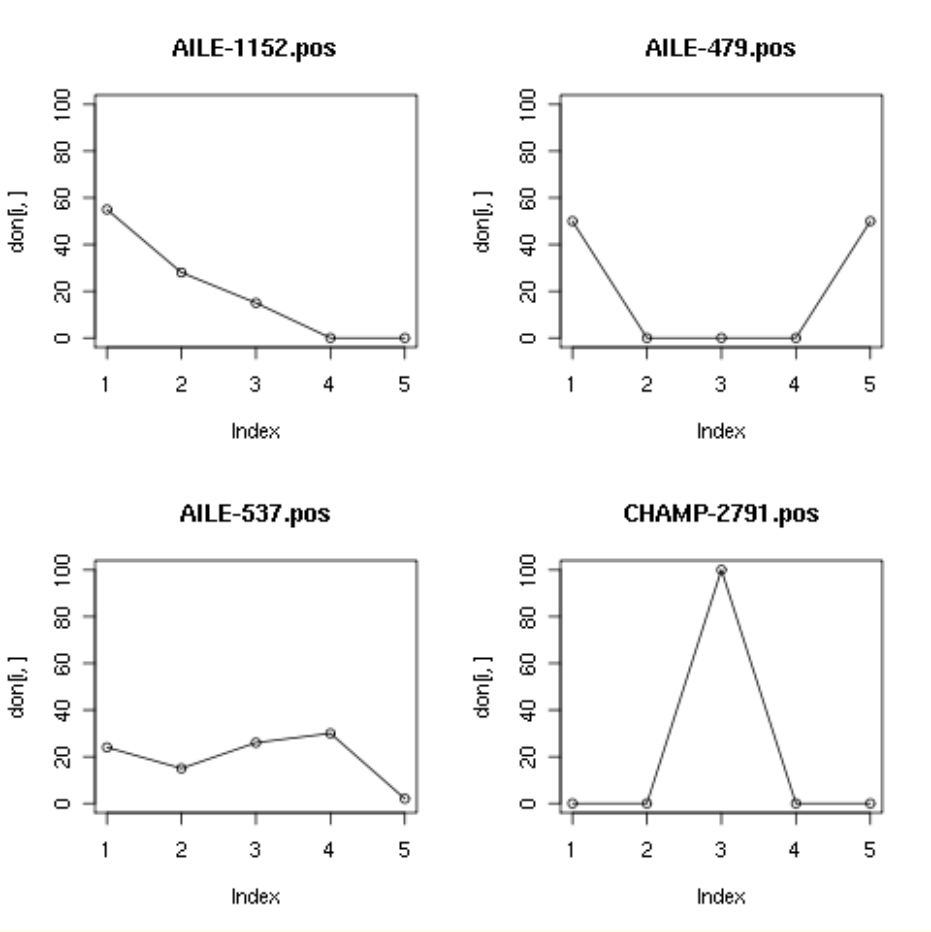
- Organiser une campagne d'annotation manuelle
  - Utiliser des annotateurs experts de leur domaine
  - Définir un protocole et une définition claire de la distinction
- Effectuer un apprentissage supervisé

Autres pistes

- Etudier les articles globalement
  - En fonction de leur profil en terme de citations
- Intuitions :
  - Différents types d'articles entraînent des interprétations différentes des indices
  - Etat de l'art / étude ciblée d'un auteur / schéma expérimental classique (état de l'art + développement d'un travail + comparaison)
- 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



Conclusion

Conclusion

- Travail encore très embryonnaire
  - Malgré l'importance des traitements mis en place
- Difficultés
  - Phénomènes visés très complexes
  - Grande variabilité (types d'articles / disciplines / style des auteurs)
  - Multiplicité des niveaux d'analyse et complexité des regroupements
    - Phénomène de déjà-vu en TAL

Perspectives à court terme

- Valoriser les premiers résultats
  - Présenter des informations associées à une bibliographie d'un article en ligne
  - Premiers contacts enthousiastes avec revues.org
  - Exemple :
    - Liste des positions et fréquence d'apparition dans le texte
    - « contextes intéressants »
    - « concepts liés »
- Travailler sur les notions de position (étalement, portée, entrelacement)