Relations discursives entre marquage faible et fort: approche multi-genre, multilingue et multimodale

Ludivine Crible, F.R.S.-FNRS & UCLouvain

17 janvier 2019, Université Toulouse Jean Jaurès

Pourquoi est-on ambigu ?

Au programme

A. Première partie: projet et premiers résultats

1. Marquage implicite-faible-fort: cadre théorique et état de l'art
2. Marqueurs du discours et variables d'association
3. Approche contrastive
4. Approche expérimentale
5. Marqueurs faibles/forts et leurs signaux: études en cours et étapes suivantes

B. Deuxième partie: ressources et méthodes sur corpus

1. Le corpus DisFrEn
2. Modèle d'annotation fonctionnelle à deux niveaux
3. Taxonomie des signaux discursifs (en cours)
4. Perspectives d'application

Les maillons faibles du discours écrit et oral

Marqueurs, contexte et cognition

Première partie: projet et premiers résultats

Le paradoxe des marqueurs discursifs

- Les MDs sont définis comme des **instructions** pragmatiques facilitant l'interprétation des relations discursives et d'autres fonctions de structuration.
- Pourtant, un petit nombre d'expressions est utilisé à **haute fréquence** pour exprimer un large panel de relations:
  - *ils trouvent pas de techniciens là-bas **et** ici ils en ont de trop (VALIBEL blaJV1l)*
  - *je l'ai cherché **et** je trouvais rien du tout (ilcCF1r)*
  - *elle avait un problème de boiler **et** j'ai été lui remettre un nouveau (blaJV1l)*
- Cette ambigüité (ou « sous-spécification ») concerne une proportion considérable des MDs les plus fréquents (*et, mais, donc*) et indique un décalage entre le **sémantisme** du MD et son usage en contexte enrichi **pragmatiquement**.

Implicite et explicite: une dichotomie ?

- Dans une perspective lexicale ou *marker-based*, les relations discursives sont souvent dites implicites (= pas de MD) ou explicites (= un ou plusieurs MDs)
  - les deux sont annotées dans les corpus comme le PDTB ou ANNODIS
  - stratégie de traduction étudiée sur corpus parallèles (Hoek et al. 2017)
- Certaines études prennent en compte d'autres signaux discursifs:
  - par ex. antonymes, temps verbaux, ponctuation, chaînes de co-référence…
  - peu de relations sont vraiment implicites (Das & Taboada 2018, Péry-Woodley et al. 2017)
- Niveau intermédiaire souvent exclu de ces études: les marqueurs « faibles »
  - *et* peut exprimer le contraste, la conséquence, le changement de sujet…
  - ➢ Polyfonctionnalité, voire sous-spécification

La notion de « force du signal »

- Certaines relations discursives sont plus fortement marquées que d'autres
  - *connective strength*, *relation markedness* (Asr & Demberg 2012, 2015)
- La force du signal correspond à l'association mutuelle entre un MD et une relation donnée ; elle augmente si:
  - le MD exprime peu d'autres relations (= monosémie)
  - la relation n'est pas exprimée par beaucoup d'autres MDs (= exclusivité)
- Par ex., l'addition est souvent marquée par *et*, mais *et* exprime beaucoup d'autres relations ➔ la relation d'addition est donc faiblement marquée

➢ Un marqueur polyfonctionnel constitue donc un signal faible

  ➢ on s'attend à plus de marqueurs faibles dans les textes oraux et spontanés

Sous-spécification: l'égoïsme discursif ?

- Décalage entre le sens (de base) du marqueur et la relation exprimée en contexte
  - « *the semantics of the connective that is used to indicate the link does not fully match the semantics of the relation that is intended by the speaker or writer* » (Spooren 1997)
- Spooren explique ce déséquilibre par le « R-principle » : ne pas dire plus que nécessaire
  - contrairement au « Q-principle » : dire autant que possible
  - ➢ économie de production (ou « égoïsme ») vs. économie d'interprétation

- Les MDs facilitent la compréhension (Haberlandt 1982, Degand & Sanders 2002)
  - ➢ Qu'en est-il des marqueurs sous-spécifiés ?

L'hypothèse de la compensation

- Les MDs ne sont pas les seuls signaux de relations discursives
- Ils peuvent être redondants avec un marquage prosodique, syntaxique ou lexical
- Des études en TAL (Yung et al. 2017) ou en psycholinguistique (Rohde et al. 2017, Koornneef & Sanders 2013, Mak et al. 2013) ont montré l'influence mutuelle entre discours, syntaxe et sémantique
- On peut s'attendre à ce que les MDs dits « faibles » ou « sous-spécifiés » soient compensés par d'autres signaux, et que ces signaux varient en nature et en fréquence selon le contexte de production:
  - compensation renforcée dans les textes écrits et préparés

➢ Comparer les MDs faibles et forts et les envisager dans leur co(n)texte

Les MDs et leur cotexte

- Beaucoup d'études ont montré l'influence de facteurs **structurels** sur la variation fonctionnelle des MDs:
  - prosodie (Kleinhans et al. 2017)
  - co-occurrence (Cuenca & Marín 2009)
  - position (Degand 2014, Salameh, Estellés & Pons 2018), etc.
- Ces études font écho à la **Grammaire des Constructions** (CxG, Goldberg 1995), appliquée aux MDs notamment par K. Fischer (2006, 2010, 2015):
  - sémantisme de base invariable + contexte structurel (position, prosodie) + situation de communication

➢ Une **construction fixe** traduit une valeur sémantique bien établie, ancrée (Crible 2018)

  ➢ la sous-spécification n'est pas associable à des régularités structurelles

Et : le maillon faible ?

- *Et* est-il si peu informatif, si dépendant du contexte qu'on le suggère ?
  - Kitis (1995): *and* véhicule des émotions en contexte concessif (*her husband is in hospital and she is seeing other men*) ➔ quels signaux de compensation ?
  - Blakemore & Carston (1999): *and* force une interprétation temporelle (*Max didn't go to school [and] he got sick*) ➔ suppression impossible
- Il y a aussi des contextes où il ne peut pas être remplacé par un MD plus fort
  - *The U.S. government makes a lot of mistakes and I think that Iraq is one of them*
- On peut aussi imaginer que l'emploi de MD faible reflète en fait une relation faible
  - ➢ un mot vague pour une idée vague, volontairement ou non
  - ➢ un *et* peut être le meilleur choix possible

Résumé intermédiaire

- Approche cognitive de la variation fonctionnelle des MDs et de leur co(n)texte
- Implicites (pas de MD)

Résumé intermédiaire

- Approche cognitive de la variation fonctionnelle des MDs et de leur co(n)texte
- Implicites (pas de MD) < sous-spécifiés

Résumé intermédiaire

- Approche cognitive de la variation fonctionnelle des MDs et de leur co(n)texte
- Implicites (pas de MD) < sous-spécifiés < polyfonctionnels/faibles

Résumé intermédiaire

- Approche cognitive de la variation fonctionnelle des MDs et de leur co(n)texte
- Implicites (pas de MD) < sous-spécifiés < polyfonctionnels/faibles < forts

Résumé intermédiaire

- Approche cognitive de la variation fonctionnelle des MDs et de leur co(n)texte
- Implicites (pas de MD) < sous-spécifiés < polyfonctionnels/faibles < forts

compensation compensation compensation

- HP1: les MDs faibles sont plus fréquents en situation de pression cognitive (principe R)
- HP2: ils sont compensés par d'autres signaux, surtout dans les textes écrits préparés
- HP3: *et* n'est pas toujours compensé ni remplaçable

MDs et variables d'association
Résultats de thèse

- Analyse contrastive (angl-fr) des MDs et des disfluences
  - identification sur corpus oral multi-genre des MDs sans liste prédéfinie
  - annotation manuelle de leur fonction, position, co-occurrence
  - annotation des disfluences adjacentes: pauses, *euh*, répétition, troncation…
- Modèle fonctionnel élaboré pour l'oral, 30 fonctions groupées en 4 domaines:
  - idéationnel: cause, conséquence, temporel, contraste, concession, condition, exception, alternative
  - rhétorique: motivation, conclusion, opposition, pertinence, reformulation, approximation, commentaire, spécification, emphase
  - séquentiel: ouverture, fermeture, retour au sujet, nouveau sujet, citation, énumération, addition, ponctuation
  - interpersonnel: contrôle, politesse, accord, désaccord, ellipse

MDs et variables d'association

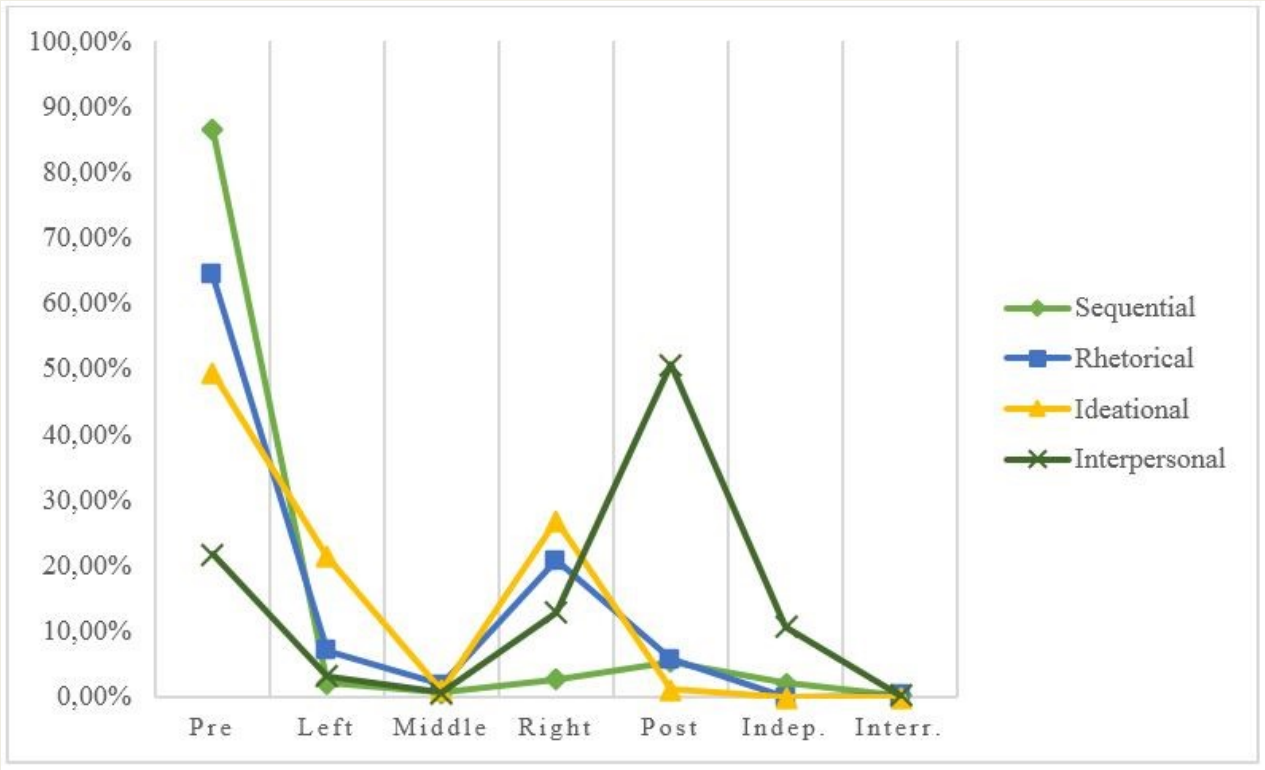
1) Fonction et position

- Hypothèse : fonctions objectives à l'initiale, fonctions intersubjectives en finale

- Séquentiel ➔ initial
- Idéationnel et rhétorique ➔ médial
- Interpersonnel ➔ final

Bon je pars en fait parce que j'ai fini hein

➢ Cycle temporel de production



MDs et variables d'association

2) Fonction et co-occurrence (Cuenca & Crible 2019)

- 24% des MDs en français font partie d'une chaîne de co-occurrence
  - *et alors* ; *et donc* ; *mais alors* ; *et comme* ; *mais si* ; *quoi tu vois* ; *ben écoutez*…
- Phénomène plus fréquent dans les **langues romanes** qu'en anglais

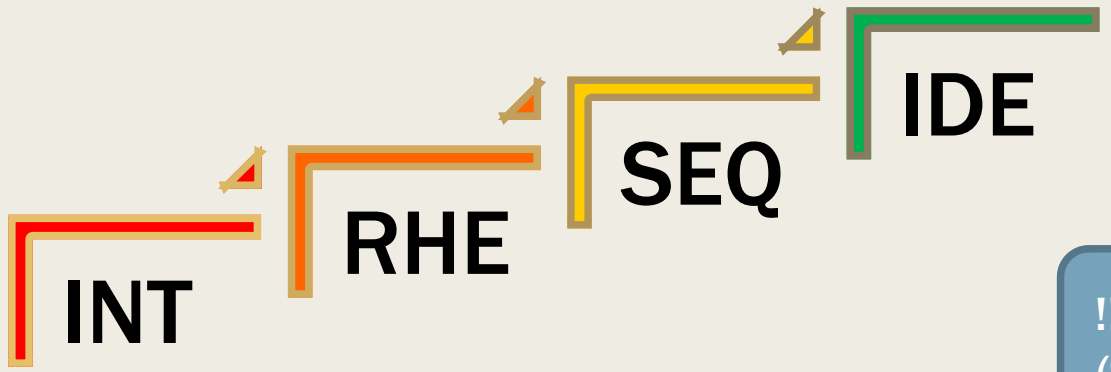
- Juxtaposition ➔ syntaxe et fonction différentes (*le truc où **tu vois quand** tu mets*)
- Addition ➔ un MD renforce l'autre (***mais d'un autre côté*** ; ***enfin j' veux dire***)
- Composition ➔ fonction unique, unité complexe, très fréquent (***et puis*** ; ***mais bon***)

➢ La présence contigüe d'un 2e MD peut **préciser voire modifier** le sens du 1er MD

MDs et variables d'association

3) Fonction et disfluence

- Chaque domaine fonctionnel montre une préférence pour certaines disfluences
  - Idéationnel ➔ **intégré** à une syntaxe/prosodie standard (pas de disfluence)
  - Rhétorique ➔ combiné avec des **répétitions** et des interruptions
  - Séquentiel ➔ très souvent combiné à des **pauses** (détachement prosodique)
  - Interpersonnel ➔ associé à des **interruptions** (faux-départs et troncations)



!! Configuration statistique, pas (forcément) la plus fréquente !!

MDs et variables d'association

4) Fonction et prosodie (Didirková et al. 2018)

- Etude de production sur *et* (addition, spécification, conséquence, temporel) et sur *alors* (conséquence, spécification, nouveau sujet, concession)
- Peu de différences prosodiques entre les différents emplois de *et*
  - profil intégré pour addition et conséquence
  - profil semi-détaché (allongement et pause) pour temporel et spécification
- Les différences sont plus marquées pour *alors*
  - profil semi-détaché pour conséquence (sens de base) et pour concession
  - profil détaché (pause, allongement, reset du pitch) pour nouveau sujet et spécification

- Kleinhans et al. (2017): seules quelques relations peuvent être identifiées grâce à la prosodie (f0, intensité, caractéristiques inter- et intra-segmentales)

Approche contrastive

1) Polyfonctionnalité en FR et LSFB (Crible & Gabarró 2018)

- Etude contrastive des marqueurs d'addition en français et en Langue des Signes de Belgique Francophone (LSFB): *et* vs. 'SAME'
- Analyse fonctionnelle sur un corpus conversationnel (90min)
- SAME a plus de « concurrents » que *et* et est plus polyfonctionnel
  - 71.8% de *et* expriment l'addition vs. 57.8% pour SAME
  - SAME est aussi utilisé pour la reformulation et l'approximation
  - ➢ polysémie de SAME (addition et alternative) vs. monosémie de *et*

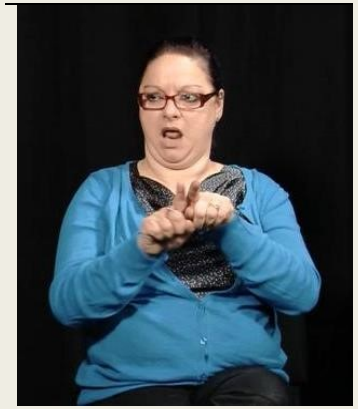


Figure 4. SAME

➢ La gamme fonctionnelle de marqueurs apparemment similaires peut varier

Approche contrastive

2) Sous-spécification en traduction (Crible et al. 2019)

- Etude de 3 types de sous-spécification sur **corpus parallèle** (TED Talks) en 5 langues
  - monolingue: déséquilibre entre le sémantisme et la fonction en contexte
  - multilingue1: traduction par un marqueur moins précis (*however* > *mais*)
  - multilingue2: omission (implicitation)
- La fréquence **d'omission** et le nombre d'équivalents de traduction dépendent de la fonction du MD dans la langue source
  - *and* est souvent omis, *but* rarement, *so* surtout hors-CSQ
  - les fonctions peu informatives sont souvent omises (addition séquentielle)
  - les emplois sous-spécifiés sont parfois traduits par des MDs plus spécifiques

Approche expérimentale

1) Prosodie et acceptabilité (Didirková, Crible & Simon 2018)

- Est-ce que le contexte prosodique facilite l'interprétation de MDs polyfonctionnels ?
- 2 études :
  - choix forcé entre deux versions prosodiques (avec question de compréhension)
  - prédiction de la relation du discours d'après le 1er segment et le MD
- Hypothèse : préférence pour la **prosodie associée à chaque relation/MD**
  - Résultats 1 : le profil associé à la relation est préféré pour *alors*; seul le profil non-marqué est acceptable pour *et*
  - Résultats 2 : la relation est bien prédite, surtout pour *alors*-conséquence; plus de variation pour *et*

➢ Le sens de base reste préféré (avec ou sans prosodie) + **biais pour la conséquence**

Approche expérimentale

2) Genre textuel et acceptabilité (Crible & Demberg 2018)

- Le MD sous-spécifié « *and* » est-il plus **acceptable** dans les genres informels ?
  - relations: **contraste** et **conséquence** (déjà désambigüisés par des naïfs)
  - genres: ***chat*** (conversationnel écrit) vs. **commentaire** d'article en ligne (préparé)
- MDs en concurrence : 0 < *and* < *so* – *but* < *therefore* – *however* (faible → fort)
- 2 études :
  - choix forcé entre deux versions (MD différent)
  - choix du MD parmi une liste de candidats (drag-and-drop)
- Hypothèse : préférence pour les marquages faibles en *chat* informel
  - ➢ effet du genre faible mais significatif, surtout pour le contraste (MD forts)

Marqueurs faibles et forts sur corpus
Premières étapes

- Identification et annotation manuelle des MDs dans différents corpus:
  - oral préparé & spontané, écrit préparé (presse) & spontané (*chat*)
- Focus sur *et* – *mais* – *donc* + les autres MDs exprimant les mêmes relations
- Désambigüisation des domaines et fonctions
- Identification des autres signaux discursifs

- Analyse quantitative des MDs et de leurs fonctions à travers les genres

Corpus utilisés

- LOCAS-F : corpus oral multi-genre (Degand, Martin & Simon 2014)
  - dialogues spontanés, années 2010: conversations, interviews (23,459 mots)
  - monologues préparés pour un public: conférences, discours académiques et politiques, journal, homélie (18,209 mots)
  - ➢ transcriptions alignées au son (format TextGrid, annotées dans EXMARaLDA)
- French Discourse Tree Bank (Danlos et al. 2015)
  - articles du *Monde*, années 1990 (535,000 mots)
  - connecteurs déjà identifiés (+ LEXCONN)
  - ➢ fichier XML (codés dans Excel)
- CoMeRe, corpus *getalp* (Falaise 2014), channel « actu »
  - conversations *chat*, années 2000 (48,188 mots)
  - ➢ fichier XML (codés dans Excel)

Premiers résultats quantitatifs

- 8437 MDs annotés au total
- MDs plus fréquents et plus variés à l'oral spontané qu'à l'oral préparé
- 73.7% des MDs expriment au moins 2 fonctions (N>10)
  - les MDs monosémiques sont rares mais très variés
- Et / Mais / Donc
  - Ils représentent plus de la moitié du total des MDs annotés
  - Proportion de *mais* et *donc* stable par corpus; *et* augmente en genres préparés
    - ET (2378) : addition, spécification, conséquence, contraste, concession, temporel, topic
    - MAIS (1382) : concession, contraste, addition, spécification
    - DONC (471) : conséquence, spécification

Marqueurs faibles et leurs concurrents

	Oral spontané	Oral préparé	Écrit spontané (CMC)	Écrit préparé (presse)	Total
Addition			en plus	d'autre part, de plus, en outre, or, par ailleurs, tandis que	23 types, 2388 tokens
Conséquence	alors	alors	alors	alors, ainsi, du coup, et, résultat	24 types, 1011 tokens
Contraste & concession	quand même, en fait		ceci dit, par contre, et	alors que, bien que, cependant, même si, pourtant, si, en revanche	52 types, 2677 tokens

- ➢ Forte spécialisation en genre et modalité
- ■ Et / Mais / Donc toujours plus fréquents que leurs concurrents, sauf:
  - *alors* > *donc* en CMC pour la conséquence (48 vs. 44)
  - *en revanche* > *mais* en presse (124 vs. 120) et *mais* = *et* à l'oral spontané (8)

Notre discours est-il égoïste ?

- Hypothèse (rappel): plus de marqueurs faibles en situation de pression cognitive
- La proportion de et/mais/donc semble indiquer que ce n'est pas systématique
  - besoin de raffiner l'analyse relation par relation (mesure de force du signal)
- Et/mais/donc ont des concurrents forts plus variés à l'écrit préparé mais ils restent globalement les MDs les plus fréquents pour leur relation de base

➢ La différence se joue peut-être au niveau des signaux de compensation

Prochaines étapes

- Calculer la force du signal pour les 3 MDs et les 3 relations

- Etablir une taxonomie exhaustive des signaux discursifs
- L'appliquer (manuellement) à tous les MDs…
- Comparer le nombre et le type de signaux par MD, par relation et par corpus

- Partie contrastive: répliquer sur des données comparables en anglais
- Partie expérimentale: tester les effets sur la production, perception et compréhension

Corpus, outils et méthodes pour l'analyse des MDs

Deuxième partie : ressources

Au programme

A. Première partie: projet et premiers résultats

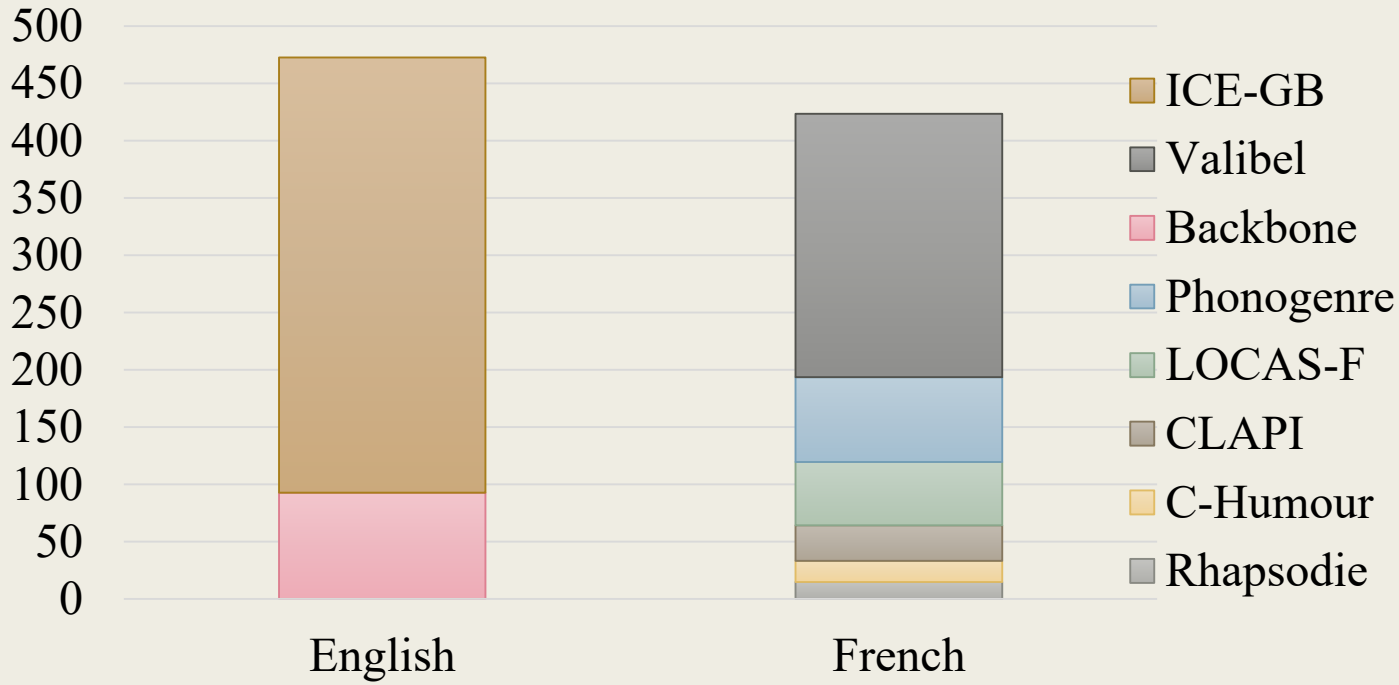
1. Marquage implicite-faible-fort: cadre théorique et état de l'art
2. Marqueurs du discours et variables d'association
3. Approche contrastive
4. Approche expérimentale
5. Marqueurs faibles/forts et leurs signaux: études en cours et étapes suivantes

B. Deuxième partie: ressources et méthodes sur corpus

1. Le corpus DisFrEn
2. Modèle d'annotation fonctionnelle à deux niveaux
3. Taxonomie des signaux discursifs (en cours)
4. Perspectives TAL

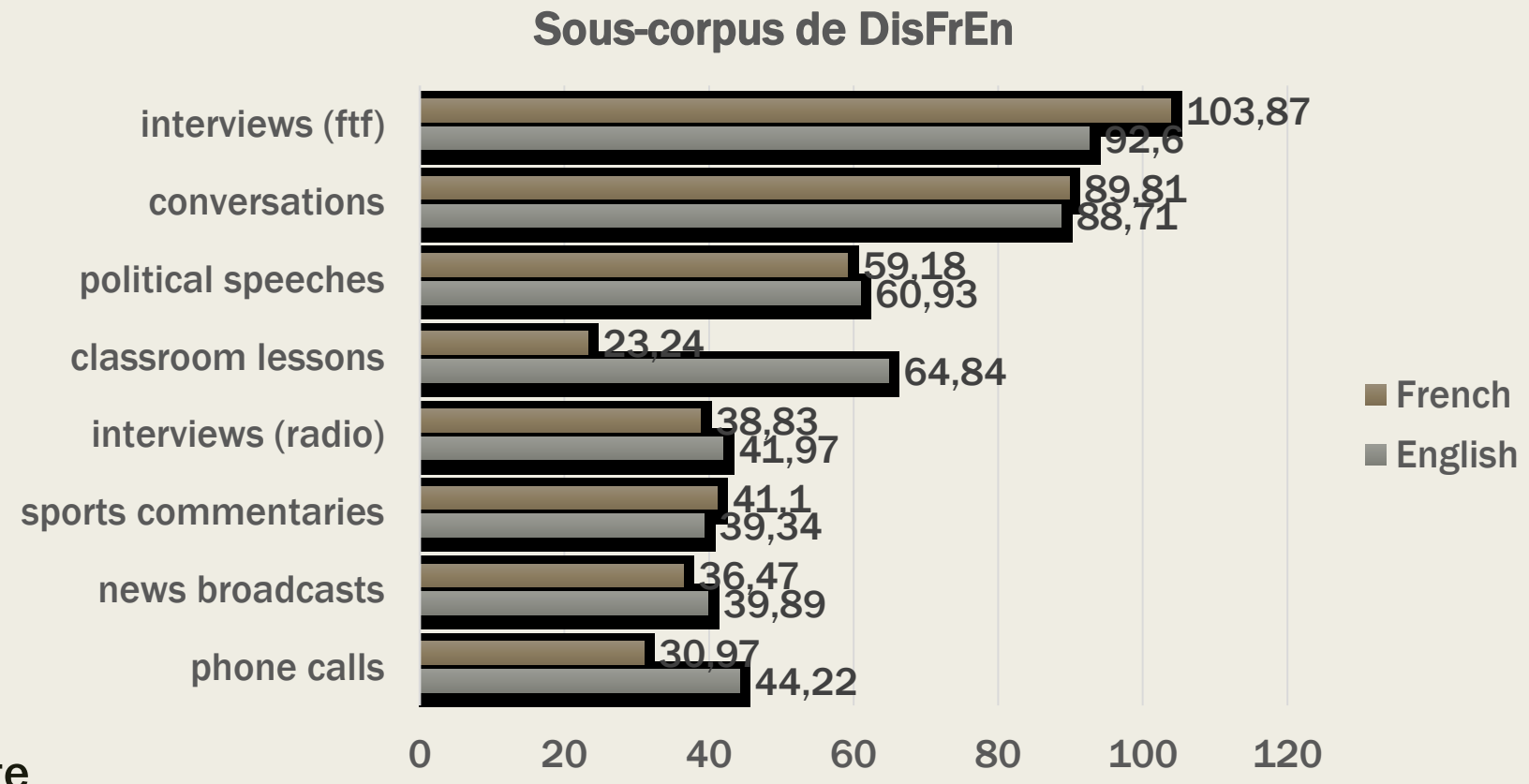
Le corpus DisFrEn

- Base de données annotées anglais-français oral
- Données « recyclées » à partir de différents corpus-sources
  - Anglais*:*
    - ICE-GB
    - Backbone
  - Français*:*
    - VALIBEL
    - C-Phonogenre
    - LOCAS-F
    - Clapi
    - C-Humour
    - Rhapsodie



Design comparable sur 8 genres

- 15h au total
- 161.700 mots
- 110 transcriptions



- Pas d'égalité parfaite
- Données manquantes dans 1 langue ou les 2

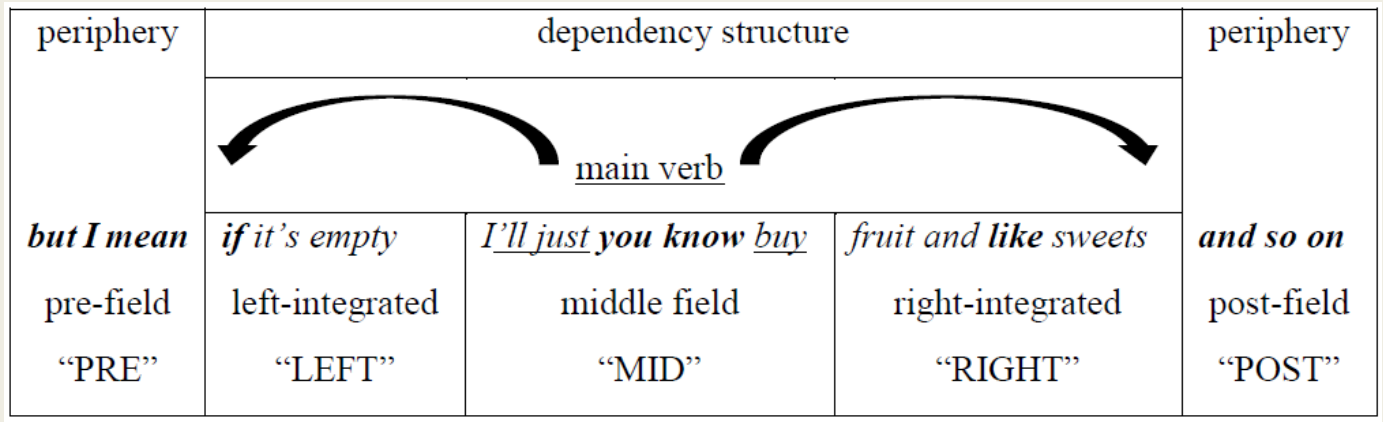
Annotations dans DisFrEn (I)
Type de MD

- Identification des MDs ouverte mais selon quelques critères:
  - optionalité syntaxique, haut degré de grammaticalisation, porte sur des unités autonomes, fonction pragmatique
  - 130 types en français : *à ce moment-là, à ce propos, à part cela, à propos, ah, ainsi, alors, alors que, après, au contraire, au fond, autrement, autrement dit, bah, ben, bien, bien que, boh, bon…*
  - exclus : pauses pleines (*euh*), modaux/epistémiques (*je crois*), tag questions…

- Origine syntaxique : CC *(et)*, SC *(parce que)*, ADV *(enfin)*, VP *(tu vois)*, ADJ (*bon*), PRO (*quoi*), NP (*sort of*), INTJ (*ouais*)

Annotations dans DisFrEn (II)
Position

- Dans le tour de parole
- Dans l'unité de dépendance (énoncé) ➔ intégré vs. non-intégré
- Dans la proposition minimale



Annotations dans DisFrEn (III)
Fonction

- Format du manuel inspiré du PDTB 2.0 + ajout de fonctions typiques de l'oral
- Domaines et fonctions sont inter-dépendants

Idéationnel	Rhétorique	Séquentiel	Interpersonnel
cause	motivation	ponctuation	contrôle
conséquence	conclusion	ouverture	politesse
concession	opposition	fermeture	désaccord
contraste	spécification	retour au thème	accord
alternative	reformulation	nouveau thème	ellipse
condition	pertinence	citation	
temporel	emphase	énumeration	
exception	commentaire	addition	
	approximation		

inter-annotator agreement κ = 0.406, 44.5%

intra-annotator agreement κ = 0.74, 75.8%

Annotations dans DisFrEn (IV)
Co-occurrences

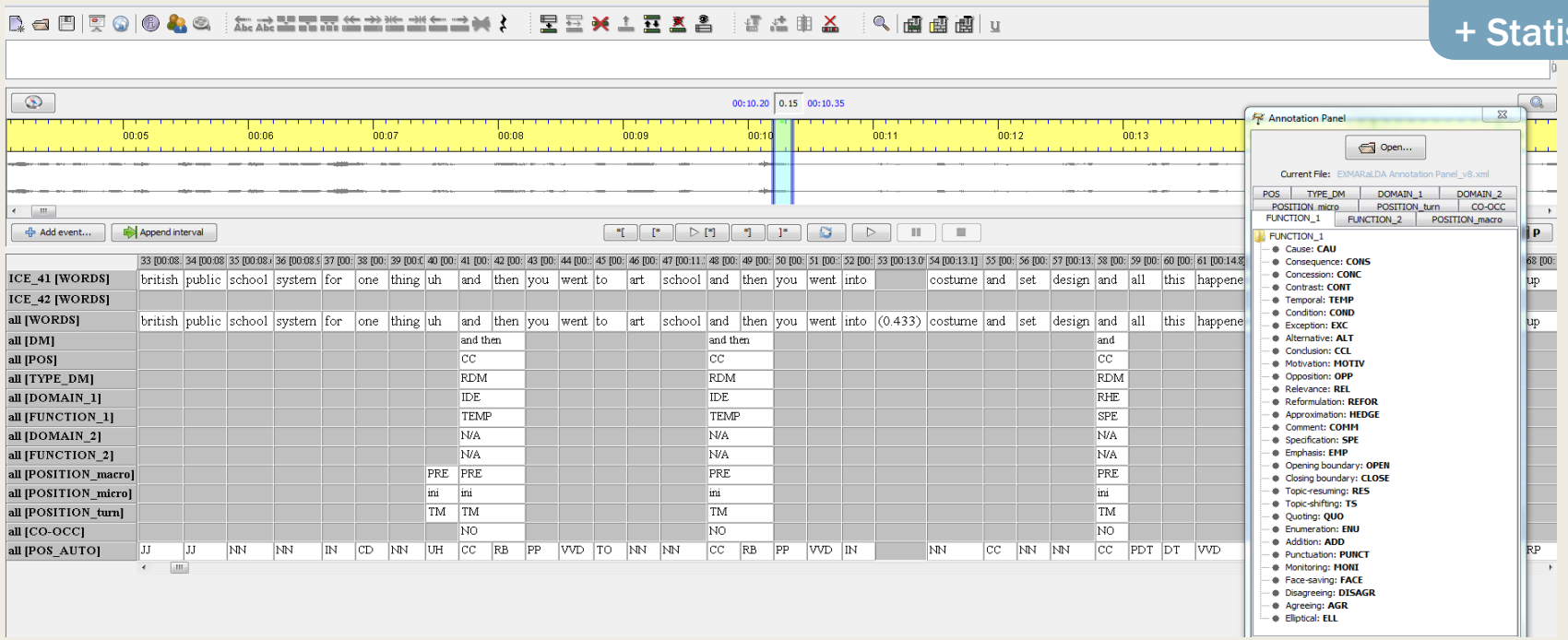
- MDs adjacents (par ex. *et pourtant*)
- Disfluences adjacentes et/ou contenant le MD (Crible et al. 2018)

Tags	Fluencemes	Examples
UP	unfilled pause (sec.)	(0.380)
FP	filled pause	*uhm, uh, euh*
DM	discourse marker	*so, because, well, I mean…*
ET	explicit editing term	*oops, what is it?…*
FS	false-start	"places are funny on (1.060) well they don't…"
TR	truncation	"tran/ uhm (0.700) transplant"
RI	identical repetition	"they go (0.630) eh they go"
RM	modified repetition	"a lot of time a lot of money"
SP	propositional substitution	"Asian speakers well no Asian people living in the UK"
SM	morphological substitution	"but there is there are"
	Related elements	
IL	lexical insertion	"I deal with disputes, so civil disputes"
IP	parenthetical insertion	"and the rainy (0.250) well touch wood the rainy"
DE	deletion	Mary didn't want to come Mary didn't come
	Diacritics	
AR	misarticulation	"to do resiv/ residential conveyancing"
WI	embedded fluenceme	"she and she"
OR	change of order	"normally would take you would normally take you"

Interface d'annotation

- Logiciel libre EXMARaLDA
  - Corpus Manager pour les métadonnées du corpus
  - Partitur Editor pour la transcription et l'annotation
  - Exakt pour les requêtes et le concordancier

+ Traitement dans Excel
+ Statistique dans R



Modèle d'annotation 2.0 (Crible & Degand ss pr.)

- Révision du modèle utilisé dans DisFrEn, avec deux changements majeurs:
  - réduction du nombre de fonctions
  - chaque fonction peut se combiner avec chacun des 4 domaines

➢ On étend le principe de la distinction objective-subjective à toutes les fonctions (en principe) et aux 4 domaines, pour en faire 2 niveaux indépendants

➢ On montre ainsi les liens entre certaines fonctions

Ideational	Rhetorical	Sequential	Interpersonal
[addition] [agreeing] [alternative] [approximation] [cause] [disagreeing] [monitoring] [concession] [condition] [consequence] [contrast] [quoting] [specification] [temporal] [topic]			

Multidimensionnalité

- En principe, chaque fonction peut se combiner avec chaque domaine
  - 4 x 15 combinaisons théoriques possibles
- En pratique, dans les données analysées jusqu'à présent, 42 sont attestées
  - le domaine interpersonnel reste marginal pour les relations du discours
  - certaines fonctions interactives n'ont pas d'équivalent idéationnel
  - 3 fonctions sur les 15 ne se combinent qu'avec un seul domaine
- Cf. Bunt (2011) : fonctions *general-purpose* vs. fonctions *dimension-specific*
- Le continuum multidimensionnel ne s'applique pas (encore) partout
  - mais évolutions possibles dans le temps ou dans d'autres langues

Fonctions multidimensionnelles

- 4 domaines: addition, alternative, cause, concession, conséquence, spécification
- 3 domaines: contraste, approximation, temporel
- 2 domaines: condition, monitoring, accord
- 1 domaine: désaccord, topic, citation

- Certaines combinaisons sont rares, mais cette répartition est encourageante
- La définition des domaines ne change pas par rapport au modèle d'origine
- Les fonctions correspondent désormais à un socle de base +/- abstrait

Exemples

	IDE	RHE	SEQ	INT
Addition	le grand frère avait un rôle de papa **et** en plus d'être papa il avait un rôle de d'essayer les choses avant nous	non je marchais pas ah non non j'ai pas couru (0.180) **et** j'ai fait encore un détour	Pacs avait fait une intendance aux baladins (0.780) **et** euh Camille lui dit euh tu oublieras pas de payer	<spk1> tu dis euh cheese pour le cliché et genre euh un peu pour se cacher <spk2> **et** un peu pour se cacher aussi ouai
Alternative	on est plusieurs **ou** tu me vouvoies	c'est pas pour ça qu'on fait de la musique mais c'est **enfin** c'est pas pour être reconnu dans la rue	euh ben j'ai fait euh deux ans **enfin** ma première et ma deuxième euh d'institutrice euh primaire	<spk1> j'avais repris euh des études en gestion des ressources humaines [...] <spk2> directement après? <spk1> ben euh **enfin** j'ai arrêté euh l'année passée euh avril et euh [...] l'année scolaire suivante
Concession	elle devait partir le lendemain **mais** elle n'est jamais partie	**si** la démocratie est un mot ancien, ici et maintenant la démocratie signifie la prospérité pour tous	c'était assez comique de les entendre parler comme ça euh des fillers (0.690) **mais** euh ouais puis après euh voilà quoi	cet auditeur euh vigilant il va vous dire tiens euh encore Jean d'Ormesson **mais** on entend Jean d'Ormesson à chaque automne

Accord inter-annotateur

- Ce nouveau système améliore nos scores d'accord en divisant la désambigüisation en 2 étapes indépendantes
- Scores: 71.16% domaines, 80.36% fonctions ; 57.45% D+F
- Variation plus importante des domaines, donc plus d'hésitation à ce niveau
  - domaine idéationnel souvent en cause des désaccords
- Les fonctions « de base » (addition, concession, conséquence) ont les meilleurs taux
  - spécification (*en fait, enfin*) et topic (*et*) ont les pires taux d'accord

Discussion du modèle

- L'exercice de regroupement de fonctions a permis de révéler des variantes jusqu'alors peu ou pas discutées dans la littérature
  - étendre le continuum objectif-(inter)subjectif à des relations comme l'addition
  - réconcilier le sens de base avec des fonctions à un niveau plus macro (SEQ)
  - ➢ distinguer le sens du MD et son rôle dans l'interaction

- Certaines combinaisons sont encore très rares pour être établies de manière fiable
- D'autres sont discutables, notamment le cas de spécification

➢ Testez-le et donnez-nous votre avis !

Vers une taxonomie des signaux de compensation

- Liste préliminaire de signaux à partir de la littérature et de mon expérience
  - prosodie/ponctuation, antonymes, sens du verbe, polarité négative, position du MD, temps verbaux, disfluences, termes évaluatifs ou epistémiques…
- Complétée au fur et à mesure de l'annotation → 25 signaux non-classés
- Annotés de manière très subjective, uniquement lorsqu'ils sont « pertinents » dans le contexte donné
  - cf. approche de Taboada & Das (2018): 1 seul signal ou 2 pour chaque relation

➢ Peu de cohérence dans l'annotation des signaux, taxonomie très hétérogène

➢ Mais a permis de couvrir de nombreux types de signaux

Du subjectif au statistique : trouver un compromis

- Adopter plutôt une approche statistique
  - caractéristiques systématiquement encodées
  - cf. Levshina & Degand (2017) pour les relations causales
- Mais tout n'est pas toujours pertinent, notamment dans les aspects sémantiques
  - multiplier les annotations du type présent/absent serait peu efficace

➢ Trouver un compromis entre le statistique pur et le subjectif pur:
  - cf. Péry-Woodley et al. (2017): signaux pré-marqués + signaux manuels
  - garder une trace de l'annotation subjective préliminaire

Proposition de taxonomie

- Sentence features (systématique)
  - modalité, polarité, temps, sujet, type d'unité + polarité-contraste, temps-contraste
- DM features (systématique)
  - position dans le tour/énoncé, MD adjacent, pause adjacente, portée
- Syntactic features (uniquement si présents)
  - parallélisme, autre construction (par ex. « c'est » présentationnel)
- Semantic features (uniquement si présents)
  - AltLex, relation sémantique, termes évaluatifs, termes épistémiques, acte de parole, deixis, noms propres, numéraux, démonstratifs
- Autres (uniquement si présents)
  - intonation, disfluences, ponctuation, oui/non, discours rapporté

Exemple annoté 1

« *si nous avons la responsabilité du pays nous donnerons des papiers à tous ceux qui n'en ont pas (0.230)* ***et*** *cessera dans ce pays la chasse aux malheureux* » (politique)

Domain	Funct	Turn	Utter	DM	Pause	Scope	Complex
IDE	CSQ	Med	Ini	0	UP	adjacent	simple

DM features

Mood	Pola	Pola-diff	Tense	Tense-diff	Subj	unit
Decl	Posi	Same	Future	Same	Diff	full

Sentence features

Syntax	Synt_spe	Sem	Sem-spe	Other	Other-spe
Constr	SV reversal	0	0	0	0

Complementary features

Dom-rel	Fun-rel
Tense	Constr, tense

Relevant features

Exemple annoté 2

« *tu vies donc tu peux savoir si la vie est bien ou pas, **mais** t'as jamais crevé pour savoir si la mort c mauvais* » (CMC)

Domain	Funct	Turn	Utter	DM	Pause	Scope	Complex
RHE	CTR	Med	Ini	0	UP	adjacent	simple

DM features

Mood	Pola	Pola-diff	Tense	Tense-diff	Subj	unit
Decl	Posi-neg	Diff	Past-pre	Diff	Same	full

Sentence features

Syntax	semant	Sem-spe	Other	Other-spe
Parallel	Anto	0	0	0

Complementary features

Dom-rel	Fun-rel
0	Anto, neg, parallel

Relevant features

Questions en suspens

- Peut-on parler de compensation quand le MD est déjà spécifique/fort ?
- Configurations rares et très variées : comment faire émerger des schémas statistiques ?
- Comment distinguer les signaux se rapportant au domaine et à la fonction ?
- Comment articuler les approches statistique et subjective ? Quel poids leur donner ?
- Comment appliquer le tout sur des milliers d'occurrences de manière efficace ?
  - possible d'automatiser les signaux sur corpus (oraux) non-parsés ?

- Toute suggestion est bienvenue !

Perspectives d'application

Lexique (Crible & Mendes 2018, Crible 2018)

- Lexique du français écrit → enrichir LEXCONN avec des annotations fonctionnelles
- Lexique du français et anglais oral & CMC → inexistant à ce jour (types + fonctions)
- Compléter ces lexiques avec des informations sur les configurations/signaux
- Distinguer les sens de base et les emplois contextuels

Entry	Core meaning	Domains of use	Contextual senses	Configurations
ET	addition	IDE, RHE, SEQ, INT	consequence, contrast, specification, topic…	…
DONC	consequence	IDE, RHE, SEQ	NA	…
	specification	IDE, RHE, SEQ	NA	…
DU COUP	consequence	IDE	NA	…

Perspectives d'application
TAL

- Corpus d'entraînement pour la désambigüisation pour l'oral et l'écrit

- Partie automatisable de la taxonomie des signaux discursifs
  - à confronter aux modèles d'identification des relations implicites
  - traduction automatique

- … d'autres idées ?

Merci pour votre attention !

ludivine.crible@uclouvain.be

UCLouvain

fnrs FREEDOM TO RESEARCH