Recueillir et exploiter des données en linguistique

Pascal Amsili

ILPGA – Lattice

avril 2026

Sorbonne Nouvelle

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation
   = des données

Les bases d'une généralisation

Généralisation

Il y a 7 099 langues dans le monde.

1. consensus sur la définition des termes/concepts
2. support empirique : observations à l'appui de la généralisation
   = des données

- la théorie doit énoncer des propositions **falsifiables**
- la théorie doit être validée par des observations **reproductibles**

Sorbonne Nouvelle

Autres exemples de généralisations

- Les enfants acquièrent les pronoms interrogatifs avant les pronoms relatifs.

Autres exemples de généralisations

- Les enfants acquièrent les pronoms interrogatifs avant les pronoms relatifs.
- Le choix entre subjonctif et indicatif est influencé par l'ordre des mots.

Autres exemples de généralisations

- Les enfants acquièrent les pronoms interrogatifs avant les pronoms relatifs.
- Le choix entre subjonctif et indicatif est influencé par l'ordre des mots.
- Le mot *aussi* en français est un *item à polarité positive*.

(1) a. Lina est malade et Janine aussi.
b. Pedro n'est pas malade, Marc non plus.
c. *Pedro n'est pas malade, Marc aussi.

Falsifiabilité

Proposition 1

Toutes les phrases du français comportent un verbe conjugué.

Falsifiable ?

Falsifiabilité

Proposition 1

Toutes les phrases du français comportent un verbe conjugué.

Falsifiable ?

Proposition 2

Il existe des phrases du français comportant trois verbes conjugués de suite.

Falsifiable ?

Falsifiabilité

Proposition 1

Toutes les phrases du français comportent un verbe conjugué.

Falsifiable ?

Proposition 2

Il existe des phrases du français comportant trois verbes conjugués de suite.

Falsifiable ?

On préfère une proposition falsifiable non (encore) falsifiée à une proposition non falsifiable.

Sorbonne Nouvelle

Reproductibilité

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on

(Ioannidis, 2005)

Selon une étude réalisée auprès de 1 500 scientifiques et publiée par Nature en 2016[6], plus de 70 % des chercheurs affirment avoir été incapables de reproduire une expérience scientifique d'un autre chercheur et plus de la moitié affirment avoir échoué à reproduire une de leur propre expérience.

Les principes FAIR

Les chercheurs s'appuient sur les connaissances scientifiques antérieures, notamment sur les résultats publiés dans les articles scientifiques.
La reproductibilité des résultats, ainsi que leur croisement, ne sont cependant envisageables qu'avec des données originelles et leurs conditions d'obtention.
C'est pourquoi la science ouverte vise à faciliter l'accès aux publications scientifiques et aux données de la recherche.
Cette facilitation s'accompagne d'un certain nombre de mesures pour rendre les données scientifiques facilement découvrables, accessibles, interopérables et réutilisables.
Ce sont les principes FAIR : Findable, Accessible, Interoperable, Reusable.

(Urfist Méditerrannée, 2019)

Exemple

Généralisation (provisoire)

Le mot *aussi* en français est un *item à polarité positive*.

(2) a. Lina est malade et Janine aussi.
b. Pedro n'est pas malade, Marc non plus.
c. *Pedro n'est pas malade, Marc aussi.

Exemple

Généralisation (provisoire)

Le mot *aussi* en français est un *item à polarité positive*.

(2) a. Lina est malade et Janine aussi.
b. Pedro n'est pas malade, Marc non plus.
c. *Pedro n'est pas malade, Marc aussi.

Généralisation (nouvelle version)

Le mot *aussi* est en concurrence avec *non plus* en contexte négatif. Le choix d'une forme est conditionné par le(s) facteur(s) suivant(s) : . . .

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Données introspectives

Jugements produits par les chercheurs ou des informateurs

- Grammaticalité
- Anomalie sémantique
- Relation sémantique (dont la paraphrase)
- Pertinence pragmatique
- Identité/différence

Données de corpus

Productions « naturelles » rassemblées par les chercheurs

- Productions langagières publiées (littérature, presse, discours...)
- Enregistrements oraux ou vidéo (multimodaux)
- Traces langagières d'activité (communication professionnelle, réseaux sociaux...)

Données expérimentales

Données recueillies dans un cadre contrôlé

- Productions langagières induites
- Jugements (grammaticalité, anomalie)
- Mesures psycho-physiques (temps de réaction, mouvements occulaires...)
- Enregistrements d'imagerie cérébrale (eeg, irm)

Données de terrain

Données recueillies dans un contexte spacial et temporel spécifique

- Enregistrements
- Enquêtes
- Entretiens

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Données introspectives

- Introspection des chercheurs, publication et échange entre linguistes
- « Chasse aux papillons »
- Recours à des informateurs
- Fabrication de données critiques

Données de corpus

- Nombreux corpus existants, toutes modalités (science ouverte)
- On constitue fréquemment un (sous) corpus pour une recherche spécifique
- Toute collection de données linguistiques naturelles ne constitue pas un corpus

Qu'est-ce qu'un corpus ?

Un corpus doit être :

- Représentatif
- Fini
- Numérisé
- Standard
- (Finalisé/stabilisé)

Définitions I

Représentatif Ne pas utiliser un seul journal pour parler de la langue journalistique, ne pas utiliser un seul auteur pour parler de la littérature...

Fini Même si les corpus incrémentaux sont de plus en plus répandu, il est difficile de comparer les analyses réalisées sur diffférents états du corpus.
$\rightarrow$ versionnage.

Numérisé Avant l'informatisation existaient des corpus aussi, mais l'utilisation et l'extraction d'information étaient extrêmement laborieux.

Définitions II

Standard

- Le corpus doit suivre les normes de la communauté :
  - Format, annotation, droits
- Une collection de textes qui n'est pas utilisée par plusieurs personnes pour des recherches variées n'est pas vraiment un corpus.
- Une recherche sur un corpus qui n'est pas accessible à d'autres chercheurs n'est pas/difficilement vérifiable/falsifiable
  $\rightarrow$ elle est moins scientifique.

Recommandations

- Partage du corpus
  - Qui travaille ? Qui paye ? Qui en profite ?
  - Ressource libre ? Disponible ? Utilisable ?
  - Principe FAIR (findable, accessible, interoperable, reusable)
- Protection de la vie privée
  - Consentement des locuteurs
  - Respect des droits d'auteur
  - Déclaration à la CNIL (+ comité d'éthique)
  - Méthodes d'anonymisation

Corpus : oppositions pertinentes

- support : papier/numérisés/discrétisés
- écrit/oral/vidéo (LSF) ; multi-modaux
- monolingue/multilingue/aligné
- synchronique/diachronique
- statique/dynamique (incrémental)

- brut *vs.* annoté

Données expérimentales

- Méthodes inspirées de la psychologie cognitive
- Champ de la psycho-linguistique
- Principe général : contrôle des variable d'intérêt ("toutes choses égales par ailleurs")
- Chaque expérience a une portée étroite, mais cumulative
- Pratique récente : *crowdsourcing* (externalisation ouverte, myriadisation)
- Autre pratique récente : utilisation des LLMs comme sujets experimentaux

Données de terrain

- Conditions de collecte souvent très spécifiques
- Dimension sociologique ou anthropologique déterminante
- Méthode partagée avec d'autres sciences sociales : possibilité de recherche croisée
- Les données collectées peuvent être introspectives, quantitatives (corpus) et même expérimentales

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Exploitation

selon catégories : exploitations distinctes, méthodes distinctes :

Données introspectives souvent exploitées dans une perspective catégorique, pour répondre à la question “est-ce que telle constuction est possible ou pas ?”

Données de corpus souvent exploitées pour répondre à des questions quantitatives : “telle forme est-elle plus répandue que telle autre”, ou “telle construction est-elle (fréquemment) associée à telle propriété ?”.
L’exploitation des données de corpus repose souvent sur une phase d’annotation

Données expérimentales très fréquemment exploitées pour trancher entre deux théories concurrentes qui font des prédictions incompatibles
N.B. Dans ces deux cas, rôle crucial : significativité statistique

Données de terrain en plus des usages semblables aux autres types de données, les données de terrain ont fréquemment un rôle patrimonial ou de documentation.

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Préparation des corpus

En partant d'un corpus "brut" :

- « Nettoyage » (suppression des balises html ; réparation des erreurs d'OCR...)
- Prétraitement : transcription (corpus oral), découpage en paragraphes ou en tours de parole...
  $\rightarrow$ des choix sont faits dès cette étape $\rightarrow$ documentation
- Annotation : enrichissement des données, par exemple :
  - Ajout d'informations morpho-syntaxiques
  - Découpage en constituants
  - Identification d'emplois spécifiques
  - Identification des intentions communicatives
  - ...

Qu'est-ce qu'annoter ?

Point de vue d'un informaticien

- Entrée : les données de départ sont constituées d'**items** (éléments) reliés entre eux
  - texte = suite linéaire de mots (1 relation d'ordre)
  - arbre = structure arborée de balises (2 relations d'ordre)
- les items appartiennent à un vocabulaire fini
- Sortie : chaque item de la donnée de départ est associé à une **étiquette**
- les étiquettes appartiennent à un (autre) vocabulaire fini

$\rightarrow$ les données et les annotations ont la même structure

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Etiquetage POS ("part of speech")

- item = "mot"
  donnée = séquence de mots (=phrase)
  annotation = séquence des catégories morpho-syntaxiques des mots dans la phrase
- plusieurs niveaux d'annotations possibles :

Le	*petit*	*chat*	*est*	*mort*
Det	Adj	NC	V	Adj
DetMSDef	AdjMS	NCMS	VIP3S	AdjMS

- difficulté : plusieurs étiquettes possibles pour chaque mot, un dictionnaire ne suffit pas !
- format textuel possible :
  phrase annotée = Le/Det petit/Adj chat/NC est/V mort/Adj

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Segmentation d'un texte en "chunks"

- chunk = constituant non récursif
- analyse syntaxique superficielle
- équivaut à un parenthésage total non récursif typé ou non
- peut être codé par une annotation B/I (Begin/In)

Il	$voit$	sa	$voisine$	$avec$	des	$jumelles$
(Il)	$(voit)$	$(sa$	$voisine)$	$(avec$	des	$jumelles)$
B	B	B	I	B	I	I
$(Il)_{GN}$	$(voit)_{GV}$	$(sa$	$voisine)_{GN}$	$(avec$	des	$jumelles)_{GP}$
GN-B	GV-B	GN-B	GN-I	GP-B	GP-I	GP-I

- format textuel possible :
  phrase annotée = <GN>Il</GN> <GV>voit</GV> <GN>sa voisine</GN> <GP>avec des jumelles</GP>

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Reconnaissance des entités nommées (extraction d'information)

- entité nommée = nom propre (personne/lieu/organisation), date, valeur numérique
- porteur de l'information factuelle d'un texte
- peut être codé par une annotation BIO (Begin/In/out)

En	2016	*les*	*Jeux*	*Olympiques*	*auront*	*lieu*	*a*	*Rio*	*de*	*Janeiro*
	date		evt	evt				lieu	lieu	lieu
O	D-B	O	E-B	E-I	O	O	O	L-B	L-I	L-I

- format textuel possible :

  phrase annotée = En <Date>2016</Date>, les <Evt>Jeux Olympiques</Evt> auront lieu à <Lieu>Rio de Janeiro</Lieu>.

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Reconnaissance des entités nommées (extraction d'information)

- format "CoNLL"

unité	étiquette
En	O
2016	D-B
,	O
les	O
Jeux	E-B
Olympiques	E-I
...	

1. Annoter pour quoi faire
Exemples d'annotations de phrases

Alignement de phrases bilingues

- nécessite des phrases alignées, traductions l'une de l'autre :

	J'	aime	le	chocolat
I	X			
like		X		
chocolate				X

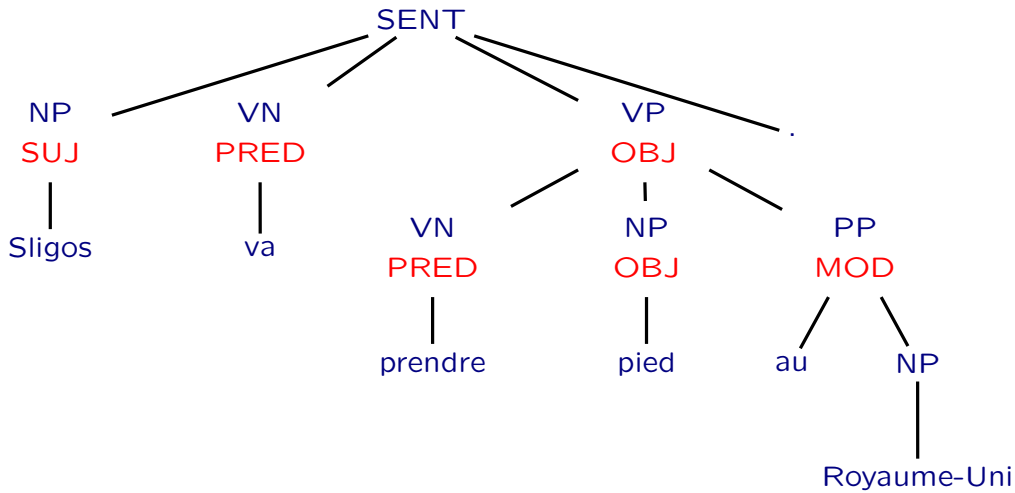
- on code les correspondances entre mots par des annotations

J'_1	$aime_2$	le_3	$chocolat_4$	I_1	$like_2$	$chocolate_3$
1	2	-	3	1	2	4

- chaque annotation réfère aux mots de l'autre phrase
- étape préliminaire des systèmes de traduction automatique statistique

1. Annoter pour quoi faire
Exemple d'annotations sur des arbres

étiquetage fonctionnel d'arbres syntaxiques



- étiquetage en rôles thématiques/sémantiques d'arbres syntaxiques : idem mais avec annotation agent, patient, etc.
- extraction d'information sur le Web ou les documents XML

1. Annoter pour quoi faire
Exemples d'annotations

Autres types d'annotations de séquences

- annotation de mots en phonèmes pour la synthèse vocale
- segmentation de textes en unités lexicales
- segmentation de phrases en clauses (propositions indépendantes)
- segmentation de dialogues en tours de paroles
- annotation de phrases successives d'une dépêche d'agence :
  item = phrase
  annotation = classe (evt présent / evt passé / commentaire)
- annotation d'une page HTML :
  item = segment de page
  annotation = classe (menu / publicité / titre / contenu)

1. Annoter pour quoi faire

Synthèse

- annoter une donnée = l'enrichir en préservant sa structure (mais sans la créer)
- permet de traiter de nombreuses tâches
- chaque tâche requiert de spécifier :
  - la nature des items (découpage initial)
  - les relations entre items : séquence, ordres dans un arbre...
  - la nature des annotations et leur interprétation
  - les relations entre annotations
  - les relations entre les items et leur annotation annotation possible (dictionnaires utilisables pour cela)
- pré-traitements et post-traitements souvent nécessaires pour coder/décoder les annotations
- différents formats d'annotation possibles, plus ou moins riches
- on peut faire des annotations successives de données

Utilisation des annotations

Les annotations servent :

- aux linguistes (phénomènes non surfaçiques) :
  - pour trouver des exemples
  - pour compter des occurrences

  dans un but de :
  - évaluation d'un modèle théorique
  - observation d'un phénomène linguistique
- aux informaticiens :
  - pour entraîner des systèmes de TAL
  - pour évaluer des systèmes de TAL

\+ comme « patrimoine »

Protocole d'une annotation manuelle multiple

Pour annoter à la main :

1. faire annoter une même portion des données par plusieurs annotateurs
2. observer les annotations qui convergent
3. discuter des annotations qui divergent jusqu'au consensus
4. mettre à jour le *guide d'annotation*
5. ré-appliquer les étapes de 1 à 4 sur les autres portions des données
6. produire un corpus annoté de référence
   (accord inter-annotateurs égal à 100%)

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité
- Plusieurs annotateurs : aucune possibilité de mesurer la qualité après adjudication (=arbitrage)

Qualité d'une annotation manuelle

Une bonne annotation est une annotation reproductible, donc généralisable

- Un seul annotateur : aucune possibilité de mesurer la qualité
- Plusieurs annotateurs : aucune possibilité de mesurer la qualité après adjudication (=arbitrage)
- Avant adjudication : mesure d'un **accord inter-annotateur**

Mesures d'accord inter-annotateurs

- on peut comparer les annotateurs entre eux
- on peut comparer les annotateurs avec la référence

Méthodes :

- Matrice de contingence

Sam \ Sim	OUI	NON
OUI	20	5
NON	10	15

- Kappa (κ) de Cohen (ou de Fleiss)

Mesures d'accord inter-annotateurs

- on peut comparer les annotateurs entre eux
- on peut comparer les annotateurs avec la référence

Méthodes :

- Matrice de contingence

Sam \ Sim	OUI	NON
OUI	20	5
NON	10	15

- Kappa (κ) de Cohen (ou de Fleiss)

accord = 0,7
κ = 0,4 ("accord modéré")

Annoter à la main...

- c'est long donc coûteux
- ça requiert des compétences linguistiques
- c'est à recommencer pour toute nouvelle donnée
- c'est la seule façon de garantir une annotation de grande qualité
- (à condition d'utiliser plusieurs annotateurs pour mesurer la qualité)

Annoter à la main...

- c'est long donc coûteux
- ça requiert des compétences linguistiques
- c'est à recommencer pour toute nouvelle donnée
- c'est la seule façon de garantir une annotation de grande qualité
- (à condition d'utiliser plusieurs annotateurs pour mesurer la qualité)

Alternative : annoter par programme !

Annotation : synthèse

- De très nombreux traitements linguistiques sur des données s'expriment par une tâche d'annotation
- Pour annoter, on peut :
  - annoter à la main (long, laborieux...)
  - chercher un programme existant (rarement parfaitement adapté au besoin) : il faut comprendre/analyser ses erreurs
  - construire soi-même un programme !
    - en programmant "à la main"
    - en annotant un échantillon et en utilisant de l'apprentissage automatique
    - en utilsant un LLM
- Dans tous les cas : des compétences linguistiques sont utiles, à un moment ou un autre...
- ... et des compétences informatiques !

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Un exemple

Une étude sur les conditions de la répétition obligatoire de *aussi*
Amsili *et al.* (2016)

(3) a. Lydia est en congé maladie, et Marcel est malade aussi.
b. ? Lydia est en congé maladie, et Marcel est malade.

(4) a. *Jo had fish and Mo did.
b. Jo had fish and Mo did too. *(Green, 1968)*

Gradation de la réduction du commentaire

Intuition : le contraste plus fort si le *commentaire* est réduit :

(5) a. Jo sent Helen a note and Mo sent Helen a note too.
b. ? Jo sent Helen a note and Mo sent Helen a note.

(6) a. Jo sent Helen a note and Mo sent Helen one (too / *∅).
b. Jo sent Helen a note and Mo did (so/it/∅) (too / *∅).

(Kaplan, 1984)

Le *commentaire* peut subir une « réduction » plus ou moins forte : arguments pronominalisés, voire ellidés, verbe réduit par une VP-ellipsis (en anglais).
Schéma de réduction progressive :

(7) a. sent Helen a note
b. sent her a note
c. sent her one
d. did so / it
e. did

Vérification expérimentale pour le français

Peut-on aller jusqu'à dire que plus le *comment* est réduit, plus *aussi* est obligatoire ?
Quel est le rôle de la répétition ?

Design :

- Expérience d'acceptabilité (AJT), sur Internet (iBexFarm). 80 participants.
- Mélangée avec 3 autres expériences, pour avoir des distracteurs.
- Jugements d'acceptabilité sur une échelle de 10 points.
- 24 exemples $\times$ 12 conditions

Réduction du commentaire

(8) Un étudiant a démontré ce théorème à Stéphane, et son collègue...

... a démontré ce théorème à Stéphane	aussi	`ful+`
... a démontré ce théorème à Stéphane		`ful-`
... l'a démontré à Stéphane	aussi	`cpt+`
... l'a démontré à Stéphane		`cpt-`
... lui a démontré ce théorème	aussi	`obl+`
... lui a démontré ce théorème		`obl-`
... le lui a démontré	aussi	`pro+`
... le lui a démontré		`pro-`
... l'a fait	aussi	`vpe+`
... l'a fait		`vpe-`
...	aussi	`vid+`
...		`vid-`

Résultats attendus

`ful+`	*problème de répétition*
`ful-`	*idem*
`cpt+`	*contraste croissant entre + et -*
`cpt-`	
`pro+`	
`pro-`	
`vpe+`	
`vpe-`	
`vid+`	*acceptabilité la plus élevée*
`vid-`	*acceptabilité la plus basse*

Table – Résultats attendus, 1re étude

Résultats (I)

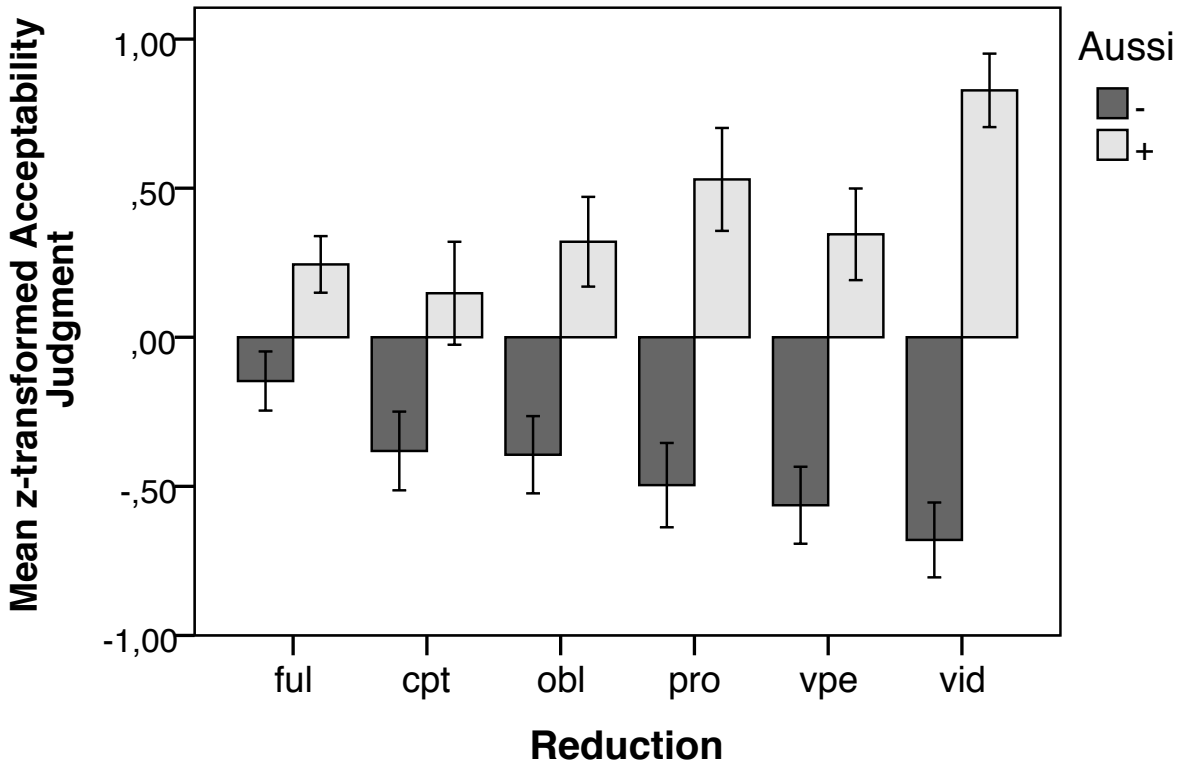
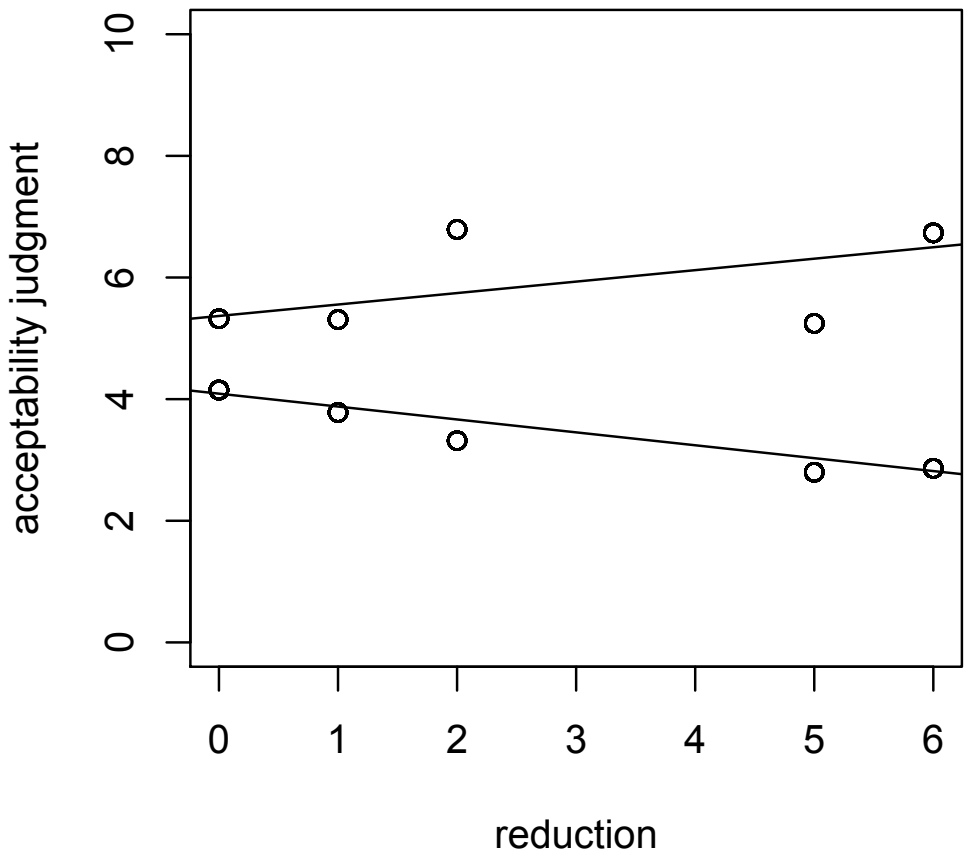


Figure – Mean answers normalized by participant : 0 denotes average answer, positive values indicate higher acceptability with 1 being one standard deviation better than the average sentence.

Résultats (II)



Modèle linéaire mixte

- **Linear mixed effects model** : réponse modélisée par rapport à
  - degré de réduction (0-6)
  - présence/absence de *aussi*
  - effets aléatoires sur les items et sur les participants

  Aussi a un effet positif très significatif sur les jugements ($\chi(1)$=415.08, p<0,001) ;
- le degré de réduction (seul) ne montre aucun effet ($\chi(1)<1$)
- les deux facteurs interagissent de manière significative ($\chi(1)$=74.31, p<0,001) :
  - Avec *aussi*, l'acceptabilité croît avec la réduction ;
  - sans *aussi*, l'acceptabilité décroit avec la réduction

Synthèse sur l'expérience

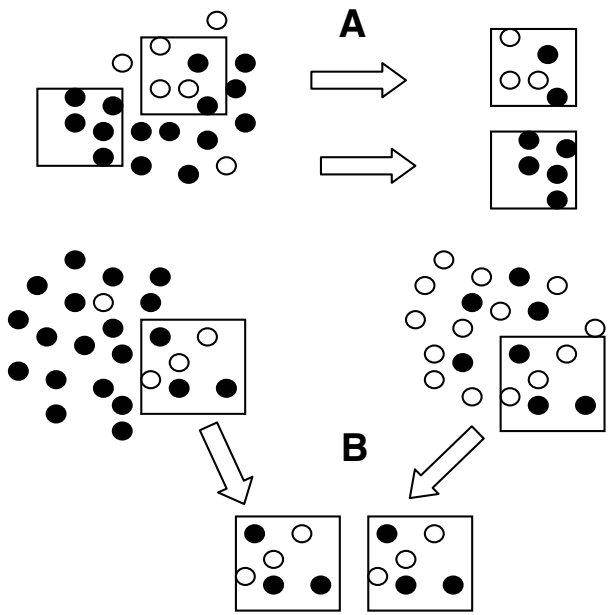
- Observation consolidée et reproductible
- Dépendance vis-à-vis des définitions (*réduction*)
- Portée limitée et espoir de généralisation
- Etude exploratoire : ne permet pas de trancher entre théories concurrentes

Pourquoi des statistiques ?

Toute expérience (ou même simple observation) visant à estimer la différence (éventuelle) entre deux groupes d'individus pour un caractère donné (poids moyen, ou temps de réaction) ne peut pas s'appuyer sur un seul exemplaire de chaque groupe pris au hasard. Dans le domaine du vivant, la grande variabilité des individus oblige à se baser sur des *échantillons* de plusieurs individus (tant mieux s'ils sont nombreux).

Les méfaits des fluctuations d'échantillonnage. **A** : Deux échantillons, même fort différents, ne proviennent pas nécessairement de deux populations différentes ; **B** : Deux échantillons, même fort semblables, ne proviennent pas nécessairement de deux populations semblables.

(Poinsot, 2004)



Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Chemins vers la connaissance linguistique

1. Introspection
   Risque : illusions, biais, manque de reproductibilité
2. Démontage du cerveau
   If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat. (Douglas Adams)

   Difficulté méthodologique : peut-on comprendre ce que vous écrivez dans votre traitement de texte en mesurant les variations d'énergie dans votre ordinateur ?
3. Observation des productions langagières
   - Expérimentalement (production induite, conditions contrôlées)
     Risque : conditions expérimentales non réalistes
   - "Naturellement" (recueil de productions, conditions écologiques)
     $\longrightarrow$ constitution de corpus
     Risque : données en vrac, des nombres et peu de compréhension

Sorbonne Nouvelle

Linguistique et relation aux données

- Linguistique introspective (« de salon », "*armchair linguistics*")
  Les données sont "dans la tête du linguiste"
- Linguistique de corpus
  Les données sont "dans la nature"
- Linguistique expérimentale
  Les données sont "obtenues en laboratoire"

Linguistique et relation aux données

- Linguistique introspective (« de salon », "*armchair linguistics*")
  Les données sont "dans la tête du linguiste"
- Linguistique de corpus
  Les données sont "dans la nature"
- Linguistique expérimentale
  Les données sont "obtenues en laboratoire"

Attention aux caricatures !

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Un vieux débat

The **corpus linguist** has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus the second word of a sentence.

The **armchair linguist** sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact !", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

(Fillmore 1992, cité par McEnery & Wilson (1996))

Un vieux débat

Le **linguiste de corpus** a tous les faits dont il a besoin, sous la forme d'environ un zilliard de mots, et il voit son travail comme consistant à dériver des faits secondaires à partir de ces faits primaires. En ce moment il est occupé à déterminer les fréquences relatives des onze parties du discours dans le premier mot d'une phrase par rapport au deuxième mot.

Le **linguiste de salon** est assis dans un fauteuil profond et confortable, les yeux fermés et les bras croisés derrière la tête. De temps en temps, il ouvre un œil, se relève brutalement et crie : « Ouah ! Quel super fait ! », attrape son crayon, et écrit quelque chose. Il s'agite alors pendant plusieurs heures dans l'excitation de s'être une fois encore approché de la nature véritable du langage.

(Fillmore 1992, cité par McEnery & Wilson (1996))

Sorbonne Nouvelle

Un vieux débat (suite)

• Du linguiste de corpus au linguiste de salon :
Qu'est-ce qui pourrait me convaincre que ce que vous dites est **vrai** ?
• Du linguiste de salon au linguiste de corpus :
Qu'est-ce qui pourrait me convaincre que ce que vous dites est **intéressant** ?

Discussion célèbre :

— Chomsky : The verb *perform* cannot be used with mass word objects : one can *perform a task* but one cannot *perform labour*.
— Hatcher : How do you know, if you don't use a corpus and have not studied the verb *perform* ?
— Chomsky : How do I know ? Because I am a native speaker of the English language !

Exemple attesté :

(9) I really want to know how hobbyist magicians perform magic in the daily life (...)

Liguistique introspective

- tire profit de l'intuition du chercheur (linguiste)
- est immédiatement accessible
- offre des exemples négatifs (contrefactuels)
- permet une étude systématique des variations → linguistique expérimentale
- permet de s'éloigner des influences extra-linguistiques → linguistique expérimentale
- tire profit de l'attention du linguiste → linguistique de corpus

mais...

- est potentiellement sensible au biais de l'expérimentateur
- n'est pas protégée contre l'influence du dialecte/idiolecte

Linguistique de corpus

- oriente vers une linguistique de l'usage
- offre toute la puissance des outils statistiques
- permet d'objectiver les observations
- permet de répliquer les observations/manipulations
- permet de faire une démarche "bottom-up"

mais...

- n'offre pas de données négatives :

  l'absence de preuve n'est pas la preuve de l'absence
- dépend de la capacité de collecter des données en grand nombre

Sorbonne Nouvelle

Linguistique expérimentale

- permet de contrôler les variables ayant une influence
- permet de “purifier” les phénomènes observés
- permet la réplication et la cumulativité en science
- permet d’objectiver les observations

mais...

- coût très élevé : chaque variation de paramètres nécessite de nouvelles expérience
- suppose une théorie faisant des prédictions (“top-down”)
- demande qu’on vérifie le passage du laboratoire au contexte écologique

Un vieux débat (fin)

Les débats sur les différentes approches ont permis de mieux réfléchir à la façon de construire des corpus et de travailler avec. Surtout, le consensus a convergé vers l'idée que les corpus et les intuitions des linguistes sont complémentaires plutôt que contradictoires.

> *"I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another."*
>
> *(Fillmore, 1992; Fort, 2011)*

Sorbonne Nouvelle

Plan

1 Introduction

2 Collecte de données
- Catégories de données
- Modes de collecte

3 Exploitation des données
- Annotation
- Expérimentation

4 Discussion
- Chemin vers la connaissance linguistique
- Un vieux débat

5 Conclusion

Conclusion

Toute investigation scientifique repose sur la collecte de données :

- bien définies
- collectées avec soin
- selon le principe FAIR
- analysées avec des méthodes envisagées dès la conception

Conclusion

Toute investigation scientifique repose sur la collecte de données :

- bien définies
- collectées avec soin
- selon le principe FAIR
- analysées avec des méthodes envisagées dès la conception

Perspectives

- Données nativement numériques
  (mais numérisation toujours nécessaire)
- Outils (TAL & IA) de collecte et d'annotation (pas parfaits)
- Nécessité de conserver contrôle et compréhension des traitements
- C'est le plus souvent la combinaison des méthodes qui produit les meilleurs résultats

Sorbonne Nouvelle

Références

Amsili, Pascal, Ellsiepen, Emilia, & Winterstein, Grégoire. 2016. Optionality in the use of 'too' : The role of reduction and similarity. *Revista da Abralin (Associação Brasileira de Linguística)*, 15(1), 229–252.

Fillmore, Charles. 1992. "Corpus linguistics" or "Computer-aided armchair linguistics". *Pages 35–60 of :* Svartvik, Jan (ed), *Directions in corpus linguistics.* Trends in Linguistics. Studies and Monographs, vol. 65. Berlin and New York : Mouton de Gruyter. Proceedings of the Nobel Symposium 82 Stockholm, 4-8 August 1991.

Fort, Karën. 2011. *Corpus Linguistics : history*. Diapos présentées le 18 novembre 2011. Inist.

Green, Georgia M. 1968. On *too* and *either*, and not just *too* and *either*, either. *Pages 22–39 of : CLS (Chicago Linguistics Society)*, vol. 4.

Ioannidis, John P. A. 2005. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8).

Kaplan, Jeff. 1984. Obligatory *too* in English. *Language*, 60(3), 510–518.

McEnery, Tony, & Wilson, Andrew. 1996. *Corpus Linguistics*. Edinburgh University Press.

Poinsot, Denis. 2004. *Statistiques pour statophobes*. En ligne : `http://perso.univ-rennes1.fr/denis.poinsot`.

Urfist Méditerrannée. 2019. *DoRANum-Enjeux et bénéfices : les principes FAIR*. DoRANum.

Sorbonne Nouvelle