Avancement du TER
– Méthodes expérimentales en linguistique –

Master de Sciences du Langage
« Linguistique Théorique & Expérimentale »
P. Amsili

17 novembre 2015

Plan

Linguistique et empirie

La linguistique a un rapport ancien avec l'empirie*

- collecte de *ressources* : dictionnaires, thésaurus, concordances...
- collecte d'*attestations*
- transcriptions/enregistrements de données de terrain

* *empirie* : *ensemble des données de l'expérience (TLFi)*

Révolution chomskyenne

- Linguistique de l'attestation = linguistique de la **performance**
- Comment atteindre la **compétence** ?
  ⇒ Nouveau type d'« évidence » linguistique : le jugement d'acceptabilité

Concepts fondamentaux

- Grammaticalité : propriété d'un objet linguistique (compatibilité avec la grammaire d'une langue donnée)
- Acceptabilité : propriété d'un stimulus linguistique tel qu'il est perçu par un locuteur
- Jugement d'acceptabilité : réponse (comportementale) d'un locuteur à la demande d'un linguiste

Bard *et al.* (1996)

Catégoricité vs gradabilité

- Acceptabilité relative : différence d'acceptabilité entre deux stimuli
  admise par tous les linguistes : * **? #?? ...
- Grammaticalité relative : différence de grammaticalité inhérente entre deux stimuli.
  controversée : l'existence d'un gradient de grammaticalité n'est pas compatible avec la plupart des théories syntaxiques actuelles

« the possibility remains that acceptability is graded because grammaticality is » (Bard *et al.* , 1996, p.33)

Problèmes avec les jugements

- reflet indirect et infidèle (?) de la grammaticalité
- influence de facteurs non pertinents :
  - prise en compte de la fréquence d'usage (estimée)
  - conformité à une norme prescriptive
  - conformité à un registre évalué socialement
  - degré de plausibilité sémantique/pragmatique
- biais de l'expérimentateur Labov (1972)

Plan

Méthodes expérimentales

- Visant à mieux contrôler les jugements d'acceptabilité
  - multiplication des sujets (banissement de l'introspection)
  - questionnaires d'acceptabilité avec échelles
  - tâches de production induite
- Visant à réduire le caractère méta-linguistique de la tâche
  $\Rightarrow$ plus indirect mais moins biaisé
  - temps de traitement (lecture auto-segmentée, mesures oculométriques...)
  - temps de réaction (tâches de décision lexicale, de déomination)
  - imagerie (ERP, IRMf...)
- Visant à introduire de nouvelles « évidences »
  - tests de perception (phonétique par exemple)
  - enregistrements psychophysiques (laryngoscopie, glottoscopie, imagerie cérébrale...)
  - enregistrements comportementaux (eye-tracking, mouse-tracking, tâches de compréhension, map task...)
  - importation de l'arsenal expérimental des sciences cognitives

Retour de la performance

- la linguistique de l'attesté a été rejetée dans le paradigme chomskyen, un des arguments étant le manque de données négatives
- les méthodes statistiques modernes répondent (en partie) à cet contre-argument :
  - l'étude statistique des **distributions** et des **correlations** dans
  - un échantillon **représentatif** permet de tirer des conclusions
  - même sur des phénomènes **rares** voire **non attestés**

$\Rightarrow$ Linguistique quantitative sur corpus (Gries, 2013)

Plan

Crowdsourcing

Cas particulier d'une méthode expérimentale : utilisation des techniques de *crowdsourcing* (externalisation participative/ouverte) pour la constitution de ressources linguistiques ou la réalisation d'expériences.

voir les tranparents qui suivent, qui proviennent d'une présentation par l'un des auteurs du papier (Munro *et al.* , 2010).

`http://web.stanford.edu/~rmelnick/files/SchnoebelenEtAlLSA2011.pdf`

Pourquoi des expériences d'acceptabilité

« L'utilisation informelle de jugements intuitifs pose un grave problème pour la justification empirique de revendications théoriques en linguistique »

(Schütze, 1996)

- manque de précision
- manque de reproductibilité
- manque de fiabilité (cohérence inter et intra-sujets)

Ce qu'il faut :

- test d'instances lexicales multiples d'une structure donnée
- test de plusieurs participants naïfs
  (ie non linguistes et non informés de l'objet du test)
- utilisation d'outils statistiques pour tester les hypothèses vis-à-vis du résultat produit.

Mesures des jugements d'acceptabilité

- jugement binaire
- echelle de valeurs
- estimation de grandeur (*magnitude estimation*)
- justement thermomètre

Autour du jugement d'acceptabilité

- Au moins 4 points de données par condition
- Prise en compte des effets d'ordre (présentation en carré latin)
- Utilisation de distracteurs (*fillers*)
  pour éviter que le participant repère des régularités
  pour éviter les effets plafond/plancher
- Utilisation de questions de compréhension
- Mesure de significativité : qu'est-ce qui explique la variance ?
  (résultat binaire : un effet/pas d'effet)
- Mesures supplémentaires possibles (η^2, estimation de la taille de l'effet = proportion de la variance expliquée par le facteur)

Jugement binaire

Caractérisation des stimuli (mots, syntagmes, phrases...) comme

- bon / mauvais
- acceptable / inacceptable
- naturel / bizarre

- jugement absolu (point de référence personnel implicite)
- effet de choix forcé
- suffisant dans de nombreux cas (ex. ordre des mots)

Jugement sur une échelle

Echelle à 5 valeurs, 7-valeurs (Lickert-scale), 10, 100...

- reste un jugement absolu (point de référence personnel implicite)
- nécessite une normalisation (z-transformation)
- biais possibles liés aux effets culturels de l'échelle (notes scolaires)
- si l'échelle est impaire, il y a un point neutre ; si l'echelle est impaire, on parle de choix forcé.
- Interprétation des niveaux intermédiaires : incertitude ou vraie gradation ?
- *central tendency bias*
- influence possible de la présentation graphique
- pas de comparaison de la force relative des violations

Magnitude estimation

Mesure d'origine psychophysique (intensité d'un stimulus perceptuel (douleur, luminosité...)), appliquée aux sciences cognitives en général et à la linguistique (Bard *et al.* , 1996).
Principe : les participants jugent l'intensité perçue d'un premier stimulus (*modulus*), et ensuite jugent l'intensité d'un second stimulus relativement au premier stimulus.
Le jugement est exprimé sur une échelle continue (par exemple en dessinant des lignes de longueur différentes, ou avec un curseur).

- jugements relatifs et non absolus
- finesse de la mesure (pas de graduations)
- réduction des effets "choix forcé"

Critique de ME

- Problème d'interprétation des mesures pour les participants et les expérimentateurs : les propriétés arithmétiques des mesures ME ne sont pas garanties (Sprouse, 2011)
- Variabilité très supérieure aux mesures par likert-scale
- Il est difficile en linguistique de faire jouer à tous les stimuli le rôle de modulus

Etude critique de Weskott & Fanselow (2011) :

- La variabilité supérieure ne correspond pas à une gradiance supplémentaire
- La comparaison de méthodes binaires/likert/ME, dans une expérience de jugement d'ordre des mots en allemand
  - donne les mêmes résultats primaires (effet des conditions)
  - ne distingue pas les méthodes sur la quantité de variance expliquée.

Jugement thermomètre

Variante de la mesure d'estimation de grandeur, avec :

- deux points de référence (20 et 30, par exemple)
- pas d'instruction du type "deux fois moins acceptable"...
- les jugements restent relatifs, à granularité fine, l'échelle reste ouverte

(Featherston, 2008)

- Intuition : le contraste plus fort si le *comment* est réduit :

(30) 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.

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

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

- Design
  - Expérience d'acceptabilité (AJT), sur Internet (iBexFarm). 80 sujets.
  - 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 comment

(32) 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-`

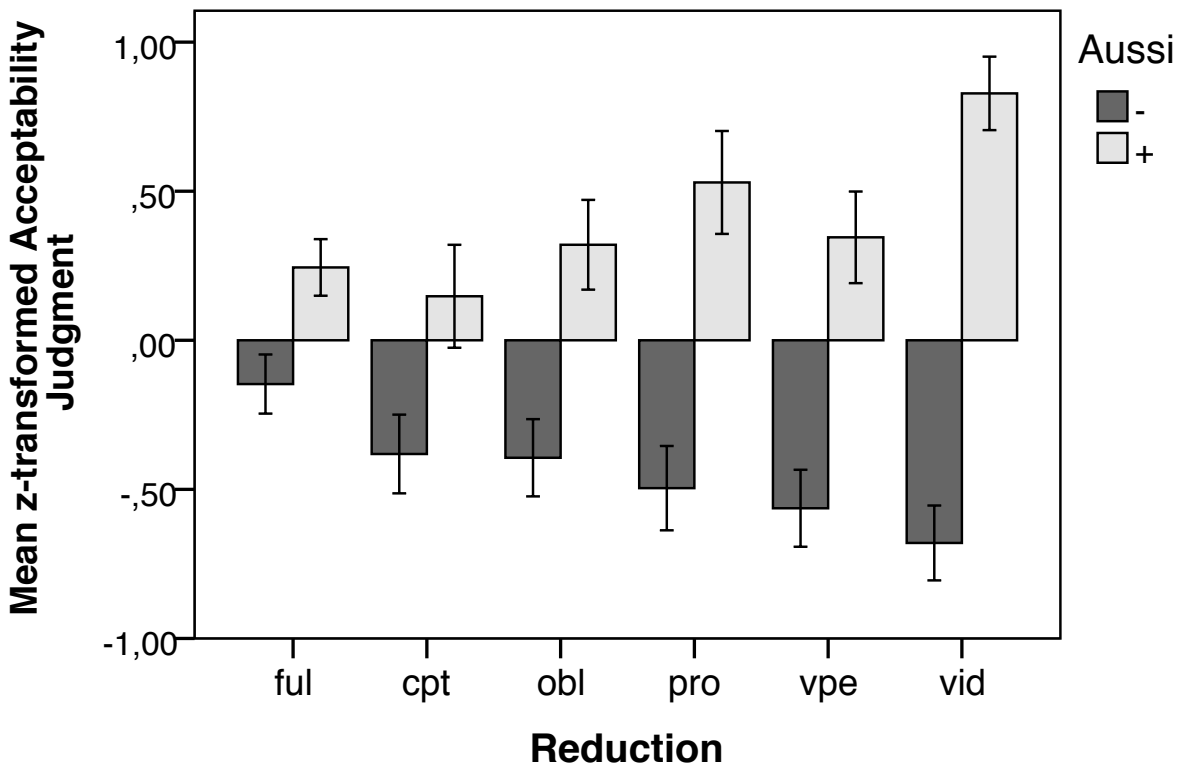
Résultats attendus

- Résultats attendus

`ful+`	*not so good, because of repetition*
`ful-`	*idem*
`cpt+` `cpt-` `pro+` `pro-` `vpe+` `vpe-`	*bigger and bigger contrast between + and −*
`vid+`	*highest acceptability*
`vid-`	*lowest acceptability*

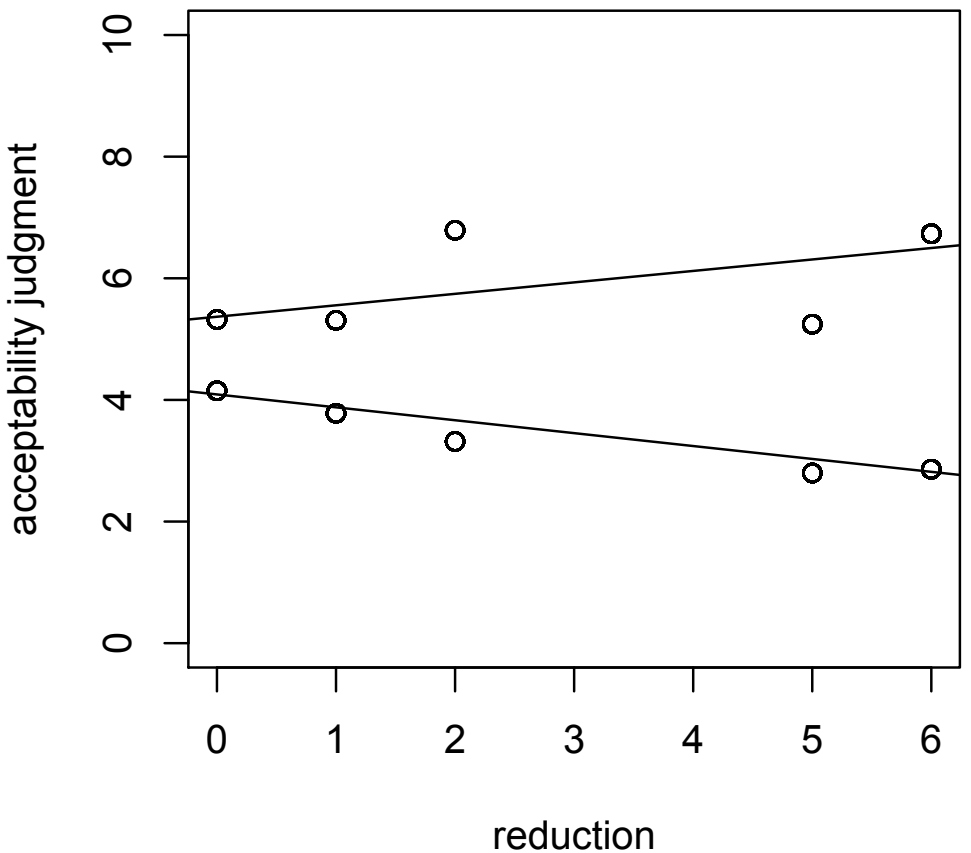
Résultats (I)

⇒ Patron des réponses conforme aux attentes :



Résultats (II)

⇒ Mesure de la corrélation : significative



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$<$ .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$<$ .001) :
  - Avec *aussi*, l'acceptabilité croît avec la réduction ;
  - sans *aussi*, l'acceptabilité décroit avec la réduction

References

Bard, Ellen Gurman, Robertson, Dan, & Sorace, Antonella. 1996. Magnitude estimation of linguistic acceptability. *Language*, **72**(1), 32–68.

Featherston, Sam. 2008. Thermometer judgements as linguistic evidence. *In :* Riehl, Claudia Maria, & Rothe, Astrid (eds), *Was ist linguistische Evidenz ?* Aachen : Shaker Verlag.

Gries, Stefan Th. 2013. *Statistics for linguistics with R : a practical introduction*. Walter de Gruyter.

Labov, William. 1972. Some principles of linguistic methodology. *Language in society*, 1(01), 97–120.

Munro, Robert, Bethard, Steven, Kuperman, Victor, Lai, Vicky Tzuyin, Melnick, Robin, Potts, Christopher, Schnoebelen, Tyler, & Tily, Harry. 2010. Crowdsourcing and language studies : the new generation of linguistic data. *Pages 122–130 of : Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics.

Schütze, Carson T. 1996. *The Empirical Base of Linguistics*. U. Chicago Press. Chap. Definitions and Historical Background, pages 19–53.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation : Commutativity does not hold for acceptability judgments. *Language*, **87**(2), 274–288.

Weskott, Thomas, & Fanselow, Gisbert. 2011. On the informativity of different measures of linguistic acceptability. *Language*, **87**(2), 249–273.

Carré latin

Un carré latin est un carré de côté n (donc avec $n \times n$ cases) comprenant n symboles différents de sorte que chaque symbole apparaisse une fois et une seule dans chaque ligne et dans chaque colonne.

a	b
b	a

a	b	c
b	c	a
c	a	b

Exemple d'utilisation : une ligne par participant, les colonnes représentant l'ordre de présentation des conditions/items.