Vers une grammaire probabiliste de microsystèmes fonctionnels en L2

Cyrielle Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu-Li, Thomas Gaillat

Vers une grammaire probabiliste de microsystèmes fonctionnels en L2

Cyriel Mallart[1], Andrew Simpkin[2], Rémi Venant[3], Nicolas Ballier[4], Bernardo Stearns[5], Jen Yu Li[1], Thomas Gaillat[1]

[1]LIDILE, Université Rennes 2
[2]School of Mathematics, Statistics and Applied Mathematics,University of Galway
[3]LIUM, Université du Mans
[4]CLILLAC-ARP, Université Paris Cité
[5]Insight, Data Science Institute, University of Galway

Motivation

Contexte

Apprendre une L2 implique l'**émission d'hypothèses** sur les correspondances formes-fonctions : (1) quelles formes sont utilisées pour réaliser quelles fonctions dans la L2 et (2) quelle importance accorder à l'utilisation de formes individuelles dans la réalisation de fonctions spécifiques. [3, p. 375]

Effet global : Modification du système interne de la L2 et stabilisation progressive des correspondances.

Effets locaux : Des microsystèmes d'apprenants

1. Hésitations entre des formes exprimant une même fonction (référence, détermination, quantification etc.)
2. Instabilité : regroupements de formes inattendues mais de paradigmes fonctionnels différents [6]
3. Nature transitoire des microsystèmes [5]: effacement progressif

Enjeux et difficultés

- Difficulté de capturer les variations entre formes d'une même fonction
- Difficulté d'attribuer ces variations à des niveaux de compétence
- Mesures traditionnelles de complexité insuffisantes

Objectifs and Questions de recherche

- Hypothèse: Association entre les variations de formes-fonctions et les niveaux de compétence des apprenants. Cela pose deux questions de recherche :

1. **QR 1 :** Comment saisir les variations entre formes associées à la même fonction linguistique ?
2. **QR 2 :** Ces variations sont-elle corrélées à la compétence opérationnalisée par le CECR ?

Application : THIS, THAT et IT proformes.

Données

EF Cambridge Open Language Database (**EFCAMDAT**) learner-English corpus [4]:

- 1,180,507 textes, 191 pays
- 16 niveaux de compétence classés selon les 6 niveaux CECR

CECR	Ecrits	Moyenne nb tokens	Écart Type
A1	626,005	39.32	21.46
A2	308,014	68.82	24.42
B1	168,473	98.88	30.23
B2	61,366	137.27	43.67
C1	14,709	171.13	49.03
C2	1,940	176.98	71.95

Table 1. Statistiques descriptives du corpus EFCAMDAT en fonction des niveaux CECR

Exemple

That is how I found the class of Sciences of Education in Paris 2 . I went to the global opening and when I was listening to the presentation of the classes , I was sure **this** was what I wanted to study for my future .

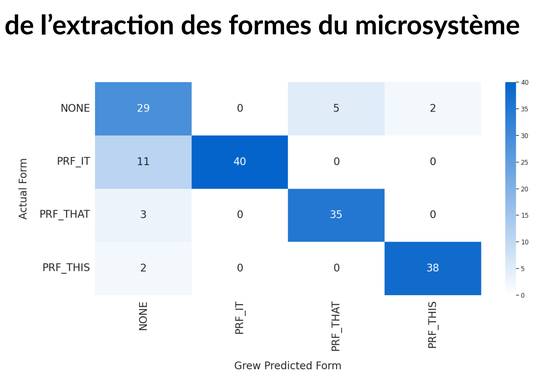
Requêtes par motifs avec Grew

Requêtes par motifs [1] fondées sur les traits linguistiques du schéma d'annotation *Universal Dependency* [2].

Exemple: THIS pattern

```
THIS-PRF::DEP[wordform="this"
|"these"|"This" |"These"]; GOV-
[nsubj|obl|nsubj:pass|nmod
|obj|nsubj:outer|conj|root] -> DEP
```

Évaluation de l'extraction des formes du microsystème



Matrice de confusion de l'évaluation de l'extraction Grew. Évaluation effectuée sur 165 proformes IT, THIS and THAT annotées manuellement par 3 experts

Références

[1] Maxime Amblard, Bruno Guillaume, Siyana Pavlova, and Guy Perrier. Graph querying for semantic annotations. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 95–101. European Language Resources Association, 2022.

[2] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, (2):255–308, 2021.

[3] Rod Ellis. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, United Kingdom, 1994.

[4] Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Miguel, A. Tseng, A. Tuninetti, and D Walter, editors, *Proceedings of the 31st Second Language Research Forum*, Carnegie Mellon, 2013. Cascadilla Press.

[5] Yves Gentilhomme. Microsystèmes et acquisition des langues. *Encrages*, (Numéro spécial):79–84, 1980.

[6] Bernard Py. Quelques réflexions sur la notion d'interlangue. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 1:31–54, 1980.

Trois modèles de mesure des variations du microsystème des proformes

- **Proportions relatives** - Collecte des fréquences brutes des proformes dans chaque texte.
- **Probabilité des proformes** - Modélisation des proformes en fonction des traits linguistiques du contexte proche. Probabilités d'occurrence des proformes masquées dans leur contextes.
  - Régression logistique sur trois niveaux CECR
  - Régression logistique sur niveau C1

Évaluation des mesures

Méthode

1. **Données** d'entraînement (N=119 904) et de test (N=22 916)
2. **Création des mesures** avec les trois modèles
3. **Évaluation statistique** de l'association entre les mesures et les niveaux du CECR avec **une régression multinomiale ordinale**

Performance des 3 modèles pour la création des mesures

- **Proportions :** *fréquence, aucun indicateur statistique*
- **Régression logistique multinomiale :** 0.62 accuracy (95% CI: (0.608, 0.635), $p < .001$)
- **Régression logistique multinomiale C1 :** 0.55 accuracy (95% CI: (0.56, 0.59), $p < .001$)

Résultats et discussion

QR 1 : Évaluation des mesures de probabilités d'occurrence

Capture de la tendance d'occurrence de chaque proforme en fonction des contextes linguistiques.

	Régression logistique multinomiale			Régression Logistique multinomiale - niveau C1		
	IT	THIS	THAT	IT	THIS	THAT
Balanced accuracy	0.72	0.68	0.70	0.68	0.61	0.67
Précision	0.64	0.57	0.62	0.59	0.42	0.60
Rappel	0.74	0.49	0.56	0.63	0.43	0.54

Table 2. Résultats des approches par prediction pour la mesure des probabilités d'occurrence du microsystème des proformes

QR 2 : Évaluation de l'association entre les mesures et les niveaux du cadre

Mesures fondées sur des probabilités **peuvent** être utilisées comme prédicteurs des niveaux CECR.

		Odds ratio	95% CI	p_value
Proportions	IT	0.996	0.996, 0.997	<0.001
	THIS	0.998	0.998, 0.999	<0.001
	THAT	1.009	1.009, 1.001	<0.001
Régression logistique ordinale	IT	0.985	0.984, 0.986	<0.001
	THIS	1.008	1.006, 1.009	<0.001
	THAT	1.02	1.019, 1.021	<0.001
Régression logistique ordinale : différence freq réelle - pred C	ITdiff	1.09	1.07, 1.11	<0.001
	THISdiff	0.93	0.91, 0.96	<0.001
	THATdiff	0.93	0.91, 0.95	<0.001

Table 3. Régression logistique ordinale du niveau CECR en fonction des mesures de IT, THIS et THAT (exprimées en *Odds ratios*)

Utilisation globale du microsystème

- Plus la probabilité des IT est forte, moins les niveaux CECR ont tendance à être élevés (cf. Figure 1)
- Plus la probabilité des THAT est forte, plus les niveaux CECR ont tendance à être élevés (cf. Figure 1)
- Plus la probabilité des THIS est forte, plus les niveaux du CECR ont tendance à être élevés (cf. Figure 1)

Sous-utilisation et sur-utilisation par rapport au niveau C

- Plus IT est sur-utilisé, plus le niveau CECR augmente
- Plus THAT est sur-utilisé, plus le niveau CECR baisse
- Plus THIS est sur-utilisé, plus le niveau CECR baisse

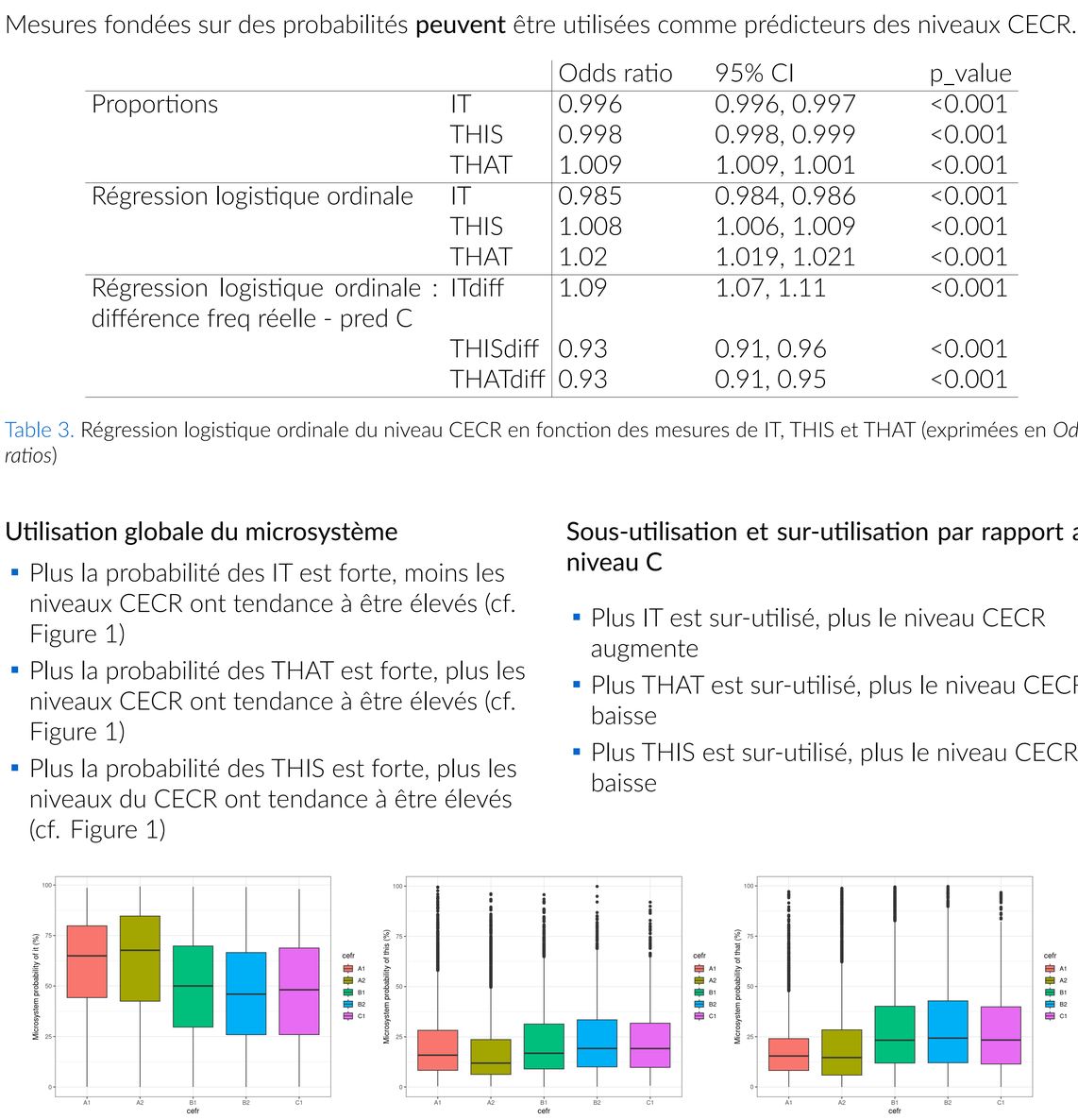


Figure 1. Probabilités de IT, THIS, THAT respectivement

Perspectives

- Autres microsystèmes : relatifs, déterminants, quantifieurs, durée, multi-noms
- Création d'un système de *learning analytics* : visualisations des tendances grammaticales en L2

thomas.gaillat@univ-rennes2.fr