Exploring linguistic complexity in two varieties of learner English: ESP and applied languages

Thomas Gaillat, Sophie Belan

▶ **To cite this version:**

Thomas Gaillat, Sophie Belan. Exploring linguistic complexity in two varieties of learner English: ESP and applied languages. Les problématiques de la spécialisation des langues en LEA, GERAS - Université de Nantes, Dec 2020, Nantes, France. hal-03337126

HAL Id: hal-03337126
https://hal.science/hal-03337126

Submitted on 7 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GERAS 2020 December 4, 2020

Exploring linguistic complexity in two varieties of learner English: ESP and applied languages

Sophie Belan
EA 1162 CRINI
UNIVERSITÉ DE NANTES

Thomas Gaillat
EA 3874 LIDILE
UNIVERSITÉ RENNES 2

Introduction

Complexity is one the three constructs that can be used for L2 proficiency and performance

Linguistic complexity is an objective given, independent from the learner, which refers to the intrinsic **formal or semantic-functional properties of L2 elements** (e.g. forms, meanings, and form-meaning mappings) or to properties of (sub-)systems of L2 elements. (Housen et al 2012)

Introduction

Operationalising the construct in terms of **specific metrics**

> it is possible to provide **points of comparisons** that can inform on possible differences or similarities of two types of L2

Research Question

- Are there differences between Applied Languages (LEA) and ESP English learner writings in terms of linguistic complexity?

Outline

1. Previous work
2. Method
    1. Corpus
    2. Annotation and metrics
    3. Processing and data set
    4. Experimental setup
3. Results
4. Discussion and next steps

Previous work

- Linguistic complexity is a construct part of the CAF triad
  - (Housen, Kuiken, and Vedder 2012; Norris and Ortega 2009)
- Used in many tasks as criterial features of proficiency in L2
  - Mostly Lexical and syntactic diversity
    - (Bulté and Roothooft 2020; Volodina, Pilán, and Alfter 2016; Yannakoudakis, Briscoe, and Medlock 2011, (Pavageau 2009, McAllister and Belan 2014)
- Readability used to identify level of difficulty in texts
  - (Xia, Kochmar, and Briscoe 2016)
- Some tools for complexity measures:
  - Vizling (Gaillat et al, 2020),
  - Coh-metrix (McNamara et al. 2010)
  - L2SCA & LCA (Lu 2014)
  - TAALES (Kyle, Crossley, and Berger 2018)

Method

Two corpora

- CELVA.Sp (University of Rennes 1). L2 Corpus of English for Specific Purposes
  - Pharmacy, Computer Science, Biology, Medicine
  - Task: writing a description and opinion on a discovery/technology
  - Time: 45 minutes
- LEA (University of Nantes). L1 Applied Foreign Languages (LEA) students
  - Domains: Applied Languages with Business
  - Task: writing a short newspaper article on the problem of counterfeiting
  - Time: 60 minutes

Subsets	Number of writings	Median # words	Standard Deviation	CEFR
CELVA.Sp	**42**	**231**	**147**	**B2**
LEA	**50**	**267**	**67.83**	**B2**

Annotation & metrics

Annotation and metrics tools

- Stanford CoreNLP (POS tagging + parsing) (Manning et al. 2014)
- L2SCA (Lu 2010)
- R Quanteda library: texstat (lexdiv & readability) (Benoit 2018)

Metrics

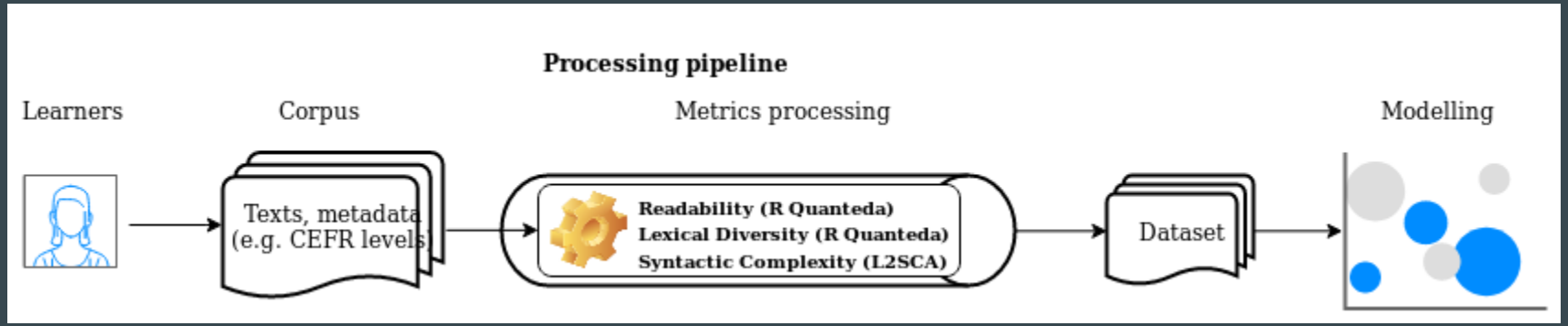
- Syntactic e.g. amount of coordination, subordination
- Lexical diversity e.g. density, sophistication
- Readability (level of difficulty of a text)

Processing pipeline and Data set

Features

- 83 metrics for each text
- CEFR levels

Pipeline



Experimental setup

Analyse statistique

H0 - There is no difference between pairs of metrics in the two independent samples

H1 - There is a difference between pairs of metrics between the two samples

Comparing group means or medians

- T-test : Welsh correction for variance difference (in case of normal distribution)
- Wilcoxon/Mann-Whitney test (in case of non normal distribution)

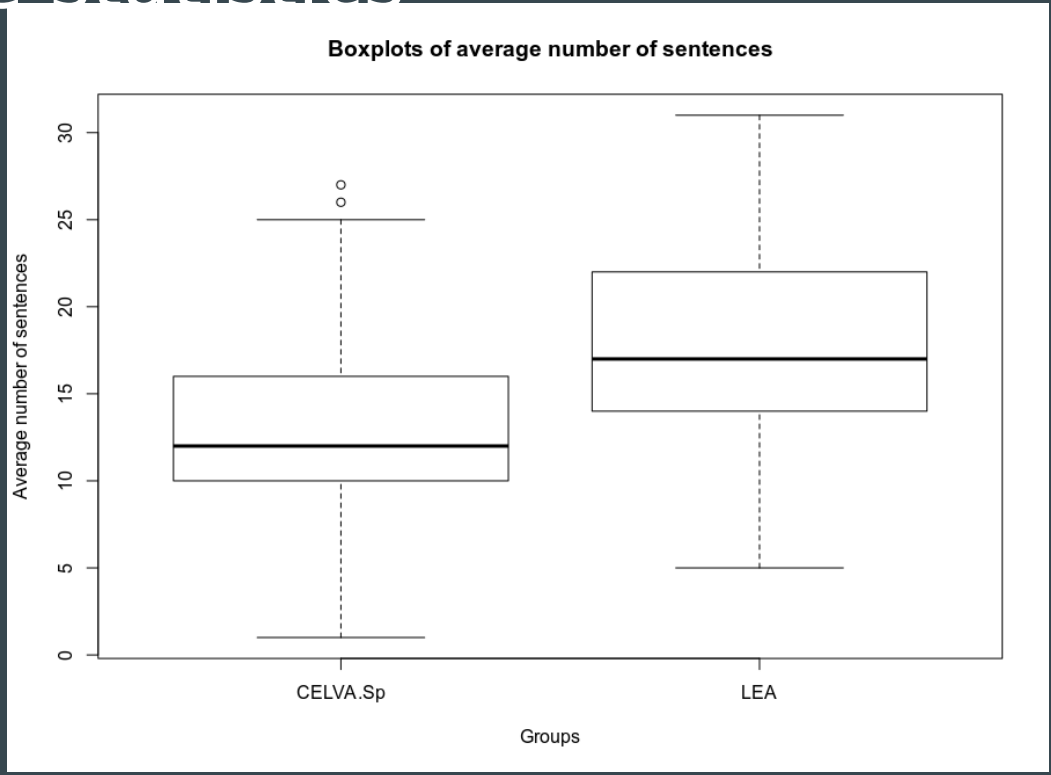
Experimental setup

1. Verify distribution normality for each metric in each sample
2. Apply corresponding test with p.value

Results

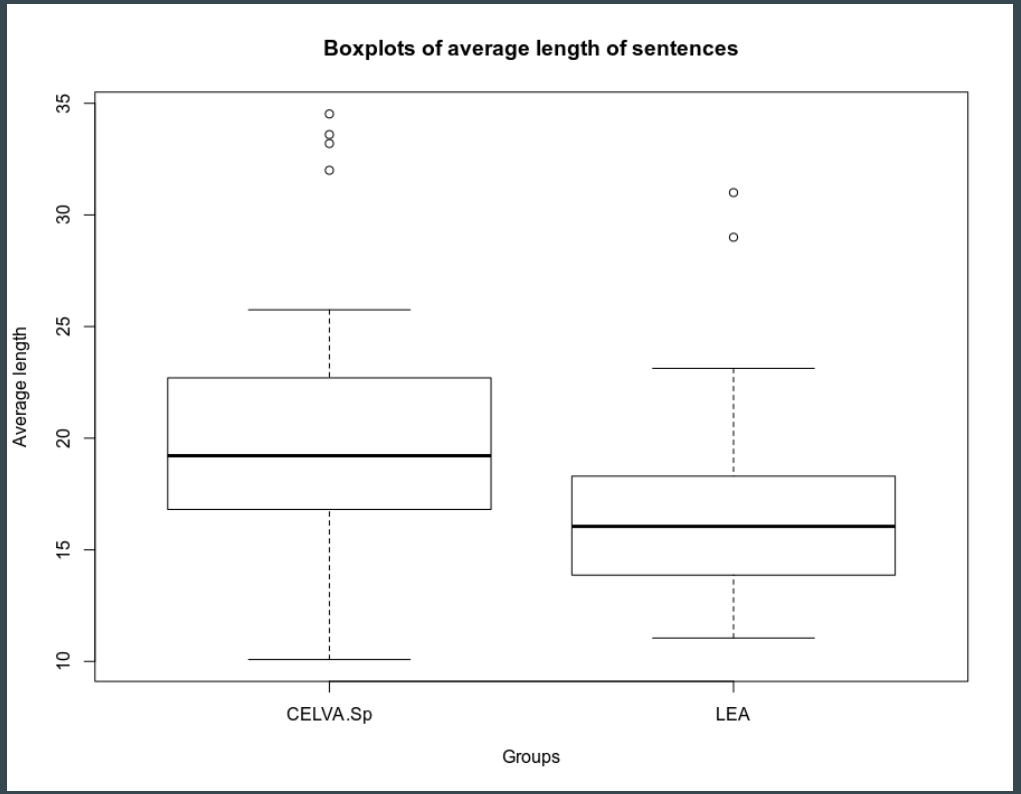
Results: descriptive statistics

```
summary(CELVA_SP_B2$S.1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   10.00   12.00   13.10   15.75   27.00 


summary(lea$S.1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.00   14.25   17.00   17.72   21.75   31.00 
```



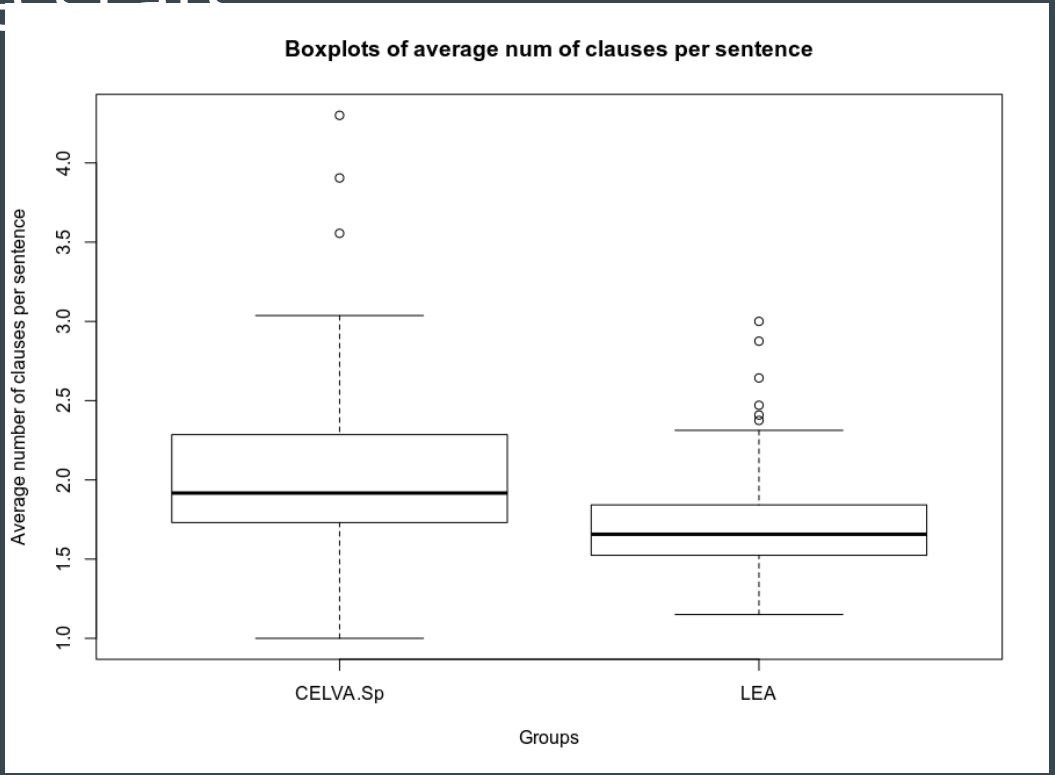
Results: descriptive statistics

```
> summary(CELVA_SP_B2$MLS)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 

  10.08   16.81   19.21   20.46   22.70   34.52 

> summary(lea$MLS)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 

  11.05   13.88   16.05   16.66   18.21   31.00 
```



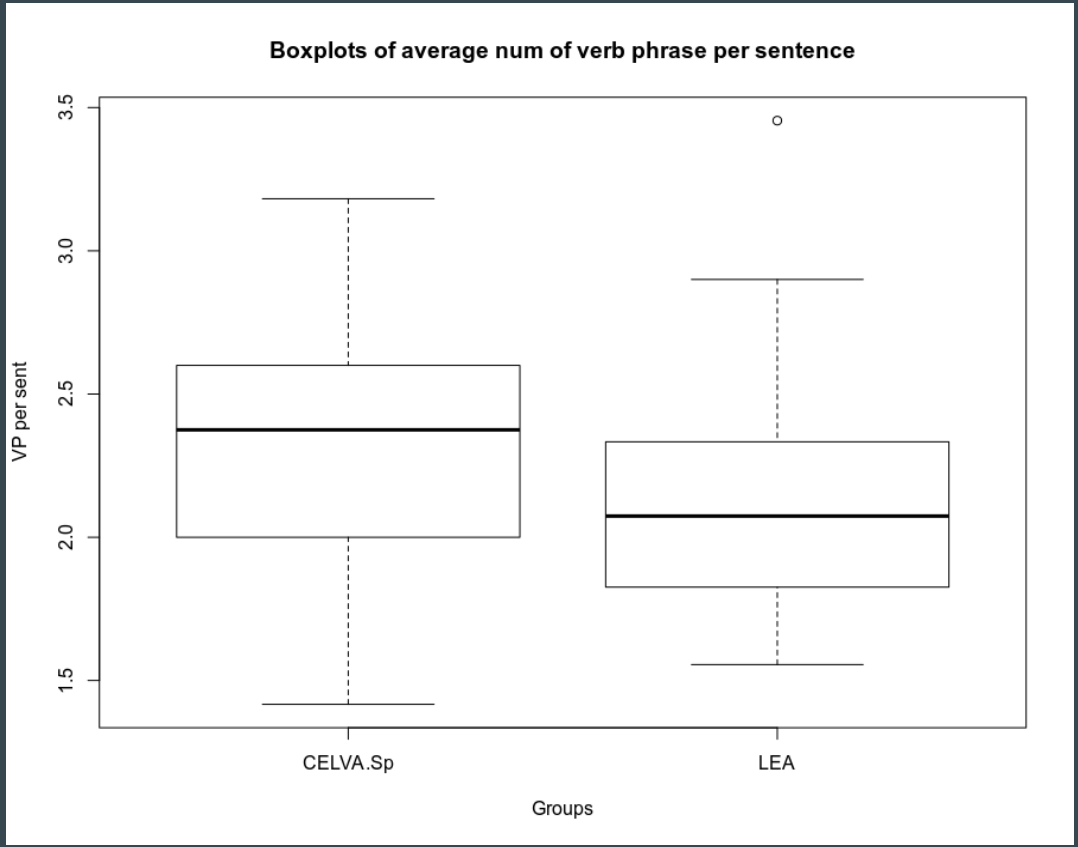
Results: descriptive stats

```
summary(CELVA_SP_B2$C.S)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 

  1.000   1.731   1.917   2.077   2.286   4.300 

> summary(lea$C.S)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 

  1.150   1.524   1.657   1.778   1.838   3.000 
```



Results: descriptive statistics

```
summary(CELVA_SP_B2$VP.T)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  1.417   2.000   2.375   2.302   2.600   3.182

> summary(lea$VP.T)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  1.556   1.830   2.074   2.127   2.317   3.454
```



Results: Testing significance of differences

Metrics that are significantly different : H0 rejected $p < 0.05$

Three examples	statistic	**p.value**	conf.int_min	conf.int_max	Metric type	Ling scope
Sentences	-3.641	**0**	-7.151	-2.098	Syntactic complexity	Text.size.count(S)
Clause per Sentence	1447.5	**0.002**	0.119	0.452	Syntactic complexity	Sentence.taxis.rate(Clause)
Verb Phrase per sentence	1310.5	**0.041**	0.015	0.401	Syntactic complexity	Sentence.verbal.rate(VP)

Download all metrics here: Wilcox test results @ https://bit.ly/geras_wilcox

Results & discussion

- Tendency among CELVA.Sp writings to complexify sentences in CELVA.Sp.
  - Longer sentences
  - More verbs and clauses per sentence

Results: Testing significance of differences

Differences found in readability metrics based on

- Combination of sentence size, word, sophistication, morphology (syllables)

Syntactic complexity metrics

- Text constituents,.e.g Clause and phrase counts
- Sentence, e.g clauses, verb phrases and complex nominals

Few differences in terms of lexical diversity

- Word repetitions

Discussion

Reasons why these differences

- Different tasks

- Readability metrics involve several variables : exploration of individual roles

Next steps

- More metrics
- Larger corpus
- More CEFR levels

Thank you

thomas.gaillat@univ-rennes2.fr

sophie.belan@univ-nantes.fr

References

Bulté, Bram, and Hanne Roothooft. 2020. “Investigating the Interrelationship between Rated L2 Proficiency and Linguistic Complexity in L2 Speech.” System 91:102246. doi: 10.1016/j.system.2020.102246.

Housen, Alex, Folkert Kuiken, and Ineke Vedder, eds. 2012. Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA. Vol. 32. Amsterdam, The Netherlands, USA: John Benjamins Publishing Company.

Kyle, Kristopher, Scott Crossley, and Cynthia Berger. 2018. “The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0.” Behavior Research Methods 50(3):1030–46. doi: 10.3758/s13428-017-0924-4.

Lu, Xiaofei. 2014. Computational Methods for Corpus Annotation and Analysis. Dordrecht: Springer.

McAllister, Julie. 2009. “Evaluation d’un Dispositif Hybride d’apprentissage d’anglais En Milieu Universitaire : Potentialités et Enjeux Pour l’acquisition d’une Langue Seconde.” These en préparation, Nantes.

McAllister, Julie. 2013. Evaluation d’un dispositif hybride d’apprentissage de l’anglais en milieu universitaire. Potentialités et enjeux pour l’acquisition d’une L2. Thèse de doctorat. Université de Nantes.

McNamara, Danielle S., Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. “Coh-Metrix: Capturing Linguistic Features of Cohesion.” Discourse Processes 47(4):292–330. doi: 10.1080/01638530902959943.

Pilán, Ildikó, and Elena Volodina. 2018. “Investigating the Importance of Linguistic Complexity Features across Different Datasets Related to Language Learning.” Pp. 49–58 in Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing. Santa Fe, New-Mexico: Association for Computational Linguistics.

Thomas, Gaillat, Anas Knefati, and Antoine Lafontaine. 2019. VizLing. https://lidile.hypotheses.org/projet-vizling

Xia, Menglin, Ekaterina Kochmar, and Ted Briscoe. 2016. “Text Readability Assessment for Second Language Learners.” Pp. 12–22 in Proceedings of the 11th Workshop on Innovative Use of NLP for Building

Many thanks to:

Xiaofei Lu

Detmar Meurers

Syntactic complexity metrics

meanSentenceLength meanWordSyllables W S VP C T DC CT CP CN MLS MLT MLC C/S VP/T C/T DC/C DC/T T/S CT/T CP/T CP/C CN/T CN/C

Readability metrics

ARI ARI.simple Bormuth Bormuth.GP Coleman Coleman.C2 Coleman.Liau Coleman.Liau.grade Coleman.Liau.short Dale.Chall Dale.Chall.old Dale.Chall.PSK Danielson.Bryan Danielson.Bryan.2 Dickes.Steiwer DRP ELF Farr.Jenkins.Paterson Flesch Flesch.PSK Flesch.Kincaid FOG FOG.PSK FOG.NRI FORCAST FORCAST.RGL Fucks Linsear.Write LIW nWS nWS.2 nWS.3 nWS.4 RIX Scrabble SMOG SMOG.C SMOG.simple SMOG.de Spache Spache.old Strain Traenkle.Bailer Traenkle.Bailer.2 Wheeler.Smith

Lexical diversity metrics

TTR C.x R CTTR U S.x K D Vm Maas (a, log V0 log eV0)

https://www.rdocumentation.org/packages/quanteda/versions/0.9.7-17/topics/lexdiv

https://quanteda.io/reference/textstat_lexdiv.html