Exploring linguistic complexity in learner English applied to business

Thomas Gaillat, Sophie Belan, Julie McAllister

Exploring linguistic complexity in learner English applied to business

Thomas Gaillat, Sophie Belan and Julie McAllister-Pavageau
University of Rennes 2 (France) | University of Nantes (France) | New Sorbonne University Paris 3 (France)

Linguistic complexity (Bulté & Housen, 2012) is one of the main dimensions used for the characterisation of an L2. It is a multidimensional construct that can be exploited in different tasks such as L2 proficiency-level identification (Ballier et al., 2020; Kyle, 2016), L2 development (McAllister & Belan, 2014) or L1 identification (Xiaofei & Haiyang, 2015). The task of L2 variety identification can also benefit from the use of the construct. By operationalising the construct with syntactic and lexical metrics it is possible to provide points of reference for L2 business English.

This study is motivated by an earlier investigation into L2 development (McAllister, 2013; McAllister & Belan, 2014) of first year Business English undergraduate students at the University of Nantes in France, which showed statistically significant positive results in learner outcomes for fluency and accuracy measures, but not for complexity measures. The initial study involved a pre-test and post-test in the form of a writing task as part of a large-scale research project concerning a task-based blended language learning programme. Further research by the team (Starkey-Perret et al., 2015, 2017) explored the effectiveness of pre-task form-focussed activities in a virtual resource centre (VRC) environment in fostering accuracy in students' written productions. The present study aims to build on this body of research by exploring the complexity dimension further through conducting a comparative corpus-based analysis of L2 Business English students' writing with authentic texts of the same genre, i.e. articles from the British National Corpus (Burnard, 2007).

Two main goals are pursued: identification of quantitative differences between the two groups and the specificity of Business English in terms of complexity. Our approach consists in analysing and visualising a group of L2 writings, classified as B2, in comparison with a control group of writings from native speakers extracted from the British National Corpus. We use a sample of 50 texts from the original corpus written in English by French 1st-year Business English students and a random sample of 50 newspaper articles categorised in the *commerce* domain of the BNC.

Syntactic complexity is operationalised with fourteen metrics. These metrics are grouped in five different types (Lu, 2014): length of production unit (e.g. sentence), sentence complexity, subordination, coordination and particular structures (e.g.complex nominals).

Readability is operationalised with forty-eight metrics. They are based on the morphological features of words used to compute different indicator values. This includes indicators such as the Coleman Liau, the Dale Chall readability score and the Flesch Kincaid grade. They all rely on word length in terms of characters and syllables as well as predetermined lists of words judged as difficult[1].

Lexical richness is operationalised with thirteen metrics. Two types of lexical diversity are included. Diversity based on word type variation is accounted for with TTR based formulae. Diversity based on type repetition is accounted for with Yule's K and similar formulae in which the frequency of word types in a sample of size *n* is relative to the total number of words in a text[2]. We acknowledge that lexical sophistication and lexical density (content vs grammar words) are not taken into account.

[1] For a detailed description of the formulae refer to https://quanteda.io/reference/textstat_readability.html
[2] For the formulae see \url{https://quanteda.io/reference/textstat_lexdiv.html

The syntactic and lexical complexity and readability metrics are computed with a pipeline program, called VizLing[3] (Gaillat et al. , 2021) composed of CoreNLP (Manning et al., 2014), L2SCA (Lu, 2010) and Quanteda (Benoit et al., 2018).

As the purpose is to explore differences between value pairs in the two corpora, we apply pairwise t-tests. The results inform the researcher on the pairs of complexity indices that do or do not show significant differences. They highlight which sub-domains of complexity tend to distinguish Non-Native Speakers from Native Speakers. The identification of differences can help describe specific features of Business English to learners and foster their writing development. This can pave the way for the integration of a corpus-based approach in English for Business Purposes/Business English (EBP/BE) courses.

References

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., & Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, *6*(1), 72–103.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Bulté, B., & Housen, A. (2012). *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Burnard, L. (2007). *The British National Corpus, version 3 (BNC XML Edition)*. Oxford University Computing Services. http://www.natcorp.ox.ac.uk/

Kyle, K. (2016). *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication* [Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. http://acl2014.org/acl2014/

McAllister, J. (2013). *Évaluation d'un dispositif hybride d'apprentissage de l'anglais en milieu universitaire: Potentialités et enjeux pour l'acquisition d'une L2*. Université de Nantes.

McAllister, J., & Belan, S. (2014). L'anglais de spécialité en LEA à la croisée des domaines: Étude de l'acquisition du lexique spécialisé. *ASp. la revue du GERAS*, *66*, 41–59. https://doi.org/10.4000/asp.4564

Starkey-Perret, R., Belan, S., Lê Ngo, T. P., & Rialland, G. (2017). The Effect of Form-Focussed Pre-Task Activities on Accuracy in L2 Production in an ESP Course in French Higher Education. In *Research-publishing.net*. Research-publishing. https://eric.ed.gov/?id=ED578668

Starkey-Perret, R., McAllister, J., Belan, S., & Ngo, T. P. L. (2015). Assessing undergraduate student engagement in a virtual resource center. Links between engagement, language learning and academic success. *Recherche et Pratiques Pédagogiques En Langues de Spécialité. Cahiers de l'Apliut*, *Vol. XXXIV N° 2*, pagination en cours. https://doi.org/10.4000/apliut.5184

Xiaofei, L., & Haiyang, A. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

[3] Available at https://lidile.hypotheses.org/projet-vizling