Conference Proceedings of the ALTE 8th International Conference

LANGUAGE ASSESSMENT FIT FOR THE FUTURE

ISSN: 2789-2344

ALTE Association of Language Testers in Europe

Conference Proceedings of the ALTE 8th International Conference, Madrid

Contents

Using a learner corpus to refresh rating scales of CELI exams

Fabio Zanda
University for Foreigners of Perugia, Italy

Danilo Rini
University for Foreigners of Perugia, Italy/Centro per la Valutazione e le Certificazioni Linguistiche (CVCL) – Centre for Language Evaluation and Certification of the University for Foreigners of Perugia, Italy

Abstract

In the last few years, we have witnessed an increase in the use of corpora to inform language testing and assessment practices. Among other purposes, the analyses of well-designed collections of real learner performances may be used as an effective counterpart to more traditional methods for the development and revision of rating scales.

In this contribution, we briefly present a learner corpus which consists of over 3,000 written texts produced by candidates of Italian L2 CELI exams, the CELI corpus (Spina et al., 2022; Spina, Fioravanti, Forti, & Zanda, 2023). (CELI stands for *Certificati di Lingua Italiana*, 'Certificates of Italian Language', issued by the University for Foreigners of Perugia, Italy). We then present a project of the Centre for Language Evaluation and Certification of the University for Foreigners of Perugia, where we explore the potentiality of the CELI corpus in informing the revision of CELI rating scales, combined with the consultation of assessment reference resources and the opinion of expert raters.

Introduction: The use of corpora to inform language testing and assessment

Corpora can generally be defined as large digital collections of authentic language productions sampled according to specific criteria to represent a certain language variety (McEnery, Xiao, & Tono, 2006). Being stored in electronic format, corpora allow for a wide range of computer-assisted queries and analyses, as well as systematic linguistic features' comparison with other similar corpora. Preliminary discussions about the potential applications of corpora in language testing and assessment (LTA) commenced in the mid-1990s since Charles Alderson outlined prospective uses to inform the development and validation of language tests with the aid of corpus data (Alderson, 1996). Following his intuitions and the noteworthy impact of corpus linguistics in linguistic analysis and pedagogy (Taylor, & Barker, 2008), corpus methods were introduced in LTA practices, signaling a steady increase in the exploitation of corpora for the development of new tests and in the maintenance and revision of existing tests (Barker, 2010; Cushing, 2021; Gablasova, 2020; Park, 2014). Another extension in the contribution of corpora to LTA was implemented with the advent of large collections of near-authentic learner texts compiled according to explicit design criteria (Granger, 2008), i.e., learner corpora. In fact, it has been reported that reliable learner corpus data 'have the potential to increase transparency, consistency and comparability in the assessment of L2 proficiency, and in particular to inform, validate, and advance the way L2 proficiency is assessed in the CEFR' (Callies, & Gotz, 2015, p. 3). Among other uses (cf. Barker, Salamoura, & Saville, 2015), learner corpus data analysis can be employed – often in combination with native corpora – for specific purposes in the testing cycle, such as to inform the development of word, phrases or structure lists (Capel 2010; 2012; La Russa, D'Alesio, & Suadoni, in print), to identify specific lexical units to inform new task types or ameliorate existing test formats (Hargreaves, 2000), provide plausible performance-based distractors for multiple choice tasks (Gyllstad & Snoder, 2021), or to supply an empirical basis to test developers when constructing or reviewing rating scales and descriptors for learner production (e.g., Barker, 2013; Hawkey & Barker, 2004).

Approaches in developing (and revising) rating scales

Fulcher (2003) and Fulcher, Davidson and Kemp (2011) oppose two major methodological approaches in developing rating scales: a *measurement-driven approach* and a *performance data-driven approach*. Both approaches present pros and cons (Fulcher et al.,

2011) and can be summarised as follows. On the one hand, the measurement-driven approach is based on intuitive methods in elaborating rating criteria, thus involving judgements of experts in language teaching and assessment. It engages in favouring clearness and usability of scales and is the most widely used. However, among the points of criticism that have been highlighted in opposition to this approach, the lack of concreteness and objectivity in the language employed in the descriptors stands out, as it may cause potential subjective misinterpretations of the scores and their meaning for raters. This raises questions about the reliability and validity of score inferences which, in addition to the absence of examination of real performances, make post-hoc quantitative or qualitative analysis of the resultant scales indispensable (Banerjee, Yan, Chapman, & Elliott, 2015). On the other hand, the performance data-driven approach is based on empirical methods, being derived from the analysis of real performance data (Fulcher, 2003, p. 92). Therefore, as a first step, this approach adopts a bottom-up method, as it 'identifies traits or features that characterize and discriminate written texts or writers across proficiency levels' (Banerjee et al., 2015, p. 6). In other words, in the performance data-driven approach, the development of rating scales is preceded by linguistic analyses of real performance data, which may be found in learner corpora purposely annotated and collected from exam data (cf. Barker et al., 2015). The scales derived from this approach do have the advantage of mirroring real performance data, which yet need to undergo time-consuming thorough analysis that 'tend[s] to generate linguistic constructs that either bear complex mathematical formulae or become extremely difficult to operationalize by human raters' (Banerjee et al., 2015, p. 6).

In light of the above, it would appear reasonable to opt for a mixed approach for the rating scale review process, relying both on real performance analysis of corpus data and on expert intuitions to improve usability.

Reviewing rating scales of the CELI exams

In this paper, we present a new research project of the Center for Language Evaluation and Certification (CVCL – Centro per la Valutazione e le Certificazioni Linguistiche[1]) of the University for Foreigners of Perugia (Italy) which aims to analyse and potentially revise the current rating scales of the *Certificati di Lingua Italiana* ('Certificates of Italian Language') (CELI), since constant monitoring and evaluation of existing scales is vital in standardised testing and assessment (Banerjee et al., 2015). As per the CELI exams corresponding to B2, C1 and C2 proficiency levels of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001), i.e., CELI 3, CELI 4, and CELI 5 exams, there exist analytic rating scales which are uniformly structured for all levels. In fact, all rating scales of CELI 3, 4, and 5 include the same four assessment criteria for the evaluation of written production tasks: *vocabulary control*, *grammar accuracy*, *sociolinguistic appropriateness*, and *text coherence and cohesion* (Grego Bolli, 2004).

The starting point of the project is to focus primarily on the vocabulary control criterion, check for updated vocabulary descriptors in assessment reference materials, i.e., in the CEFR Companion Volume (CEFR CV, Council of Europe, 2020), analyse the vocabulary actually produced by learners in written productions by candidates of the CELI at each proficiency level under scrutiny, and subject the existing scale descriptors of CELI 3, 4, and 5 (CEFR Levels B2, C1, C2) to a critical examination by CELI expert raters.

CEFR CV vocabulary descriptors

The publication of CEFR CV and the presence of entirely newly released or accurately refreshed descriptors reflects how recent studies and considerations over second language proficiency tend to stress the importance dedicated to word combinations and phraseological units in both language acquisition and L2 production (Ebeling, & Hasselgård, 2015; Siyanova-Chanturia, & Pellicer-Sánchez, 2019). In the CEFR CV descriptors concerning 'vocabulary range', in the levels of interest (B2, C1 and C2), reference is made to 'idiomatic expressions' for C1 and C2, but, very interestingly, at B2 level, the production of 'appropriate collocations of many words in most contexts fairly systematically' (Council of Europe, 2020, p. 132) is introduced as being characteristic to the level. Such examples clearly show how L2 assessment cannot set these features aside.

By comparing CEFR CV descriptors and CELI rating scales, a few expressions turned out to be overlapping, but a few others seemed to be missing in scales, whereas others were introduced there, probably with the aim of facilitating the work of raters. For instance, in CELI scales reference is made to the presence of errors in written production by candidates, a reference which, by the very nature of the approach chosen by the CEFR, is absent in the latter.

[1] CVCL webpage: https://www.unistrapg.it/en/certification-of-italian-as-a-foreign-language

Real exam data: The CELI corpus

In order to also base our reviewing process on real performance data, we chose to rely on the CELI corpus (Spina et al., 2022, 2023). The CELI corpus has been designed to systematically compile the written texts produced by different candidates of Italian L2 who have passed the CELI exams at B1, B2, C1 and C2 proficiency levels of the CEFR. Over 3,000 texts, elicited out of more than 60 comparable task assignments, were included in the corpus, with a balanced distribution of the tokens in terms of proficiency level, totaling c.600,000 tokens, thus featuring a pseudo-longitudinal design (Meunier, 2015), with 150,000 tokens per proficiency level (Spina et al., 2022, 2023).

Preliminary analysis of the CELI corpus

Research has shown that vocabulary is a key component in overall language competence development (Milton, 2013) and that phraseological competence plays a crucial role in language acquisition, processing, fluency and idiomaticity (Ellis, Simpson-Vlach, & Maynard, 2008; Wray, 2002). In view of this, we based our preliminary data analysis on recent vocabulary studies performed on CELI corpus data. First, we referred to a recent study (Forti, Fioravanti, & Zanda, 2022) which investigates one of the most popular constructs to analyse vocabulary knowledge, i.e., lexical complexity (Bulté, & Housen, 2012). Lexical complexity is defined as a multifaceted construct that includes the main dimensions of lexical diversity (the number of different words in a sample), lexical sophistication (the number of less frequent or unusual words in a sample) and lexical density (the ratio of content words on total words in a sample) (Kyle, 2019). Second, we also resorted to the investigation of phraseological competence, i.e., the ability to use phraseology and word combinations, operationalised as phraseological units, that is

> the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance (Gries, 2008, p. 6).

We computed different measures of phraseological complexity, which, echoing Ortega (2003), is defined as 'the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units' (Paquot, 2019, p. 124). The dimensions of phraseological complexity taken into account are phraseological diversity and phraseological sophistication for three typologies of co-occurrences that appear in specific syntactic relations (verb+direct object, adjective+noun, adverbial modifier+verb).

As for lexical complexity, the results of the analysis of B2, C1, and C2 sub-corpora indicate that there are differences in the development of complexity across proficiency levels, with a statistically significant linear development of lexical diversity. Concerning lexical sophistication and lexical density, although there are significant differences between the proficiency bands (B and C), these are not significant between the C1 and the C2 levels (Forti et al., 2022). With regard to phraseological complexity, we found that there are significant differences in the development of phraseological diversity across levels for all syntactic relations considered, while, again, for measures of phraseological sophistication, the results show significant differences for the relation of verb+direct object between the B and the C bands[2], yet not between C1 and C2. Conversely, for the adjective+noun and the adverbial modifier+verb relations the development appears not to be linear.

In a nutshell, according to the CELI corpus data, we could say that empirical research showed that learners present a development in the variety and originality of words and word combinations used in their texts as the proficiency level grows, but not necessarily in the 'rarity' of the lexical units employed in the context of a written production task in a standardised certification exam.

Impressions of raters on existing rating scales

On the basis of the qualitative analysis conducted on CEFR CV descriptors and CELI rating scales, and of the quantitative analysis of the CELI corpus, we proceeded to involve expert raters in the reviewing process of the current CELI 3, CELI 4, and CELI 5 rating scales[3]. Five texts per each level investigated were selected from CELI corpus and eight experienced raters were asked to assess them, using the vocabulary control criterion only. Afterward, questionnaires were submitted to raters. Questionnaires were built by dividing them into two main sections, the first one concerning the usage of CELI scale descriptors in assessing papers, and the second one concerning the structure and wording of CELI scales themselves.

---

[2] The B band comprises B1 and B2 levels together, while the C band includes C1 and C2.

[3] Current CELI 3 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-3-B2-scritto.pdf
Current CELI 4 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-4-C1-scritto.pdf
Current CELI 5 rating scale: https://www.unistrapg.it/sites/default/files/docs/certificazioni/competenze-punteggi-CELI-5-C2-scritto.pdf

From the first section it turned out that raters, in assessing the paper, put particular emphasis on the use of a lexical repertoire coherent with input and expected register (above 87% of cases), but also gave importance to the presence of errors in lexical usage (above 62%). The use of idiomatic expressions was not considered as important, but it is worth noting that, when asked, raters considered the appropriateness of phraseological units as being very important in assessment, just as vocabulary control and appropriateness, and its extent and variety. On the other hand, they underlined how those aspects were often not present in scales, and possibly should be included.

From the second section, mainly concerning the wording of vocabulary control descriptors in CELI scales, it turned out that scales were considered generally clear, but some of the terms used there are considered ambiguous, such as 'adequate', and the reference to the number of errors present in scales is considered as 'misleading'. Scales are considered generally exhaustive, but the absence of reference to appropriateness and metaphoric use of language was stressed, while, when it comes to easiness of use, raters underlined how too much is left to raters' interpretation, and too many aspects are to be taken into consideration.

Further comments stressed the difficulty of using scales while referring to a single component/criterion in assessment, without any reference to other aspects of written production, thus arising the ever-present questions about the use of analytic vs holistic scales in language assessment. Moreover, issues such as variety, originality, and appropriateness were mentioned as relevant features to be included in scales, while a few raters considered referring to errors particularly misleading in assessment.

Conclusion

In summary, corpora derived from learner productions can be indeed helpful to inform language testing and assessment practices. In the project that we presented, we chose to adopt a mixed approach to the review of the existing rating scales of the CELI exams, starting from the vocabulary control criterion. In this context, the CELI corpus was used effectively as an empirical basis: being a collection of real exam performances and thanks to its pseudo-longitudinal design, it served to identify and compute several vocabulary features across levels. In combination with corpus data, we resorted to the analysis of CEFR CV renovated descriptors concerning vocabulary and to expert raters' judgement on the existing descriptors in CELI rating scales. The analysis of rater questionnaires and of the opinion of CELI raters represent a precious instrument in determining how existing scales may be amended in order to eventually rephrase descriptors and thus achieve scales with easier applicability, possibly leading to a fairer assessment. Future steps in the project involve a within-level quantitative corpus analysis in order to possibly identify the features that actually discriminate between higher quality and lower quality productions of the same CEFR level; an in-depth analysis of raters' behaviour when assessing with the current scales; and the creation of a large database with in-text examples of lexical features indicated by raters as determining in their assessment, which could be included in future scales.

Acknowledgements

We would like to extend our gratitude to Daniela Alessandroni, Maria Cristina Bricchi, Claudia Fedeli, Marina Mancinotti, Elisabetta Marchetti, Franco Romano, and Roberta Rondoni (Centre for Language Evaluation and Certification, University for Foreigners of Perugia) for their contribution in the rating process and raters' questionnaires, and to Irene Fioravanti (University for Foreigners of Perugia) for the preliminary corpus analyses used in our paper.

Author contributions

The present paper is a joint effort by the co-authors. Zanda contributed to all sections except 'Impressions of raters on existing rating scales', which Rini wrote alone. Both authors contributed to the study design and to the final manuscript.

References

Alderson, C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds), *Using Corpora for Language Research. Studies in Honour of Geoffrey Leech* (pp. 3–14). New York: Longman.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, *26*, 5–19.

Barker, F. (2010). How can corpora be used in language testing?. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 633–645). New York: Routledge.

Barker, F. (2013). Using corpora to design assessment. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1,013–1,028). Hoboken: Wiley-Blackwell.

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 511–534). Cambridge: Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency Volume 32* (pp. 21–46). Amsterdam: John Benjamins.

Callies, M., & Götz, S. (2015). Learner corpora in language testing and assessment: Prospects and challenges. In M. Callies & S. Götz (Eds.), *Learner Corpora in Language Testing and Assessment* (pp. 1–9). Amsterdam: John Benjamins.

Capel, A. (2010). A1–B2 vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*, E3.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal, 3*, E1.

Council of Europe. (2001). *Common European Framework of Reference for Languages. Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume*. Strasbourg: Council of Europe Publishing.

Cushing, S. T. (2021). Corpus linguistics and language testing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 545–560). New York/London: Routledge.

Ebeling, S.O., & Hasselgård, H. (2015). Phraseology in learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 207–230). Cambridge: Cambridge University Press.

Ellis, N.C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*, 375–96.

Forti, L., Fioravanti, I., & Zanda, F. (2022, September 22–24). *Lexical complexity across proficiency levels in L2 Italian: some preliminary findings* [Poster presentation]. 6th International Conference for Learner Corpus Research (LCR 2022), University of Padua, Padua, Italy.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29.

Gablasova, D. (2020). Corpora for second language assessments. In P. Winke & T. Brunfaut (Eds)., *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 45–53). New York/London: Routledge.

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*. *Volume 1* (pp. 259–275). Berlin/New York: Walter de Gruyter.

Grego Bolli, G. (2004). Measuring and evaluating the competence in Italian as a foreign language. In M. Milanovic & C J.. Weir (Eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference, July 2001* (pp. 271–83). Studies in Language Testing Volume 18. Cambridge: UCLES/Cambridge University Press.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 3–25). Amsterdam/Philadelphia: John Benjamins.

Gyllstad, H., & Snoder, P. (2021). Exploring learner corpus data for language testing and assessment purposes: The case of verb + noun collocations. In S. Granger (Ed.), *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (pp. 49–71). Bristol: Multilingual Matters.

Hargreaves P. (2000). How important is collocation in testing the learner's language proficiency? In M. Lewis (Ed.), *Teaching collocation – Further Developments in the Lexical Approach* (pp. 205–223). Hove: Language Teaching Publications.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.

Kyle, K. (2019). Measuring lexical richness. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 454–475). London/New York: Routledge.

La Russa, F., D'Alesio, V., & Suadoni, A. (in print). Designing a corpus based syllabus of Italian collocations. Criteria, methods and procedures. *Revue Roumaine de linguistique*.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meunier, F. (2015). Developmental patterns in learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 379–400). Cambridge: Cambridge University Press.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 Vocabulary Acquisition, Knowledge and Use. New Perspectives on Assessment and Corpus Analysis* (pp. 57–78). Amsterdam: Eurosla Monographs Series.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly, 11*(1), 27–44.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (Eds.) (2019). *Understanding Formulaic Language: A Second Language Acquisition Perspective*. New York/London: Routledge.

Spina, S., Fioravanti, I., Forti, L., & Zanda, F. (2023). The CELI Corpus: Design and linguistic annotation of a new online learner corpus. *Second Language Research*.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il corpus CELI: Una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, *14*(1), 116–138.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy and N. H. Hornberger (Eds.), *Encyclopedia of Language and Education Volume 7. Language testing and assessment* (Second edition) (pp. 241–254). New York: Springer.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.