Bridging Corpora: Creating learner pathway across texts

Hugh Paterson III
University of North Texas / Denton, Texas
University of Oregon / Eugene, Oregon
Drexel University / Philadelphia, Pennsylvania
i@hp3.me

Bret Mulligan and **Anna Lacy** and **Patricia Guardiola**
Haverford College / Haverford, Pennsylvania
bmulliga, alacy, pguardiola@haverford.edu

Abstract

The Bridge, a linked data application supporting curriculum development is presented. It was developed with Latin in mind, but has been extended to Greek as well. It quickly helps instructors and students find new vocabulary words in newly assigned texts, based on texts they have already encountered in their curriculum.

1 Introduction

In this paper we present *The Bridge*, a linked data application, started in 2014 (Pistone, 2020) with on-going development designed for use by participants in language pedagogy processes.[1] *The Bridge* and its supporting tool-chains facilitate web-based interactions with texts as instructors and students navigate the learning and acquisition of new lexical items.

The Bridge is written in Python 3. It uses Python-based Natural Language Processing on texts to lemmatize them and then link lemmas across texts. The user interface allows users to query and receive reports regarding lexeme similarity across several selected texts. In this way, instructors, grounding their curriculum in texts, can map out the new vocabulary from text to text as they craft lesson plans. Likewise learners can look for new-to-them words, on the basis of the texts they have already been exposed to. In this way, learner pathways can be "charted" based on texts learners have already encountered. Our success in facilitating the acquisition of Latin and Greek has led us to believe that the application can be used in more languages than just English, Latin, and Greek. The code running *The Bridge* is available via Github.[2]

2 CEFR Applicability

Measuring an individual's language proficiency and language-learning progress is important for a host of reasons. The *Common European Framework of Reference for Languages* (CEFR) is a standard developed and widely used in the European Union for language competency description (Council of Europe, 2001). It is applied in the context of language proficiency assessment and language-learning curriculum development. Given the market position of the EU and its national languages, CEFR carries a significant presence in the area of language competency certification and language pedagogy, especially in the government and business sectors. Other systems for indicating language competencies have been mapped to CEFR. For example, the Cambridge English Scale used in the UK[3] and the dominant system in the USA, the *American Council on the Teaching of Foreign Languages* (ACTFL) system (American Council on the Teaching of Foreign Languages, 2016). In contrast to the ACTFL system, which is designed primarily for assessing oral language fluency, the framework consists of a set of competency descriptions covering the areas of speaking, reading, and writing.[4] The CEFR competencies are laid out in progressively increasing capabilities from the perspective of the pedagogical trajectory found in curriculum of commonly taught languages (CTL). CTLs are languages which have generally undergone substantial language development activities (Fishman, 1968; Ferguson, 1968). For example, languages such as English, German, Chinese, Russian, and Italian all have strong ethno-linguistic populations and are

[1] https://bridge.haverford.edu

[2] https://github.com/HCDigitalScholarship/FastBridge

[3] https://www.cambridgeenglish.org

[4] https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

languages that benefit from national-government level support. They are also marked by being used in communities that engage in intergenerational transmission. It is easy to apply the CEFR competencies to CTLs because they frequently rank at 0 or 1 on the *Expanded Graded Intergenerational Disruption Scale* (EGIDS) (Lewis and Simons, 2010; Bickford et al., 2015). That is, language use occurs in all the scenarios outlined in CEFR. However, for languages which score at a level between EGIDS 8a and 10, it is harder to consistently apply the CEFR competencies, assessments, and associated pedagogical methods. There are several reasons for this which vary by circumstances. Many of the *Less Commonly Taught Languages* of the world are also technologically under-resourced and do not yet have significant literary materials. Therefore, measuring language competency on the basis of a person's reading skills in a language as required by CEFR presents a challenge. In other cases—such as sign languages, endangered languages, and languages of antiquity (LA)—oral user communities do not exist. It is a challenge to prove CEFR B1 level competency under the requirement: “Can deal with most situations likely to arise whilst travelling in an area where the language is spoken”. These language use contexts appear to be at odds with the CEFR presumed relationship between oral/aural methods of communication and the written/reading methods of communication. More recent work has helped extend CEFR concepts to sign languages (Council of Europe, 2018). However, as sign languages are not the only non-oral languages, challenges exist in aligning curriculum and assessments to CEFR for endangered languages and LAs. Unlike many endangered languages, LAs such as Ancient Greek, Latin, Classical Chinese, Hittite, or Ancient Egyptian have large exploitable corpora. Endangered languages and LAs also differ in that LAs often have a significant educational presence but lack communities with current oral communication practices; although some argue that even for LAs, oral-first approaches support learners more effectively (Buth, 2020; Halcomb, 2020). Curriculum developers working with more commonly taught languages also use texts. Some have mapped texts or corpora according to a CEFR level (Xia et al., 2016; Wilkens et al., 2018) even though mapping text to CEFR levels and student capabilities to specific texts is challenging (Escobar-Acevedo et al., 2022). Using graded texts has some drawbacks as texts are not the same as performative communication which CEFR is supposed to be assessing. Nevertheless, it has long been the practice for the languages of antiquity to be taught through the use of texts—without the requirements for oral competency, and literacy in some language has been a presumed foundational competency.

3 Instructional Goals and Classroom Context

Our current classroom context involves the instruction of languages of antiquity through text based approaches. Considering both communicative (oral/aural/signed) and text based approaches, a rather uncontroversial assertion is that sufficient vocabulary acquisition is essential if a language learner is to gain fluency in the new language. This is true whether a student's learning environment prioritizes *Comprehension* or *Skill-Building* in fostering language acquisition (Krashen, 2017). Vocabulary knowledge is not sufficient for comprehension, as cultural context, grammar, and discourse structures also need to be acquired. Ultimately, successful language learners must possess an operational vocabulary that allows them to understand a text (or utterance). This common-sense observation is well-supported by research into second language acquisition in several languages. Vocabulary knowledge is repeatedly claimed as the single best predictor of reading comprehension (Hu Hsueh-chao and Nation, 2000; Stæhr, 2008). Within the context of English, Chall (1958, 156–158) showed that vocabulary difficulty accounts for as much as 80% of the variability in reading scores, far outpacing syntactical elements. While these findings have been supported by research in inflected languages—e.g., on German (Röthlisberger et al., 2023)—the effect in highly-inflected historical languages like Latin and Ancient Greek remains to be assayed. For instructors focused on fostering successful reading of historical languages, these robust findings strongly suggest the importance of matching reading activities with lexical knowledge.

Yet the reading and instruction of many historical languages are on the horns of a dilemma. These languages often comprise vast corpora—in the case of Latin estimated at over a trillion words—yet a typical Latin student might engage texts totaling just a few tens of thousands of words (or a mere 0.000002% of the total corpus). Within this small slice, novice readers routinely move directly from fabricated Latin in textbooks to difficult historical texts, whose reading grade level is akin to college-level texts (Gruber-Miller and Mulligan, 2022). To attain full comprehension, readers must typically know 95 to 98% of the words in that text (Hu Hsueh-chao and Nation, 2000). Yet many novice readers routinely know only 25% of the words in commonly-taught texts. While the statistics vary across language fields, the overarching concerns are the same. Instructors and independent learners have begun to pay attention to this dilemma, but lacked accurate and easily accessible tools to help them bridge the gap between their individual lexical knowledge and the lexical competence expected by the target text,[5] as other tools routinely provide full vocabularies. These often automatically-generated and so prone to provide inaccurate information, especially for homonyms and inflected forms.

4 *The Bridge*

While *The Bridge* currently exists and can be exemplified by use cases, it is also undergoing active development based on classroom support needs.

4.1 Example use case

Imagine a class in which students completed an elementary sequence in the language using a standard textbook (e.g., *Wheelock's Latin* Wheelock and LaFleur, 2011), but turned to reading a historical text after finishing only 36 of the 40 chapters in the textbook (a common scenario, either because instructors run out of school year or because the final chapters of textbooks often present less common grammatical constructions that can be glossed in reading). Imagine this same class aimed to read the open-access version of Nepos' *Life of Hannibal* at Dickinson College Commentaries (DCC).[6] The DCC version of the text includes vocabulary, but only other words that are not among the 997 most common words in Classical Latin that it has identified as the DCC Latin Core. Students using Wheelock have been exposed to a core vocabulary of 829 words (fewer if, as in our imagined scenario they have not yet finished the book); yet only 489 of these are also in the DCC Latin Core. Thus instructors who wished to know what words were known and unknown for their student would have a great deal of time consuming work to identify words for their students—or cast them to the lexical wolves and let them fend for themselves, which will almost certainly lead them to use suboptimal resources that provide both too much and inaccurate lexical support. Also, while it might be useful to know the global vocabulary needed for Nepos, our instructor and students might instead wish to focus only on the first assignment.

The Bridge can quickly produce exactly this list. The first chapter of Nepos' *Life of Hannibal* contains 77 unique words. By default, *The Bridge* list appears with macrons but one can easily toggle between macronized and unmacronized entries. One can display basic English definitions or more full definitions—or create a practice or self-quiz list by removing the dictionary entries or definitions entirely. One can also reveal more information about each word, its importance in the text, or its frequency in Latin more generally. One can reveal the first time every word appears in the text—and sort by that appearance, creating a running vocabulary for each sub-division of the text. One can reveal the number of appearances in the entire text (and also sort), creating a quick reference for those words that will reappear frequently or are unique within the text [toggle up/down]. One can reveal the part of speech; and add a link to powerful open-source dictionaries like *Logeion*, connecting our list with an authoritative lexical resource. Finally, one can also reveal the rank of the work within the *Bridge Corpus*, which boasts over 1.5 million words in a range of poetry and prose from antiquity to neo-Latin texts.[7] Every column of data is sortable.

But what makes *The Bridge* such a powerful tool is that it empowers users to customize the words that appear in the list. To return to our original scenario, students were not reading Nepos 1 with no lexical knowledge but having (supposedly) mastered vocabulary from the first 36 chapters of Wheelock. Instead of 77 words, there are only 25 unfamiliar words—still too many to expect students to divine from context but a much more manageable set, if one were to seek to prepare students to encounter them. But, of course, DCC commentaries already assume that students will not know any words that are not already among the 997 most common Latin words. So one could create a list

[5] Here we mean a competence with a finer granularity than CEFR competencies imply.

[6] https://dcc.dickinson.edu/nepos-hannibal/chapter-1

[7] Currently there are about 300 Greek and Latin texts, textbooks segments, and core vocabulary lists.

that shows only those words in the DCC that also appear in this section of our reading. This returns a list of the 22 words (17 if we exclude proper nouns) that could be the foundation for preparatory activities—a supplemental list. One can also use the *The Bridge* to create a list of the 55 words in the text that students have already seen for review or assessment purposes.

This process can then be sequenced as students continue to read and gain familiarity with new words. To take another possible scenario: imagine students are engaging with text in the Advanced Placement Program (AP)[8] selections of the *Aeneid*—or to better align with a typical weekly assignment, the first 100 lines of *Aeneid, Book 1*. One could construct a vocabulary list by excluding multiple sources of vocabulary: say, (1) the 50 most common Latin verbs; (2) the 400-most common words in the DCC Latin core; (3) all of the words from the *Cambridge Latin Course* textbook (Cambridge School Classics Project, 1998); and (4) any word that appeared in a text that you have already read, e.g., *Catullus 1* and the AP selections of *Caesar's Gallic Wars*. The resulting vocabulary list results in a useful learning aid.

The Bridge lists can be further customized using morphological filters: e.g., a list of just nouns, or just 3rd declension nouns, 3rd declension nouns and adjectives, or a list that excludes proper names (or just proper names). These lists can be printed or exported (as CSV files) for further manipulation or transfer to a flashcard program, question bank, or other media.

4.2 Usage

The Bridge has been well reviewed (Pistone, 2020) and has seen significant use among classicists. Usage growth beyond Haverford College resulted in over 24,000 unique user sessions in 2022.

4.3 Active development

To support this lexical tool, we are further developing *The Bridge* ecosystem to enable users to: (1) encode texts for analysis in this and other digital ecosystems; (2) analyze and compare the readability of texts; and (3) discover readable texts for data-informed lesson plans, syllabi, and curricula. Integration with Linking Latin (LiLa)[9] and its scheme is part of ongoing NEH grant funded work. The current vision for *The Bridge* ecosystem includes *Bridge/Lemmatizer*, *Bridge/Stats*, and *Bridge/Oracle*.

4.3.1 Bridge/Lemmatizer

Bridge/Lemmatizer will be a web-based environment, allowing more rapid, accurate, and detailed lexical and syntactic encoding of texts, and facilitating collaboration by faculty, students, and other contributors. Lemmatizers can be optimized for different languages. Our plan is to enable different lemmatizers for different language requirements.

4.3.2 Bridge/Stats

Bridge/Stats will be a web-based dashboard that displays information about lexical and syntactic difficulty—i.e., readability—for texts, and the effect that user-defined knowledge has on textual readability for one or more texts and/or sections based on their (1) generic readability; and (2) readability that factors in personalized lexical knowledge using metrics such as: (a) word length; (b) word frequency, or the prevalence of very common words; (c) lexical sophistication, or the percentage of rarer words; (d) lexical variation, or the variety of different words; (e) hapax legomena, or words that appear only once; (f) the corpus frequency of rare and/or unknown words; (g) the number of words per sentence; and (h) the number and length of subordinate clauses.

4.3.3 Bridge/Oracle

Bridge/Oracle will be a web-based app that allows users to discover lexically readable texts in the Bridge Corpus by revealing the authors, texts, and passages that have the highest percentage of familiar vocabulary alongside basic readability data, with users selecting the author(s), text(s), or genre(s) they would like to explore and then indicate their known vocabulary by selecting textbooks used, lists mastered, and texts previously read.

5 Conclusion

Early development of *The Bridge* ecosystem has focused on Latin but its framework has been designed to be language agnostic. This allows the development of Latin to serve as a model system for the longer-term goal of supporting the teaching and

[8] The *Advanced Placement Program* is a commercial educational program available through secondary schools in the United States. Passing students are generally given university level credit for course completion. The AP Latin curriculum is well known by classicists in the United States. https://apcentral.collegeboard.org/courses/ap-latin/course/ap-latin-reading-list

[9] https://lila-erc.eu

accessibility of other languages, beginning with Ancient Greek and then other historical languages. This can be further extended to other commonly taught modern languages, across a global spectrum. *The Bridge Readability Apps* will be designed for use with any language for which Natural Language Processing (NLP) resources exist, creating the potential of use cases far beyond its initial target audiences at schools, colleges, and universities around the world.

Acknowledgements

Bret Mulligan is *The Bridge* project director. Hugh Paterson III's involvement is sponsored via the *LEADING Data Science Fellowship* through the *Institute of Museum and Library Services* (IMLS) RE-246450-OLS-20. Michael Rabayda and over twenty other contributors have made *The Bridge* possible. Their names are in the project credits.[10] We are grateful to Paul Unger for editorial comments and to an anonymous reviewer for relevant suggestions.

References

American Council on the Teaching of Foreign Languages. 2016. *Assigning CEFR Ratings to ACTFL Assessments*. American Council on the Teaching of Foreign Languages, Alexandria, Virgina.

J. Albert Bickford, M. Paul Lewis, and Gary F. Simons. 2015. Rating the vitality of sign languages. *Journal of Multilingual and Multicultural Development*, 36(5):513–527.

Randall Buth. 2020. The Role of Pronunciation in New Testament Greek Studies. In David Alan Black and Benjamin L. Merkle, editors, *Linguistics and New Testament Greek: Key Issues in the Current Debate*, pages 169–194. Baker Academic, a division of Baker Publishing Group, Grand Rapids, Michigan.

Cambridge School Classics Project. 1998. *Cambridge Latin Course*, 4th edition. Cambridge University Press, Cambridge, UK.

Jeanne Sternlicht Chall. 1958. *Readability: An Appraisal of Research and Application*. Number 34 in Educational Research Monographs. The Ohio State University Bureau of Educational Research, Columbus, Ohio.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press; [CoE] Modern Languages Division, Strasbourg, France.

Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment — Companion Volume with New Descriptors*. Language Policy Programme, Education Policy Division, Education Department, Council of Europe, Strasbourg, France.

Adelina Escobar-Acevedo, Josefina Guerrero-García, and Rafael Guzmán-Cabrera. 2022. A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 14(3):1–8.

Charles A. Ferguson. 1968. Language Development. In Joshua A. Fishman, Charles A. Ferguson, and J. Das Gupta, editors, *Language Problems of Developing Nations*, pages 27–36. Wiley and Sons, New York.

Joshua A. Fishman. 1968. Language Problems and Types of Political and Socio-Cultural Integration: A Conceptual Postscript. In *Report on the Ninth International Conference on Second Language Problems, Tunis, 24–27 April*. English-Teaching Information Centre, London, England.

John Gruber-Miller and Bret Mulligan. 2022. Latin Vocabulary Knowledge and the Readability of Latin Texts: A Preliminary Study. *New England Classical Journal*, 49(1):80–101.

T. Michael W. Halcomb. 2020. Living Language Approaches. In David Alan Black and Benjamin L. Merkle, editors, *Linguistics and New Testament Greek: Key Issues in the Current Debate*, pages 147–168. Baker Academic, a division of Baker Publishing Group, Grand Rapids, Michigan.

Marcella Hu Hsueh-chao and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1):403–430.

Stephen Krashen. 2017. The Case for Comprehensible Input. *Language Magazine*, July.

M. Paul Lewis and Gary F. Simons. 2010. Assessing Endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique*, 55(2):103–120.

Amy Pistone. 2020. Review: A Digital Tool that Helps Teachers Generate Latin and Greek Vocabulary Lists. *Society for Classical Studies Blog*.

Martina Röthlisberger, Christoph Zangger, and Britta Juska-Bacher. 2023. The role of vocabulary components in second language learners' early reading comprehension. *Journal of Research in Reading*, 46(1):1–21.

Lars Stenius Stæhr. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2):139–152.

[10] https://bridge.haverford.edu/about/people

Frederic M. Wheelock and Richard A. LaFleur. 2011. *Wheelock's Latin*, 7th edition. The Wheelock's Latin Series. Collins Reference, New York.

Rodrigo Wilkens, Leonardo Zilio, and Cédrick Fairon. 2018. SW4ALL: A CEFR Classified and Aligned Corpus for Language Learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.