Leksis, an Adaptive Vocabulary Test For Low-Resource Languages

Alan Kersaudy

1884

PRIFYSGOL
BANGOR
UNIVERSITY

School of Arts, Culture and Language
College of Arts, Humanities and Business

Submitted in partial satisfaction of the requirements for the Degree of Master of Science in Language Technologies.

Supervisor Dr. G. Bovolenta

September, 2025

Statement of Originality

The work presented in this report is entirely from the studies of the individual student, except where otherwise stated. Where derivations are presented and the origin of the work is either wholly or in part from other sources, then full reference is given to the original author. This work has not been presented previously for any degree, nor is it at present under consideration by any other degree awarding body.

Student:

Alan Kersaudy

Statement of Availability

I hereby acknowledge the availability of any part of this report for viewing, photocopying or incorporation into future studies, providing that full reference is given to the origins of any information contained herein.

Student:

Alan Kersaudy

Abstract

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents the stakes, scope and purpose of the present dissertation. Particularly, the third section brings to light the role that educational technologies have to play in either supporting or further endangering low resource languages (LRLs), depending on whether the technology is meant to teach languages already endangering other languages. This section can be read as a general introduction to the field of educational technologies for those concerned with the fate of LRLs or as an introduction to the concerns of LRLs for those involved in the field of educational technologies.

1.1 Structure of the Dissertation

This dissertation introduces Leksis, a new recognition vocabulary designed test tailored for LRLs. This first chapter explains the rational behind such a test. The second chapter brings together parts of the available literature from different fields ranging from applied linguistics to information theory in order to set the ground for scalable vocabulary tests adapted to the limitations and context of LRLs. The third chapter present an initial test design for the Breton language. The fourth chapter analyses the results from the test to assess the relevance of the design choices. Finally, the fifth chapters assesses the value and limitations of the test, as well as presenting an informed opinion on the needs specific to LRLs in regards to both educational and language technologies.

1.2 Aim, Objectives and Research Question

LRLs face peculiar challenges in a world where data science made quantity the mother of all qualities. The ultimate aim motivating the present work is low resource languages teaching optimisation. The essential problem of any optimisation being the metric one aims to optimize, this lead to the development of rapid, minimalist language tests that will be introduced here. Especially, the objective is to find ways to make up for the resource scarcity problem by developing methods and techniques designed to work in this scarcity context first, instead of porting to LRLs methods and techniques that too data intensive.

For these reason, we propose the following research question. *Can language proficiency progress be measured reliably in low resource languages?*

Of course, the time constraints of this dissertation cannot allow for a large scale study of the progress in language proficiency of entire groups of learners during the full extent of a course. But by studying thoroughly the literature and presenting early evaluation, we intend to be able to propose a solid argument by the end of this work.

1.3 Background and Motivation

1.3.1 Terminology: AIED and EdTech

Modern academic research on educational technologies primarily falls under the "AI in Education" (AIED or AIEd) umbrella. This terminology dominates the field because of the "International AIED Society", founded in 1993, and the structuring impact of its journal issues and conferences. However, AIED may at time be used somewhat interchangeably with EdTech, for "Educational Technologies", which is a more product and market oriented terminology, a term that relates more to other neologisms like "FinTech", "BioTech" etc... Educational companies such as Duolingo or Rocket Language may be considered as EdTech companies for the industry, but belonging to the field of AIED for researchers. In yet another formulation, EdTech is AIED with a business model.

1.3.2 Lower Resource Language in Educational Technologies

The question of LRLs in AIED is tightly correlated with their general situation in the field of natural language processing (NLP). The situation is best described in Magueresse, Carles and Heetderks (2020), as statistical, connectionist, methods became dominant in NLP, the question of data scarcity becomes the main limiting factor in the application of modern NLP solutions for LRLs. This problem is also compounded with a general WEIRD bias in cognitive science (Henrich, Heine and Norenzayan 2010), where languages from cultures that are wealthy, educated, industrialised, rich and democratic tend to be privileged in all fields of cognitive sciences. However, if the adoption of these technologies is the most limited for LRLs, it is ironically these languages that stands the most to lose from not adopting them. Not adopting these technologies can cause a loss of visibility, prestige and desirability, which in turn leads to a lesser adoption and usage, leading to a vicious circle where less training resources are available to adapt these technologies to LRLs. This phenomenon as been described as the digital stagnation, or death, of a language, which is the online signature of socially extinct languages (Kornai 2013).

The role educational technologies could play in breaking this vicious circle, at least for some of the languages concerned, cannot be understated. On the one hand, helping to adapt existing educational technologies to LRLs languages can help to maintain their relevance as a teaching medium for parents who wish the best educational standards for their children and offer people seeking intellectual fulfilment an alternative to simply abandoning their mother tongue to keep learning new things. It is understood that NLP technologies such as automated translation can help to port established educational technologies to a large number of linguistic communities which do not possess the resources to otherwise develop their own educational tools (Haddow et al. 2022). A study by Horbach et al. (2024) supports the idea that educational equality can be achieved through cross-lingual scoring systems, in the context where open questions are used to assess skills, and where different linguistic backgrounds may impact the fluency of the students answers regardless of their understanding of the concept assessed. On the other hand, when it comes to language oriented educational technologies, the field is almost entirely dominated by research to teach English, and even coming in concurrence to already endangered languages. A paper by Henkel et al. (2025) is symptomatic of those risks. In this study, English speech recognition technologies are used in an AIED system to improve literacy in Ghanaian schools, a country home to more than 70 indigenous languages (*Ghana Languages, Literacy, Maps, Endangered Languages, Population, Official Use (GH)* 2025). To the best of our knowledge, it seems that little effort have been engaged in the academic literature to support the development of educational technologies specifically tailored for the needs of LRLs and their speaking communities, despite all the progress made in recent years to develop these languages in NLP. This lack of evidence may be caused by a language barrier, but this only reinforce the idea that more should, if must not, be done to support LRLs in AIED.

1.3.3 Artificial Intelligence and Education

As exposed by Doroudi (2023), artificial intelligence (AI) and research in education entertained a 70 years long dialectic that benefited the two fields of cognitive sciences. If early works on AI initially drew from developmental psychology and even developed educational tools as part of their endeavour to emulate human intelligence with machines, it is now the field of education that benefits from the possibilities unlocked by AI technologies.

Early research in artificial intelligence explored two different approaches to try to emulate cognitive processes. The first is commonly known as Good Old-Fashioned AI (GOFAI), it was centred around a symbolic approach that stemmed from Allen Newell, Herbert A. Simon and Cliff Shaw's seminary work on the Logic Theorist (Newell and Simon 1956). This approach sought to understand how experts solve problems using rules-based systems and symbolic abstractions. The second, connectionist, approach was centred around neural networks and focused on the acquisition of cognitive skills over performance proper. It was developed by people like Marvin Minsky, Seymour Papert and many others (Doroudi 2023). Papert notably, came to the AI world after having studied children cognitive development in Jean Piaget's laboratory in Geneva. He brought to the connectionist paradigm in AI a consequent influence from Piaget's constructionism, which is a theory that posits that learners build their skills and understanding on the knowledge and skills already acquired.

Both approaches led to attempts to create interactive educational systems early on. Examples of early educational software programs based on GOFAI comprise the GUIDON system, which relied on the Mycin engine, an infection diagnosis system, to teach students diagnosing pathologies (William J. 1983). The connectionist branch privileged the development of educative "micro-worlds", such as educative programming languages, in which children could learn unspecified problem-solving skills. Instances of such approach comprise the Logo programming language, designed to learn about relative positioning and geometry by designing programs to guide (drawing) robot turtles. Many systems followed Logo, like the Scratch programming language and the Lego Mindstorms kits. But the necessary specialization in AI led later research to strictly focus on computer systems performance, especially as the advent of back-propagation gave rise to deep learning (DL), achieving to establish the supremacy of the connectionist paradigm in AI.

At this point, the focus definitely shifted from using developmental psychology to support AI, to integrate AI technical solutions in educational tools. A meta-analysis by Schmid et al. (2023) now supports the benefits of constructivist educational approaches like Blended Learning (BL) and the Flipped Classroom (FC) , which give more of a coaching role to teachers, with the charge of the instruction being deported to online interactive systems, most often used outside the classroom.

In this section, we saw how Piaget's constructivist ideas in education first infused in the connectionist approach to AI through Seymour Papert's works. Then, when this connectionist approach took the world by storm with the advent of DL, AI came back to education in the form of adaptive learning platforms to support constructionist practices development in schools. Learning about this combined history goes beyond a mere inquiry for historical anecdotes, it gives us the scope and epistemological framework to fix the goals and methods of educational technologies, which is a necessary step to ensure that such tools could one day achieve real-world success. This is, not as an isolated system evolving in the vacuum, but tools in the service of a holistic learning environment.

1.3.4 Adaptivity and Knowledge Models

The Promise of Adaptivity

The key difference between classic textbooks or lecture-based education and most of the recent learning technologies is the promise of adaptivity. This means that the system adapts its behaviour based on the learners' performance, ideally with the goal to maximize their learning intake. In most modern systems, but not all, this maximization is done by a recommender system, the most sophisticated forms of which resolve an instance of the multi-armed bandit problem. Problem which may be solved by one of several different algorithms (Chen et al. 2017). The multi-armed bandit problem is the mathematical formulation of a situation where different actions are proposed, in our case, recommending different learning materials with uncertain pedagogical values, and an agent must decide which actions will maximize a specified reward, here, the students' growth in knowledge. Those systems must make an arbitration between exploiting actions with known, but limited rewards and exploring actions with unknown rewards.

This paradigm allows systems designers to free themselves from the headache caused by having to arbitrate the question relating to the selection of learning material, like their relative difficulty; one exactly on par with the level of the student, or one leveraging other teaching paradigms such as desirable difficulty, or a combination of the two. Depending on the algorithm selected, the promise of adaptive learning is to enable the construction of an individualized profile of the learners' skills, possibly also including a description of theirs learning capacity or rhythm, and to have the system build an optimized curriculum to reach the specified pedagogical goal.

It must be pointed out that more rule-based systems still exist, and are widely implemented, where the curriculum is designed in advance based on a pedagogical model, playing the role of the recommender systems presented above. Those systems may be relevant when the goal is to teach a specific, well-defined, sets of skills, like primary and secondary schools programs. Pelánek (2025) mentions the Umíme platform in the Czech Republic, that seems to be largely adopted by schools and relies on such architecture. Others systems may not even have adaptivity systems, but simply interactive properties, like the educational programming languages mentioned above, but those are not the focus of the present work.

Knowledge Model and Instrumental Goal

Where recommender systems can make the promise to optimize any given instruction, from a YouTube video watch-time to paperclip manufacturing (Bostrom 2003), AI systems do not bear the responsibility to define these intermediary instructions, what we call the instrumental goal. This question is at the core of all alignment considerations, and educational systems are no stranger to this probematic. In educational systems, this proxy is based on a knowledge model, also called student model, which are psychometric data from which can be derived a learning model (the evolution of that knowledge through time) which may in turn be used to define the pedagogical value of a teaching material, this metrics being the reward that a multi-armed bandit algorithm would be charged to optimize. The definition of this knowledge model and the nature of the psychometric construct it collects is quintessential to the success of an adaptive learning system, and this definition is the responsibility of the field that the system is intended to teach and psychological models, not AI directly.

1.3.5 Conclusion

In this section, we analysed the history of educational technologies since the cognitive revolution in the 1950s. We saw the invaluable potential of the still emerging field of AIED, and its promise of adaptivity, together with the risks and opportunities it brings to LRLs. We identified a gap in the literature on LRLs teaching in AIED. If the translation of AIED systems in LRLs may work as long as the topic it is intended to teach is not a language itself, when it comes to teaching languages, the innovations in the field of language educational technologies seems dominated by English, which is an outlier in terms of resources' availability when compared with the majority of the 7000 other languages spoken around the globe. In this context, it seems necessary to rethink how adaptivity can be achieved when most languages around the world don't even possess a scientific descriptive grammar, let alone the dozens of hours of annotated recordings necessary to train speech recognition systems.

Chapter 2

Literature Review

This literature review is divided in two main sections. The first section is dedicated to the analysis of the constructs that have been investigated in assessing language proficiency, and among those, which ones could serve in an adaptive learning system, while the second one dwell on the statistical ways to score and analyse a given construct. The guiding criterion throughout this chapter will be the simplicity of the solutions proposed, because, in any case, it is always easier to fix a simple system's shortcomings than those of a complex system.

2.1 The Proficiency Constructs and Where to Find Them

The introduction exposed how the definition of the instrumental goals that a recommender system has to optimize belongs in the domain of expertise relating to the final goal of the system, rather than in the technology itself. Language testing has traditionally been the matter of Second Language Acquisition (SLA) research, which can be seen as a subdomain of learning science (LS), but this field takes inputs from – and is closely related to – psycholinguistics, applied linguistics, and as we shall see, neuroscience, on which it depends for a general understanding of the processes involved in language use and acquisition. Without pretend to an exhaustive review, this section will attempt to provide a unified overview of language proficiency and the ways to measure it.

This question has been widely studied within various theoretical frameworks and for several practical purposes. Most noticeably, on the mastery of a language can depend the access to citizenship, educational institutions or work positions, which are live-making opportunities that have made its validation a social mobility issue. In this section, we start with the most widely accepted and used way to assess language skills, before moving towards alternative solutions that would fit the needs of a scalable adaptive learning system. Finally, we critically assess these alternatives. The second section is dedicated to finding ways to address the shortcomings of these alternatives.

2.1.1 The Holistic Approach to Testing (CEFR)

The complex latent traits like language proficiency can be assessed by two testing paradigms, the first being described as maximalist, comprehensive or holistic, and the second minimalist, proxy-based or reductionist. Commercial and institutional language tests such as the IELTS and the Cambridge English Qualifications for English or the DELF and DALF for French, to only cite these, follow a maximalist approach defined by the Common European Framework of Reference for languages (CEFR) (Europe 2020). This framework not only define the now famous six alphanumeric degrees of language mastery, but also the four usage contexts in which it ought to be measured, two modes of usage, oral and written, for two types of activities, reception and production. It measures the linguistic knowledge (vocabulary, grammar and their constituents) together with the four common language skills that language users engage with: listening, speaking, reading and writing. This framework is considered standard beyond the borders of Europe, but in spite of its strengths, it may not suit all contexts in which languages need to be tested.

The main critique that could be levelled at this testing paradigm is the fact that only ten European languages can brag about having CEFR-compliant tests spanning the six proficiency levels it defines (*Common European Framework of Reference for Languages* 2025; *Cadre européen commun de référence pour les langues* 2025). After twenty-five years of existence, even national languages of leading economies of the EU like Dutch or Czech, do not belong in this list. This is a fundamental flaw for a paradigm that was explicitly designed not to favour the main languages of the Union. The reasons for this are obvious, only the most "marketable" languages can develop an educational ecosystem strong enough to make these tests economically viable. Sometime, political will can breach the gap like for Spain's regional languages (Galician and Catalan are in the ten languages mentioned above, when Basque only lacks a test for the A1-2 levels), but this will is built on strong institutions and expertise that only a handful of languages have at their disposal in Europe, let alone in the rest of the world. Despite its theoretical grounding, the scarcity of resources (time, money, expertise and interest) that most languages have to deal with make a comprehensive language testing paradigm impractical. Once again, it is the languages that have the most to benefit from these tools, and the most to lose by not using them, that are facing the greatest difficulties to access them. Furthermore, in the case of an adaptive learning system, which is the main motivation for this dissertation, a comprehensive test would be both redundant and unpractical, as the testing would take too much time from the learning experience, unless the testing were to be part of the pedagogy.

We must thus look at more efficient ways to measure proficiency, but before this, we need to develop a deeper understanding of what language acquisition means at a neurological level, and how the abstract theoretical knowledge present in the unread dictionaries and grammars articulates with the two or four practical skills that characterize daily language use of all known human cultures around the globe.

2.1.2 The Intertwined Nature of the Proficiency Constructs

Most theories in linguistics, especially structuralism and generativism, are based on an analytic approach, first taking language in isolation, then separating its conceptual constituents, lexis from grammar, from performance (Chomsky 1965) and repeating process with their constituents and subconstituents, to then study the ways to combine them together. In a way, the CEFR paradigm follows the same epidemiologic trend, by dividing production and perception skills, oral and written usages. The main benefit of these analytical methods is obvious, by separating aspects and categories, one can cover an exhaustive understanding of the constituents and rules of a complex systems such as languages. But despite its strength, this analytical approach brings a biased view as to what a language is, as it brings a static and isolated representation to the systems it studies. However, languages, or for this matter language knowledge, never are a fully static structure nor a succession of synchronic states, because languages live the human flesh, they have to be acquired and forgotten by every passing generation and are never stagnant, nor limited to their internal structure. This is where modern approaches, like functionalism or cognitive linguistics Evans and Green (2009) come into play, along by developmental psycholinguistics, by bringing the focus to the acquisition and use of the language and it's relation to the body, rather than its structure. Bybee (1999) argues that usage-based linguistics can produce formal models, but with a twist. By stating that the competence comes as the formalisation of usage, almost as an emerging property, and this usage of the language being primarily a social, physical, embodied and cognitive activity, this new paradigm brings new considerations into light. Where generativism view performance as the materialisation of innate structures of the brain giving the structure precedence over anything linguistic, usage-based approaches consider structures as generalisation made by the language learning brain. This view goes beyond the simple inversion of precedence in what is an obvious chicken-egg situation. By insisting that cognitive processes always have some degree of dependence on embodied, sensorimotor processes, this view also breaks the Cartesian mind-body duality (Varela, Thompson and Rosch 1991) as well as Chomsky's competence-performance duality. In simple words, everything in the brain is (or eventually becomes) connected based on usage, and structures always come a posteriori.

These developments in linguistics proper are also supported by recent advance in neurology. Since their discovery by Vermon Mountcastle in the 1950's, it has been debated whether the cortical columns inuformally structuring the the grey matter in the neocortex play a role as a modular unit of computation (Horton and Adams 2005). The thousand brains hypothesis (Hawkins, Ahmad and Cui 2017; Hawkins and Dawkins 2021) is the latest iteration of this idea and proposes a model on how this unique architecture can, through a voting mechanisms, progressively map sensorimotor inputs towards and from different degrees of abstractions and to refine a unified representation of the world, and thus better engage with it in a continuous feedback loop. This produces a compelling argument on how abstract thinking and language can progressively emerge from sensorimotor interactions (Constantinescu, O'Reilly and Behrens 2016), when Chomsky's genes of a Universal Grammar are still waiting to be found anywhere.

Implications for Language Testing

At this point, the parallel between the CEFR testing paradigm and formal linguistics has to be clarified, because in the CEFR paradigm, in a way, we measure performance to deduce competence, so the link between those is never denied. But the epistemological critique of the quest exhaustiveness as undermining the understanding of the dynamics of the acquisition process still stands. If we are interested in the acquisition process and its dynamics, a complete, static representation of the skills is counter-productive. Furthermore, if the competence does not exist independently from the performance, could the skills be deduced from knowledge itself? This is what functionalist linguistics seems to argue for.

If everything is connected, if all is one (though one is not all), that is, if more practice leads to better practical skills, or performance, which leads to better theoretical knowledge, or competence, then, performance could in theory be measured through any construct describing competence, such as vocabulary knowledge. Vocabulary is especially interesting as its acquisition is a discrete, yet, never-ending process during a language learning journey. Eun Hee Jeon and Yo In'nami (2022) published a series of meta-analyses on the correlates of the different practical skills defined by the CEFR, all pointing towards this direction, with vocabulary knowledge being cited as a strong correlate for proficiency in listening (In'nami et al. 2022), speaking (Jeon, In'nami and Koizumi 2022), reading (Jeon and Yamashita 2022) and writing (Kojima, In'nami and Kaneta 2022). Note however that this does not mean that vocabulary knowledge causes fluency, although it contributes to it to the extent that fluency does not come without an advanced level vocabulary knowledge. This basic premise opens the door for low stake, low-cost, LRLs-friendly quick testing which may be more scalable and applicable in many areas, from self-assessment, to the development of automated language learning tracing systems mentioned in the introduction. Notably, in the context of LRLs, that some may call "oral languages", the idea that higher vocabulary level is linked to practical skill becomes even more likely, because the dominant way to access knowledge is a "more integrated usage" (one doesn't learn Rapa Nui in the books). This way, one may even posit that vocabulary testing becomes increasingly relevant as less written and digital resources are available to a given language.

The last implication of this first-principle and connectionist view of language acquisition is the absence of practical difference between the way competence in the first language (L1) and a second language (L2) are acquired, that is, through usage. Once the circuitry responsible for verbal communication is unlocked between age 1 and 6, either through monolingual (including a sign language) or multilingual education, the way new words are acquired is consistent across the languages spoken by a multilingual. If a word or a feature is discovered through integrated usage and the piece of knowledge in the brain stems from a sensorial experience present during the acquisition of the term, and if a word in L2 is learned as the translation of a word in L1, its representation in the brain will stem from the L1 word as its synonym within another "register" which is the network of the L2. The two scenari implying a formation of knowledge from the usage context but with no difference in status between the L1 and L2 networks. A word can be learned in L2 as the product of an integrated experience, and its L1 equivalent can be learned at a later stage as a "synonym in another space". As someone who learned about back-propagation in English first, my third language, I can assure the reader that I still need to think about the English word before finding its translations during a conversation in French or Breton. Once again, this equivalence between L1 and L2s is convenient in the context of LRLs, because these languages are often the low variety in diglossic regions, where the notion of native speaker and the line between L1 and L2 are often blurred.

2.1.3 A topography of Vocabulary Tests

It has often been shown that well-chosen proxies can give a reliable understanding of complex processes that one tries to measure. Economists have for example shown how nightlight measurement from space can serve as a reliable growth indicator in countries where official statistics may be lacking in quality or honesty (Henderson and Storeygard 2009), even without providing a causal mechanism for why this may work. Linguists imagined many ways to define and measure vocabulary knowledge, as they understood and demonstrated the strong correlation it had with other constituents of language proficiency. This last part of this first section of the literature review will give an overview of the different ways linguists attempted to measure vocabulary so far.

Productive Vocabulary Tests

The most integrated ways to test vocabulary consist of asking the test takers to give a synonym of a word, thus assessing the productive vocabulary skills, the words that the testees can, not only recognize and understand, but also retrieve from its meaning only. It is one of the strategies used to measure the vocabulary index, which is combined with three other indices to calculate the so-called IQ of the test taker in the Wechsler adult and children intelligence scales (Wechsler 2025).

Receptive Vocabulary Tests

In the middle are found a series of tests that aim to measure receptive vocabulary skills, the words that can be associated with their meaning by the test takers. The most widely used of those is the Vocabulary Levels Test (VLT), developed in the 1980s by Nation (1990) (see Kremmel and Schmitt 2017 for more details on its implementation, evolution and application). This test was designed for widespread use in schools as a placement tests for students. VLT is somewhat adaptive too, as it is testing the skills to associated terms related in meaning from different frequency ranges. An interesting receptive vocabulary test design is the Peabody Picture Vocabulary Test (L. Dunn and D. Dunn 2025). As it is based on pictures instead of written words, it allows testing children who could not otherwise read the words assessed. This picture-based approach could seem to make this testing design an ideal candidate for translation, and thus a candidate for a universal standard that could be applied even in environments where literacy is not widespread. However, this idea may be good only on appearance, as the calibration for the pictures-words mapping took place in an English-speaking country, and the words that may be used to describe similar situations may vary greatly between different linguistic spaces. This is what Kartushina et al. (2022) learned the hard way as they tried to translate the test in Russian for preschoolers, somewhat accidentally demonstrating that the Peabody test may be one of the hardest vocabulary tests to port to other languages, even ones spoken in a somewhat WEIRD society like Russia.

Recognition Vocabulary Tests

Finally, the simplest family of vocabulary tests are recognition vocabulary tests, sometime simply called simple vocabulary tests, they measure the aptitude to merely recognize the presence of a word, without requiring the justification of a further understanding of the meaning of the word. For an overview and assessment of different designs, see P. Meara 1994. The most successful design of this vocabulary testing family are the lexical decision task (LDT) vocabulary test, they were given many other names such as "Yes/No" or "binary" vocabulary tests, but all follow the same principle; a sequence of testing items, either real words of a pseudo-words (Paul Meara 2012) are presented to the test takers, who is systematically asked whether they think the item belongs to the lexis of language concerned. The results come in a combination of the four output defined by a confusion matrix, hits, misses, false alarm and correct rejection and different methodologies have been proposed to treat the results, from subtracting the percentage of the wrong answers from the percentage of correct answers, up to applying more complicated systems from Signal Detection Theory (SDT) (Huibregtse, Admiraal and Paul Meara 2002).

Many such tests have been built so far include at least one online version, and, encouraging fact, available in several languages English, Dutch and German (Lemhöfer and Broersma 2012). This paper showed encouraging results, with strong correlation of the vocabulary result with other traditional tests, thus supporting the idea that proficiency can be effectively measure through vocabulary testing. Another test has apparently been made for Croatian (**srce_how_2025**), although more information is not yet available. And this is in parallel with the numerous systems developed by Meara over the yearsP. Meara 1994. The main limitation of these systems is the fact that their items are limited and static, so they are never designed for a repeated usage, which would help measure the dynamics of vocabulary acquisition. This is a problem to be fixed, because the main interest of a minimalist test is to allow recurrent testing.

2.1.4 Relevance and Limitations of Vocabulary Tests

All the vocabulary tests presented above had their load of commercial or academic success due to their reliability in capturing different aspects of vocabulary acquisition. This shared reliability even works against the idea of seeing any of those becoming a standard, because they would all play an equally relevant part in this matter. We already explained the reasons why this should be so in section 2.1.2. If one admits that any sub-construct of proficiency is linked in the brain in a way defined by usage, that "all is one", then the same logic applies to vocabulary. Recognition comes as the first stage of vocabulary acquisition, without which any further development towards a more integrated usage is impossible. All these testing families measure different stage of the same integrated process of vocabulary acquisition. Nightlight measurement does not only measure the "nighttime electric consumption dedicated to street lightning of a territory" construct, but, as the statistics showed, it can be used as a GDP indicator, which is itself an indicator of economic health. The same goes for these vocabulary tests, they all are different constructs measuring the same phenomenon of vocabulary acquisition, which is an integral part of language acquisition.

The main differences between these tests are how resource-intensive they are and how integrated the constructs they measure are. Simple indicators like mere vocabulary recognition have weaknesses and can be subject to cheating or manipulation. The Economist's famous Big Mac index for inflation was allegedly the target of manipulation attempts by the Argentinian government in 2011 (Politi 2011) for this very reason. Similarly, the simpler to acquire a construct used as an indicator is, the more likely it is to become subject to manipulation attempts. But this does not mean that the construct as no value, indeed, both nightlight and Big Mac prices levels are still used today, but in scopes and at stakes relevant to their complexity. The same goes for psychometrics. In the context of vocabulary tests, the Peabody picture test's resource intensive design requirement make it need commercial use to support its complex development. The other, simpler tests achieve only academic success because they are so simple to put in place that they never need commercialisation, which limits their scaling potential and in turn their development. Nonetheless, they are all equally useful in measuring their respective stages of vocabulary acquisition.

2.1.5 Conclusion

In the context of an automated and adaptive testing with the purpose of tracing the acquisition of language skills, the vocabulary tests advantages outweigh largely other methods, and among them, the simpler vocabulary recognition tests designs truly shine, especially when considering the problem posed by LRLs. LDT vocabulary tests are simpler to administer in a fully automated way, and they are easier to port to LRLs because they can be derived from a simple list of dictionary entries. Yet, significant challenges remain before enabling a widespread implementation of LDT vocabulary test. The main limiting factor being the number of items proposed in the tests, both real and pseudo-words had to be selected from a larger set during a preliminary study in Lemhöfer and Broersma 2012. If an LDT vocabulary test is to be used in a recurrent way, to trace vocabulary progress through time, the items available for testing must be plentiful, maybe cover the whole lexis of a language or at least a significant portion of it. But then the question of the items calibration kicks in. There can be no question of thinking of scaling the preliminary study done for selecting the items in LexTALE to get enough items to allow reliable recurrent testing, already for a language with tremendous resources like English, let alone LRLs. Solving this problem of the items calibration would open the door to scaling both vertically (allow recurrent testing for the same language) and horizontally (allow porting the test to many languages). The next section will be dedicated to finding such a solution.

2.2 Knowledge Tracing

To paraphrase P. Meara (1994), many assessment tasks may be valid ways to assess vocabulary recognition skills, be provided the appropriate method of analysis. This section is dedicated to this problematic. Measuring latent traits from a tests items responses is a complex task known as Knowledge Tracing (KT) (Shen et al. 2024), which is a fundamental concept in Computerized Adaptive Testing (CAT). Part of this complexity depends on the assumptions one makes on the latent traits, are they a continuous construct or a set of discrete skills, which combine together in a multidimensional knowledge space, and if so, which skill depends on which others? These dimensions and the relationships between them can be defined manually or based on data, using Bayesian techniques or DL. Other assumption may include the influence of the testing process on the learning process, in which case one may factor in the half-life of new memories formed during previous assessment rounds. Fundamentally, this complicated choice of the model is an arbitration between accuracy and interpretability (Pelánek 2025). More qualitative models may be appropriate to inform recommendations of learning material, but presenting a proficiency vector as the result of a stand-alone test may be impossible to interpret than a single result.

Since this dissertation primarily focus on testing, a unidimensional, quantitative index seems more appropriate. Furthermore, the calibration of a qualitative paradigm would require large amount of data or resources like time and expertise, which are unavailable for LRLs. The end of this chapter will lay down the theoretical basis a this quantitative interpretation of the results of a LDT vocabulary test.

2.2.1 Theoretical Capacity of a Noiseless Unidimensional Test

The goal of the knowledge tracing model in a CAT is to make predictions on the outcome future test items in order to select the items whose answer are the most uncertain based on previous results. In the information theory jargon, this is called maximizing the entropy, which maximizes the gain of information by the model by minimizing its uncertainty. Drawing from Shannon (1948), one can define the theoretical absolute capacity of a noiseless binary test, before adapting it to a noisy environment. In a simple, unidimensional scale, finding this spot of highest uncertainty can be achieved with the binary search algorithm. Take a list of items ordered by difficulty, take an item in the middle, repeat the the process with the second half of the original list if the answer is right, else, with the first half. Repeat the process until the list is one item long. This algorithm has a time complexity of $\theta(\log n)$, which means that for n number of items, $log_2(n)$ steps are required to reach the last item. This is 10 items need to be tested for a scale containing 1 024 items, 11 for 2 048 items, 12 for 4 096 and so on...

Supposing that all the words in a dictionary of 30 000 words could be ordered by "difficulty", and that half the items of a test have to be pseudo words to deter cheating, a test using this algorithm would find the test taker's current level in only 30 rounds of testing, to compare with the 60 items used by a test like LexTALE (Lemhöfer and Broersma 2012). Even if we take into account the need for error corrections, the total number of step required will remain a proportional to this logarithmic progression. This setup has obvious limitations that we will address in the following subsection, but it bring interesting insights concerning the scaling problem of previous tests. Primarily, it is possible to test a really large number of items in a time efficient way, which opens the door to using the whole lexis of a language as testing items, rather than a selected list of words. This possibility in turn opens the door to unique testing experience, where the chances of going twice through the same testing experience are virtually non-existent. This unlocks the vertical scaling problem that was highlighted earlier in this chapter.

2.2.2 The Elo Rating System

Elo rating and Rasch Model

The first obvious limitation of the model previously proposed is the calibration of the items. One cannot get the relative difficulty directly from a dictionary, and the order in which words are acquired by learners may vary greatly depending on various factors. Most vocabulary tests go around this problem by grouping the items by frequency ranges (Nation 1990; P. Meara 1994; Dudley, Marsden and Bovolenta 2024). However, possessing frequency lists is often a high-resource language's privilege, and most LRLs don't have such resources at their disposal. For this reason, we propose that the difficulty rating of the words items be directly updated based on the results of the test.

In standardised test, this calibration of the items difficulty if achieved by Item Response Theory (IRT), which is a set of models derived from the Rasch model (Rasch 1980). The maths behind the Rasch model were rediscovered many times, including outside of the psychometric world, like in chess with the Elo rating system (Elo 1961; Elo 1986). The key equations for these models are presented below.

$$P(X_{AB} = 1) = \frac{1}{1 + e^{R_b - R_a}}$$

Figure 2.1 – Rasch formula

$$P(X_{AB} = 1) = \frac{1}{1 + 10^{\frac{R_b - R_a}{400}}}$$

Figure 2.2 – Elo rating system

In the Elo rating system, $P(X_{AB} = 1)$ is the probability of player A of rating R_a winning by checkmate against a player B of rating R_b. In the Rasch model, $P(X_{AB} = 1)$ is the probability of a test taker of rating R_a to successfully answer at a questionnaire item of difficulty rating R_b. Since both follow a logarithmic progression, the rating from a "Rasch rating" to an Elo rating is done by multiplying it by $400/ln(10)$ and inverting the nominator and the denominator to go from Elo to Rasch. The difference in the logarithm base and the addition of a spread factor of 400 in chess was meant to increase readability and interpretability, while matching rating systems previously used in the chess world. A 400 difference in Elo rating means a

1:11 vs 10:11 chance of victory, which is more interpretable than a 1 point difference meaning a 1:2.718 versus 1.718:2.718 odds distribution.

In practice, the main difference between the two systems lies more in the update mechanisms. Since IRT was developed for static tests (without real time adaptive features), it relies on more computationally intensive techniques, which are not well suited for the purpose of a CAT. Its simple updating system is why the Elo rating system has been gaining more attention in the AIED community over the years, Pelánek 2016 mentions several successful integrations of this model in adaptive educative setups, although never for stand-alone tests. The same article also present various update mechanisms that take into account different assumptions, such as correction for cheating strategies or short and middle term memory half-life. The update of an Elo rating is given by the following formula.

$$R'_A = R_A + K \times (S - P) \tag{2.1}$$

The actual score (1 or 0) S is subtracted by the prediction P of the outcome based on the score difference given in 2.2 (value between 0 and 1). If an outcome is certain (more than 800 rating difference) and the result follows the prediction, this value will be close to zero and the change in rating will be close to 0. If the opposite happens, the score increases by a value close to K, names the K-factor a value akin to the learning rate in the DL world. This value that may vary depending on the implementations of the rating system, but is often around 20 in the chess world. Sometime, an uncertainty function is used to progressively change the rate of update based on the number of updates (cf. equation 2.2).

Error Correction and Degeneracy

MCQ use three category of component, queries (the questions), keys (the right answers) and distractors (the wrong answers). Fundamentally, recognition vocabulary tests are a subset of MCQs, with a unique query for the whole test, and the real words as keys and the pseudo words as distractors. It is acknowledged that there may be different reasons why a test taker may select right or wrong answers. The most obvious one is that a test taker recognises the keys and ignores the distractors. But two other course of actions must be taken into consideration.

1. The test taker knows the answer but mistakenly selects a wrong answer (e.g. by answering too quickly and noticing the mistake too late).

2. The test taker does not know the right answer, and answers properly by pure chance.
3. The item rating does not correspond to its actual difficulty level because the calibration is not over.

It is understood that these effects add noise to the system and that the test should be made more redundant to compensate these effects. It is understood that if an answer is given for a good reason more than half of the time, the rating of the test taker would still converge towards its real value, although more slowly. Even in a setup where more than half the answers are given for wrong reasons, but the distribution of right and wrong answers is balanced, the model would still be able to avoid degeneracy. But in any case, the number of items tested in a test session shouldn't be made as short as theoretically possible, but take these noise into consideration. Once again, the Elo rating system does this seamlessly with an "uncertainty function". Pelánek 2016 proposes the following uncertainty function to update the rate of the ratings update in function of the number of previous answers.

$$(n) = a/(1+bn) \tag{2.2}$$

Where a and b are positive constants and n the number of previously answered items. The resulting number is used as the K value that multiplies the correction of a rating after an answer. Once again, we'll come back on this aspect in the next chapter.

2.3 Conclusion

This literature review introduced ideas from several fields and attempted to organized them in a coherent whole. From a psycholinguistic argument supporting the idea that vocabulary can be used as a proxy for general language proficiency. To proposing a knowledge tracing model that optimizes the information gained by the results of a binary test. In the next chapter, we shall put these pieces together to build a working recognition vocabulary test.

Chapter 3

Methodology

This chapter describe how a binary vocabulary test was designed based on the points highlighted in the literature review. When language-specific aspects of the methodology are mentioned, such as the sourcing of the types from an online Breton dictionary, it is understood that an analogue method can or has been used for a test in another language.

3.1 Sourcing the Keys

One thing to understand about vocabulary recognition tests is that the they don't test words knowledge per se, because they dismiss the meaning of the words. As long as the string of character is associated with a dictionary entry, what one calls a word type, the item is "real", and is expected to be recognised. In Breton, the word *brec'h* means arm (the part of the body) like in Welsh *braich*, but also (small)pox, like in Welsh *brech*. Those are two distinct words and have always been. But when testing the type recognition skills, most test taker will most likely think "arm" and completely ignore the "smallpox" meaning, the knowledge of which would mean a higher vocabulary knowledge. It is even expectable that many would think that the word *brec'hadur*, vaccine, is related to the meaning "arm", because they were inoculated vaccine in their arms and could not think of other etymologies. However, when facing the type *brec'h-vihan*, "small-pox", what would these people think? Most likely something around this line: "little-arm? in one word? which is the big-arm? this does not make any sense, it is not a real word!". Only the people aware of the second meaning of the word *brec'h* would recognize *brec'h-vihan*.

This little example shows how how the types differs from words proper, and how their rating is expected to be levelled down to their simplest interpretation, although lesser known meaning can style be expected to be found in derived, more advanced terms.

For the Breton vocabulary test, all the entries of the Breton diachronic dictionary Devri were fetched, and rules where designed to remove proper nouns and affixes. Since the monolingual dictionary Meurgorf classifies its entries in one of three categories: frequent, common and

Category	in Meurgorf	also in Devri	only in Devri
Frequent	1 108	946	
Common	47 740	26 197	
Rare	6 867	4 868	
Total	55 715	32 011	30 158

Table 3.1 – Categories of word types extracted from Devri (filtered) and Meurgorf

rare, it was possible to organise the items in four categories, the three previous ones, and the items that were in Devri but not in Meurgorf. The distribution of the items is shown in 3.1. The reasons why so many words seems to be absent in Devri, is that Meurgorf entries contain many proper nouns and affixes, as well as neologisms built with common affixes. The total number of available keys for the test was 62 169, half of which were given a rough difficulty rating between 1 and 3. The entries from Devri not found in Meurgorf were added the the category of the rare types.

Obviously these categories are not perfect, but they are still a precious help to the difficult questions of calibration. Other methods of sourcing types and different ranges of frequencies for other languages could include fetching the entries from dictionaries of different sizes. The entries present in the smaller dictionaries would be understood to be the most useful and frequent. The section on the rating initialization shows how these initial ratings are used. The code for these steps can be found on GitHub[1].

3.2 Generating the Distractors

3.2.1 Training the Model

For a study of the scale of this project, manually crafting the non-words is not an option. Different methods of computationally generate pseudo-words have been developed over the years, most of them chaining n-grams taken from a training dataset of various sizes (New et al. 2023; Keuleers and Brysbaert 2010). However, since some languages are known to exhibit features of phonotactic long distance relationships, such as vowel harmony in Turkic languages, n-gram-based models were deemed non-optimal to generate unlikely pseudo-words. For this reason, the use of Long Short-Term Memory (LSTM) was privileged (Hochreiter and Schmidhuber 1997). The design will be straight-forward for people familiar with Recurrent

[1]For the sourcing of Devri's entries and their filtering, see this Jupyter notebook, for the the range of frequencies, see this other Jupyter notebook

Neural Network (RNN), but some optimization technics were developed to increase the speed of training. Since the words in the training dataset (the keys from the previous section) are of various lengths, no the batches are of length 1, which means that no parallelization was possible during training. To circumvent this problem, the words were concatenated in a hundred longer strings, with a new line character used as the special token, to start a sequence of words, to separate each word and to end each sequence, thus making these sequences both compact and human readable. 15 such sequences were kept for validation and 85 for training proper. Around 10 embedding dimensions (to represent the characters) and 180 hidden dimensions (to memorize the patterns) in the one LSTM cell were more than sufficient to train an "orthographic" language model able to generate good quality pseudo-words. Between 10 and 20 epochs are enough to obtain a low cross-entropy of below 1.8, and thank to the batching technique mentioned above, the training barely took around 10 to 12 second per epoch. Note however that a low cross-entropy was not systematically obtained, even with the same hyper-parameters. This is where another optimisation technique comes into play, a sensible effect on the loss function progression by to reshuffling the words in a different order and remake different sequences of around 620 word types. In a way, this was "generating more training data" where the only common point between the previous sequences and the new was the internal structure of the words, and the relationship between the words would not be taken into account by the LSTM hidden vector. This effectively swapped some words from the validating to the training set, but as the training dataset's loss function was consistently lower for all the trainings, thus showing no sign of overfitting even with large numbers of epoch, this was deemed not to be a problem.

3.2.2 Generation

Once the character-based language model trained with a satisfying cross-entropy, it is ready to generate new words. Dividing the probability of the next token by an increasing temperature value increases the entropy of the softmax function distribution (last layer of the network) and thus tends to equalize the chance of the next token selected. It was found that a temperature of 0.7 was the sweet spot for a good balance between diversity and correctness of the characters generation. This sweet spot was found by looking at the proportion of words starting by the letter z, in Breton 1:2000 types start by a z. Obviously, different languages, especially languages with other alphabets would need another temperature.

As the goal for the network is to reproduce the training data with the biggest fidelity as possible, it will try to generated real words. The real words have to be filtered out, which was done in two different ways. Every time a new word was generated, when a new line character is generated, a new word is generated, the string is then compared with the available types in the training dataset, if it is not in the training dataset, the word is checked against a Hunspell spelling dictionary, and only if the spelling is not recognised, the word is added to a set generated pseudowords, with a high degree of confidence that the word is meaningless. The code for the generation of the pseudo-words is available on GitHub[2]. This method was used to generate an equally large number of pseudo-words as real words, which was later used for statistical comparision of the two sets of strings.

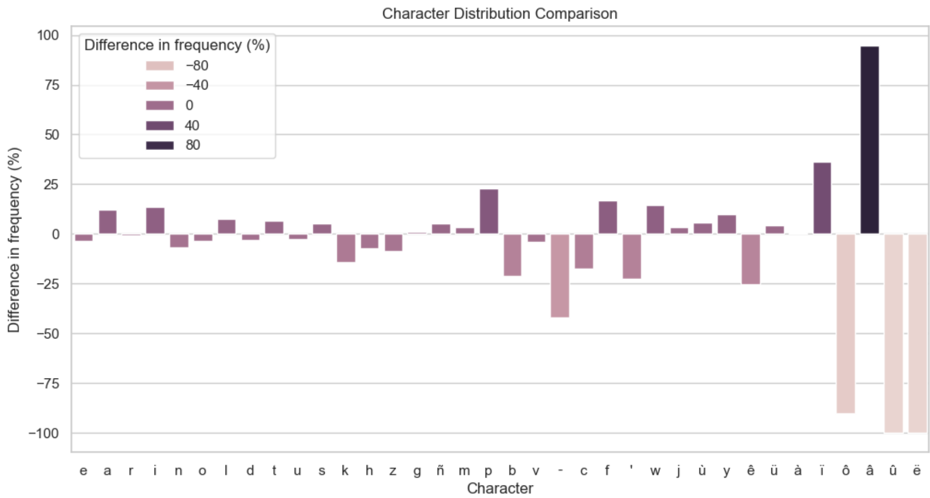

Figure 3.1 – Distribution of characters (pseudo-words / real words)

As one thing that could give away a pseudo-word is the disproportion of some characters in the words, the Figure 3.1 is used to inspect the distribution of characters throughout the sets of words and pseudo-words. If the value for a given letter is positive, it means that a character is over-represented in the pseudo-words set, and the reverse if the value goes negative. The characters are ordered by frequency, *e* being the most common character and *ë* the rarest in Breton (found in once in the real words, and never in the pseudo words, hence the 100% difference). Overall, the distribution of the character in the generated pseudo-words seems coherent with that of the real words.

[2]See this Jupyter notebook for details

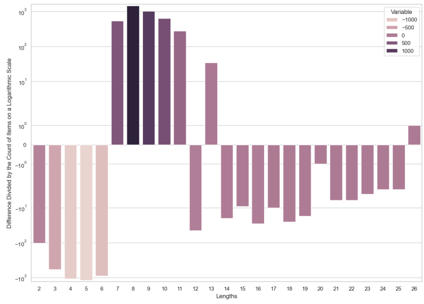

Figure 3.2 – The difference between the count of pseudo-words over real words on a logarithmic scale for a given length.

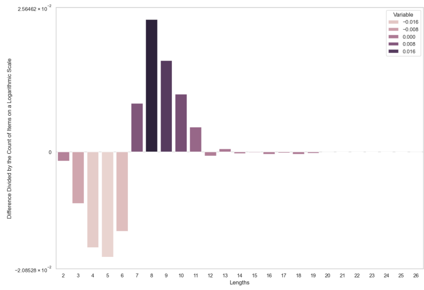

Figure 3.3 – The same difference as in 3.2 divided by the total number of items to bring the differences in the context of a test session.

The figures 3.2 and 3.3 are of particular interest. As the length differences could also give away clues to the test takers on whether a word is real or not. One can see that the network did not produce as many short pseudo-words as expected, where items of a length between 7 and 11 are over-represented to some degrees. This may be due to the fact that less "antimatroid" (combinations) of characters are possible for smaller length, compounded with the fact that many of the possible words are already "taken" by real words. This excess in one direction being caused by the limitation in the other direction. Knowing this, different rules could be designed when generating new words in order to compensate this phenomenon, like manually increase the likelihood for a new line character below a given length threshold. However, this may be considered as over-engineering. When scoped down to the total number of items, in 3.3, one can see that the lengths biases are irrelevant and unlikely to give away a pseudo-word. With a maximum variation of 1.6% there is no way a test taker would be able to rely on lengths difference to guess whether an item is real, even less so consistently, throughout several tests. The detailed methodology for these figures can be found on GitHub[3].

3.3 Initialization the Items Rating

The absence of initial calibration poses a tremendous challenge for a well functioning test. The test needs to be calibrated enough so that speakers with a limited vocabulary range be presented words that they will recognise. If test takers feel demotivated by a testing session, they are unlikely to take the test again, which will impede further the calibration process. To avoid this vicious circle, we need a calibration without calibration. As already stated, LRLs often lack a words frequency lists, so this technique will not be developed here, although it

[3] See this Jupyter notebook for more details.

may prove useful for some languages.

3.3.1 How the Initial Ratings Impact Adaptivity

A simulation by Pelánek (2016) provides insightful, if not surprising information on the question of calibrating the items difficulty scores with the Elo rating system. The idea that a fully adaptive system is leveraging uncertainty to maximize the gain of information is challenge by his results, which are improved when some randomness is added to the selection. Unfortunately the paper does not give informations on the initial distribution of the items, whether it was random, or set to a unique value for all items would undoubtedly change the benefits of an adaptive selection of the items. Considering a setup where all the items are set to have the same initial difficulty score, a fully adaptive test would be biased to select items that have already been selected, as the items which have not been yet selected would still be clustered at there initial value, a value that is unlikely to be reached by the test takers as the start deviating from the norm. The question of the initial rating of the items may be why adding randomness to the selection of the items had a positive impact on the correlation of the estimated items difficulty with their ground value.

3.3.2 The Modulo Clustering

The idea of clustering several items around a single value can however be leveraged to optimize the testing experience and the precision of the estimation of the test takers, if not the difficulty rating of the items themselves. To do so, we propose to randomly spread the rating of the items in three difficulty ranges (from the frequency category mentioned in the first section), and then to cluster these initial ratings around the closest multiple of a value. This "modulo clustering" leaves gaps in the initial estimated difficulty distribution of the items. The items that will fill these gaps are items that will have been assessed already, and as the rating of the test taker evolves, there will be more chance that a fully adaptive selection assesses items that have been already tested (whose rating is closer to the ground truth), but this bias will be balanced with the fact that new items will still be selected from time to time at all ranges of proficiency being tested. The selection of a low multiple, 2 or 3, will mean that more uncalibrated items will be shown to the first test takers, while a higher value, around 5, 10, or higher, will privilege the selection of items that have already been assessed often, thus limiting the diversity of the test sessions. For the Breton test, a value of

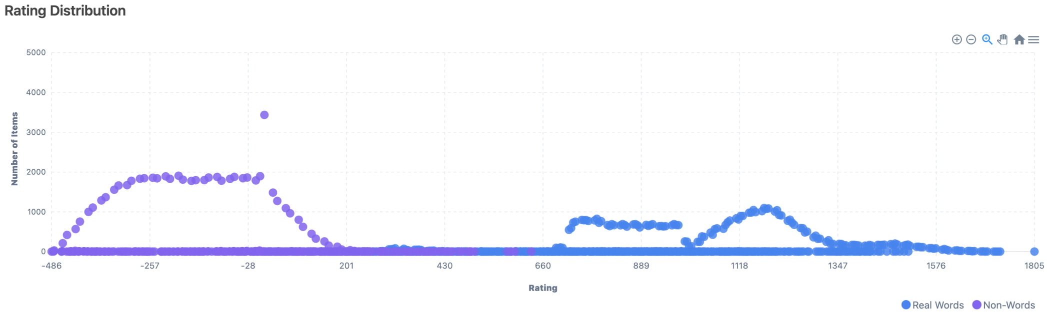

Figure 3.4 – Items distributions after the calibration process is initiated.

5 was selected for the clustering, as a way to balance accuracy with diversity. As explained above, having clusters to far away from each others may influence the selection process to such an extend that the calibration process itself is degraded. Such a system progressively transition from giving test takers a score mainly dependant on the ratio of word types they can recognise in a random distribution towards a score that depends on the recognition skills in regards of the other test takers scores. It is understood that for a test with as many items to be calibrated and so few potential test takers as in a LRL context, the test will never be fully one or the other, but a proportion-based system constantly moving towards a calibrated logistic scale, where the most calibrated part is the range representing the lower level of proficiency. This system is only possible by a fully adaptive system using this modulo clustering technique.

3.3.3 Initialization of the Distractors Rating

The distractors, or pseudo-words being expected to be recognised less frequently, their ratings is expected to go downwards. If the the ratings of the keys (real word types), is capped above zero, in order to show that a test score above zero is symbolically a non-null knowledge of the language, the rating of the distractors can be negative. Otherwise, the distractors would cluster at a rating of zero. For these reason, it was decided to "take advance" on the calibration and give non-words items a random ratings below zero, which means a difference in the average rating between the keys and the distractors. During the selection of the items, this difference is corrected by adding the difference between these means to the rating of the test takers. This has for effect to "punish" more severely the recognition of a non-word that the increase in rating rewarding the recognition a real word. This way, a cheater who consistently pretend to recognize all items (real or not) would face a steep decline in rating, instead of a stable rating. The figure 3.4 gives a representation of the distribution of the items by ratings. The clusters can be seen hovering over the items which are in the process of being calibrated, which may not number above 10 for a given rating. In a more calibrated distribution, one would see the two line merging in one. As we can see in the same figure, the calibration well advanced for the words of the highest frequency range (the small blue bump below 500 rating), which is exactly the range of test takers for which a simple proportion of correct answer (PC) rating-based test would not be suitable, and for whom a logistic scale such as the Elo rating system is needed.

3.4 Items Shortlisting

From this section onwards, we transition away from the question of the items to focus on the mechanic of the test proper. The test was deployed on a web platform openly accessible without requiring users to create an account[4]. The behaviours presented in this section happen on the front end of the application. As a test session is expected to use only a small portion of the available items, the idea of shortlisting the available items emerged. Instead of randomly sampling items from the lists of items, which would select almost exclusively items that have not been calibrated, the test select items by unique rating. The items that are not selected during this shortlisting are thus the items clustered at their initial modulo based ratings. Those items would have little chance of being selected without this shortlisting anyway, because in an adaptive setup, they belong in clusters of several hundreds of items. This shortlisting of the available items thus increase the performances for a test session without weighting on the quality of the test. Since the items are shuffled before being selected by unique ratings, the items shortlisted are never exactly the same (especially those which have not been selected yet), thus contributing to the uniqueness of each testing session. In practice, several items with the same rating may still exist after the shortlisting because this selection of item by unique rating is repeated as many times as necessary until the lists of keys and distractors a length of 4000 elements each.

In the case a test taker wants to retake the test after finishing a session, the same (shortlisted) list of items is used, note however that every time an item is selected, it is taken out of the list of available items. This means that someone taking the test a second or third time will never see again the same items. This feature could be used to assess the reliability of the test scores. If the items vary, the only common point between the different testing sessions is the (never fully calibrated) item scores themselves.

[4] See https://leksis.bzh

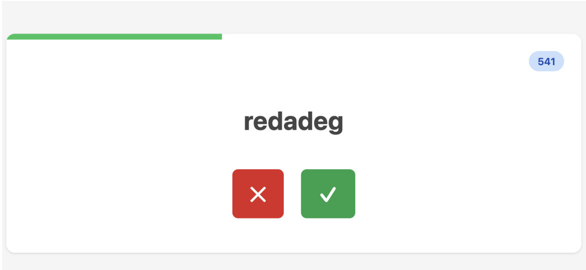

Figure 3.5 – A screenshot of the test's interface in the middle of Breton language test session. The word "redadeg" is famous in Brittany because of a biannual relay race taking place across the country. Many non-speaker would still recognise this word, which makes the answer from this term a case of the previously mentioned third reason for rating update given for the wrong reason: the rating of the item does not match its real proficiency range. This problem is expected to fade as more people answer the test.

3.5 The Testing Session

3.5.1 User Rating Updates and Session Duration

The actualisation of the test takers rating takes place in real time and the current rating is shown to them. There are two ways to loose points, by not recognising a real word or by recognising a non-word. Not recognising a non-word does not influence the rating and only recognising real words increases the rating.

The logarithmic base for the progression is 10, with a spreading factor of 400, like in chess in order to keep the rating human-readable. The rating is always shown to It uses the uncertainty function 2.2, with $a = 100$ and $b = 0.5$. This means that a correct recognition of a real word brings 47 points the first time a real word is presented to the test taker, and roughly 8 points after a 100 times, that is, half the result of the uncertainty function because the probabilities of correct answers are always around 50%. However the uncertainty function is capped to 20 point in order to maintain a steady growth for better performers. The pace of the rating growth is important as the length of a testing session is determined by the current rating, see below the equation that determine the number of real words to be answered.

$$f(x) = 10 + x/14 \tag{3.1}$$

Where x is the current score. This way, a poor performer is not expected to go through a long testing session. Consider a score averaging around 146 (obtained after merely three consecutive good answers), the testing session would only last around 22 items shown (11 real words and around 11 non-words). So ideally, the test would spend three real words to climb up to the test taker's rating, and the 9 remaining items would be used to sort what words do a 150-ish level learner knows and ignores.

When the last real word item is answered, the test results are stored and sent anonymously to the website's data base and the final score is shown to the test taker.

3.5.2 Items Selection

When a test session starts, the program first randomly decides which list it is going to select an item from. If the list selected is the keys, then the items with the closest ratings to the test taker's current rating is selected. As mentioned earlier, when the distractors list is selected, then the difference between the average of the distractors rating and the average of the keys rating is added to the test taker's current rating. If the testee's current rating is 500 and the difference between the average keys and distractors rating is 600, then the test will look for an item with -100 rating. This ensures the diversity of the non-words selected as the test taker's current rating move away from the distractors range.

For performance reasons, the test does not wait an answer to find the next item. It as soon as a new item is displayed on the screen, the test computes the next ratings for the two outcomes, good or bad answers, and selects two items based on these changes in rating. This system, along with the fact that the lists of items are shortened before the start of a session, ensures a smooth and seamless transition after each answer.

3.6 Items Rating Update

The rating of the items is recomputed in the back-end, once a day at midnight, based on the detailed test score stored in the database during the previous day. The system follows the Elo rating system update, also, based on the fact that these updates are asynchronous, one could imagine other systems to update the rating. For example, a real word whose initial rating was 100 not recognised by a test taker whose ultimate score is 500 should probably be increased at least to 500. There must be better ways to update a given item's rating, but ultimately, the time pressure led to a simple update of base on the difference between the prediction and the actual score multiplied by a K-factor of 20 (instead of an uncertainty function). Note however that the expected score of an item is calculated in function of the final score of the test session, not the rating at the moment the item was being answered.

Since the generation of the pseudo-words is expected to produce strings of various degrees of credibility it is admitted that some pseudo-words will be completely unlikely and other pseudo-words will be actual, meaningful words which were never added to a dictionary. This variable credibility would be considered a problem in most applied linguistics experiments, as an equal degree of nonsense is expected from all non-words, but it is an inevitability when items are generated by the tens of thousands. By updating the rating of the non-words, this test instead recognises the fact that all non-words are not created equal, and that unforeseen properties of a pseudo-word may allow people to recognise these strings as actual words. This means that pseudo-words whose rating increases to a point where more than half the test takers would consider it a real word (including more advanced speakers) may eventually be identified and moved from one list to the other. The test does not provide mechanisms to do this yet, apart from downloading the JSON file of the items with their current rating and manipulate the file manually before loading it to the web application again. But the possibility that generated pseudo-words may make sense for test takers is well taken in consideration and the problem is at least partially solved by the current design, as frequently recognised items will deviate from the norm and be shown less and less thank to the update system.

3.7 Adding Languages

3.8 Feedback

Since a widespread use of the test is essential to better calibrate the items ratings, two strategies were developed to increase engagement. First, the ability for users to share their scores with a link to the test. Second, an elaborated large language model (LLM) prompt that integrates the results of the test, destined to build a constructive feedback by teaching the meaning of the unrecognised words. This personalised and interactive lesson focuses on the words with the lowest rating, then asking the user to build sentences using the new words. After which it proposes to go deeper in elaborating on these word, by showing multimedia content that use the words words, or to keep going in learning about the other unrecognised real words. The prompt can be copied and pasted to the user's favourite LLM, or, if the navigator allows it, directly shared with the LLM app with the navigator.share() API. The version of the prompt at time of writing can be found in Appendice A, along with an example of answer from GPT-5.

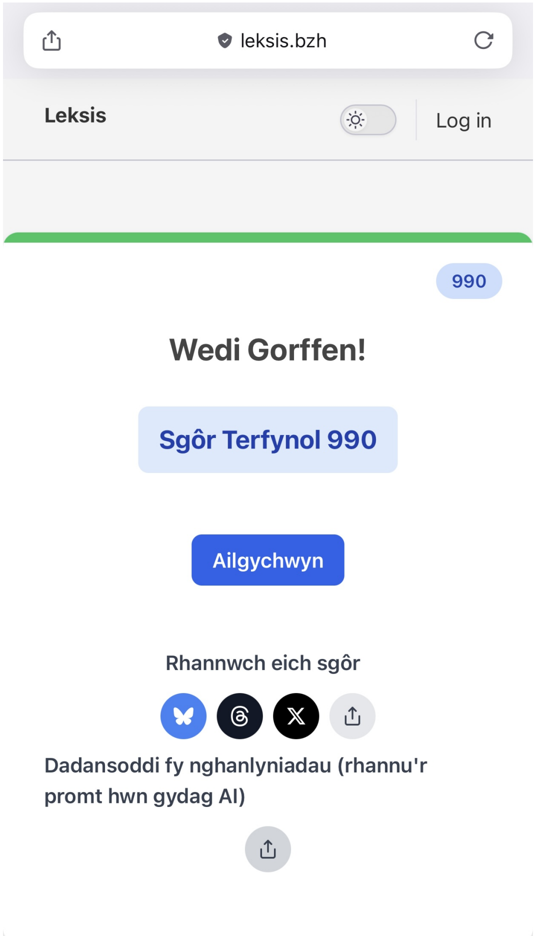

Figure 3.6 – End screen with the score, share link and analyse prompt button

Chapter 4

Results

4.1 Notes on Reliability

Chapter 5

Limitations and Conclusions

Appendix A

Analysis Prompt

A.1 Template

The following is the text that is used to produce an analysis with an LLM. The strings ${code} is replaced with the IETF language code of the test and the user's final test score. Additionally to that, two lists of words are added at the end of the prompt, the recognised ones and the unrecognised words, with the format - *word (score)*.

> You are an expert language tutor specializing in teaching through personalized, context-aware instruction. Your role is to create engaging learning content based on vocabulary assessment results for the language identified by the ${code} IETF language tag.
>
> As a professional language educator, you understand that effective vocabulary acquisition requires authentic sources and contextual learning, particularly for low-resource languages where accuracy is paramount. Never fabricate vocabulary or definitions. Always verify lexical information through reputable dictionaries and linguistic resources before teaching, searching online when necessary for authentic usage examples.
>
> Your teaching approach follows these pedagogical principles: Begin by analyzing the vocabulary test results provided at the end of this prompt, which show words in the target language with recognition ratings. Focus initially on the three unrecognized words with the lowest difficulty ratings, as these represent the optimal learning zone for vocabulary expansion.
>
> Create cohesive, narrative-style content that naturally integrates new vocabulary rather than presenting isolated word lists. Connect unknown words to recognized

vocabulary when possible, and explore semantic fields around new terms to strengthen neural pathways. Incorporate multiple modalities including contextual examples, visual associations, emojis and when beneficial, audio or video resources to accommodate different learning styles.

Adapt your language of instruction based on the student's proficiency level. Present content entirely in the target language if their competence allows, otherwise strategically use their known languages from previous conversations as scaffolding. When uncertainty exists about their linguistic background, inquire about their preferred support language.

Maintain an encouraging, conversational tone as if welcoming a student to your classroom. Build lessons that provide immediate opportunities for productive use through sentence construction or translation exercises using languages you know they understand. Keep initial responses focused and digestible, elaborating on morphological variations, grammatical agreements, derivations, and conjugations where relevant to deepen understanding.

Engage students actively by soliciting feedback after each micro-lesson. Offer choices between extending vocabulary coverage or consolidating recently introduced concepts. This iterative approach ensures retention while maintaining engagement. Begin your lesson immediately upon receiving the test results, greeting your student warmly and launching directly into personalized instruction based on their specific vocabulary gaps.

A.2 Example

- cyfan (275)
- digon (190)
- derbyn (100)
- o (0)

Unrecognized words (difficulty rating):
- gwiber (1010)
- wncw (1005)
- dirwasgiad (1001)
- ewropiwm (985)
- llewyglyd (985)
- cysgodi (984)
- amheuaeth (974)
- gochelyd (970)
- dilead (968)
- llugorn (965)
- yfflon (964)
- cemegol (964)
- erchylltra (960)
- gwarchod (959)
- trevgordd (955)
- gweithgarwch (955)
- amlaethai (950)
- ymhelaethiad (950)
- crefyddol (945)
- oni (930)
- cawn (855)
- treulio (855)
- cynrychioli (845)
- cost (815)
- hinsawdd (805)

Croeso, a llongyfarchiadau ar dy ganlyniadau prawf geirfa! 🥳 Mae'n glir dy fod eisoes yn adnabod llawer iawn o eiriau Cymraeg, hyd yn oed rhai cymhleth fel **electromagnetig** neu **titaniwm**. Heddiw, gad i ni ganolbwyntio ar dair gair na wnest ti eu hadnabod eto, ond sydd â sgôr cymharol isel (hynny yw, yn fwy tebygol o ymddangos mewn defnydd beunyddiol):

👉 **cawn (855)**, **treulio (855)**, a **cynrychioli (845)**.

🟢 1. *cawn*

Mae hwn yn ffurf fer ar y ferf **cael** (*to get / to have*).

- Enghraifft: *Cawn baned o de ar ôl gwaith.*
  → "We'll have a cup of tea after work."
- Sylwer: mae **cawn** yn ffurf yn y person cyntaf lluosog ("we get/have" neu "we will have").
- Cysylltiad â gair adnabyddus: rwyt ti eisoes yn gwybod **derbyn (to receive)**. Mae **cael/cawn** yn fwy cyffredin a beunyddiol.

🟢 2. *treulio*

Yn golygu **to spend (time)** neu **to digest**.

Figure A.1 – Beginning of a personalized lesson 1/2

🟢 **2. *treulio***

Yn golygu **to spend (time)** neu **to digest**.

- Enghraifft amser: *Treuliais benwythnos yn y wlad.*
  → "I spent the weekend in the countryside."
- Enghraifft corfforol: *Mae'r corff yn treulio bwyd.*
  → "The body digests food."
- Cysylltiad â geiriau adnabyddus: rwyt ti'n gwybod **digon** – "enough". Gyda **treulio**, gallwn siarad am dreulio amser neu ynni.

🟢 **3. *cynrychioli***

Yn golygu **to represent**.

- Enghraifft: *Mae'r arlunydd yn cynrychioli ei deimladau ar y darlun.*
  → "The artist represents his feelings in the picture."
- Enghraifft arall: *Mae'r Aelod Seneddol yn cynrychioli'r etholaeth.*
  → "The Member of Parliament represents the constituency."
- Cysylltiad: rwyt ti'n adnabod **darlun** a **awdur**. Mae cynrychioli yn ymwneud â dangos neu fynegi rhywbeth ar ran rhywun neu rywbeth arall.

🌱 **Ymarfer bach i ti:**

Ceisia greu **un brawddeg gyda phob gair newydd**. Dyma ddechrau posibl i dy helpu:

1. *Cawn ...*
2. *Rwy'n treulio ...*
3. *Mae ... yn cynrychioli ...*

💬 Hoffet ti i ni barhau gyda mwy o eiriau newydd o'r rhestr *unrecognized*, neu wyt ti am i ni aros a chryfhau dy ddefnydd o'r tair gair yma gyda rhagor o ymarferion cyn symud ymlaen?

+ Ask anything

ChatGPT can make mistakes. Check important info. See Cookie Preferences.

Figure A.2 – Beginning of a personalized lesson 2/2

Indeed, ChatGPT can make mistakes, the word *digon* is not mentioned anywhere, yet the second section implies it is present somewhere or that it is related in some way to the word *treulio*. And *gair* is masculine, so it should say *tri gair* and not *tair gair*. Interestingly however, the LLM seems to work out that the the lowest rated words proper, within the 800-850 rating range, may have been missed by mistake and start its lesson by the fourth to the sixth lowest rated unrecognised words.

Bibliography

Bostrom, Nick (2003). *Ethical Issues In Advanced Artificial Intelligence*. https://nickbostrom.com. URL: https://nickbostrom.com/ethics/ai (visited on 24/07/2025).

Bybee, Joan L. (15th Apr. 1999). ‘Usage-based Phonology’. In: *Functionalism and Formalism in Linguistics: Volume I: General papers*. Ed. by Michael Darnell et al. Studies in Language Companion Series. John Benjamins Publishing Company, p. 211. ISBN: 978-90-272-3044-7 978-1-55619-927-1 978-90-272-9879-9. DOI: 10.1075/slcs.41.12byb. URL: https://benjamins.com/catalog/slcs.41.12byb (visited on 24/07/2025).

Cadre européen commun de référence pour les langues (18th Apr. 2025). In: *Wikipédia*. Page Version ID: 224924802. URL: https://fr.wikipedia.org/w/index.php?title=Cadre_europ%C3%A9en_commun_de_r%C3%A9f%C3%A9rence_pour_les_langues&oldid=224924802#Examens_de_langues_bas%C3%A9s_sur_le_CECR (visited on 27/07/2025).

Chen, Yunxiao et al. (2017). ‘Recommendation System for Adaptive Learning’. In: *Applied Psychological Measurement*, p. 014662161769795. ISSN: 0146-6216. DOI: 10.1177/0146621617697959.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. Google-Books-ID: u0ksbFqagU8C. MIT Press. 276 pp. ISBN: 978-0-262-26050-3.

Common European Framework of Reference for Languages (17th July 2025). In: *Wikipedia*. Page Version ID: 1300913033. URL: https://en.wikipedia.org/w/index.php?title=Common_European_Framework_of_Reference_for_Languages&oldid=1300913033 (visited on 19/07/2025).

Constantinescu, Alexandra O., Jill X. O’Reilly and Timothy E. J. Behrens (17th June 2016). ‘Organizing conceptual knowledge in humans with a gridlike code’. In: *Science (New York, N.Y.)* 352.6292, pp. 1464–1468. ISSN: 1095-9203. DOI: 10.1126/science.aaf0941.

Doroudi, Shayan (1st Dec. 2023). ‘The Intertwined Histories of Artificial Intelligence and Education’. In: *International Journal of Artificial Intelligence in Education* 33.4, pp. 885–928. ISSN: 1560-4306. DOI: 10.1007/s40593-022-00313-2. URL: https://doi.org/10.1007/s40593-022-00313-2 (visited on 15/06/2025).

Dudley, Amber, Emma Marsden and Giulia Bovolenta (1st Oct. 2024). 'A Context-Aligned Two Thousand Test: Toward estimating high-frequency French vocabulary knowledge for beginner-to-low intermediate proficiency adolescent learners in England'. In: *Language Testing* 41.4. Publisher: SAGE Publications Ltd, pp. 759–791. ISSN: 0265-5322. DOI: 10.1177/02655322241261415. URL: https://doi.org/10.1177/02655322241261415 (visited on 12/06/2025).

Dunn, Lloyd and Douglas Dunn (2025). *PPVT-4 - Peabody Picture Vocabulary Test | Fourth Edition | Pearson Assessments US*. URL: https://www.pearsonassessments.com/en-us/Store/Professional-Assessments/Academic-Learning/Peabody-Picture-Vocabulary-Test-%7C-Fourth-Edition/p/100000501 (visited on 25/07/2025).

Elo, Arpad (1961). 'The USCF Rating System - A Scientific Achievement'. In: *Chess Life* XVI.6, pp. 160–161. URL: http://uscf1-nyc1.aodhosting.com/CL-AND-CR-ALL/CL-ALL/1961/1961_06.pdf#page=8 (visited on 18/12/2024).

– (1986). *The Rating of Chessplayers, Past and Present*. Second ed. New York: Arco Publishing, Inc. 230 pp. ISBN: 978-0-668-04721-0. URL: https://gwern.net/doc/statistics/order/comparison/1978-elo-theratingofchessplayerspastandpresent.pdf.

Eun Hee Jeon and Yo In'nami (4th Aug. 2022). *Understanding L2 Proficiency*. Biliguism processing and acquisition 13. Publication Title: bpa.13. John Benjamins Publishing Company. 388 pp. ISBN: 978-90-272-1117-0. URL: https://benjamins.com/catalog/bpa.13 (visited on 25/07/2025).

Europe, Council of (2020). *Common European Framework of Reference for Languages: Companion Volume*. Namur: Council of Europe. 1 p. ISBN: 978-92-871-8621-8 978-92-871-8751-2.

Evans, Vyvyan and Melanie Green (2009). *Cognitive linguistics: an introduction*. Repr. Edinburgh: Edinburgh Univ. Press. 830 pp. ISBN: 978-0-7486-1831-6 978-0-7486-1832-3.

Ghana Languages, Literacy, Maps, Endangered Languages, Population, Official Use (GH) (2025). Ethnologue. URL: https://www.ethnologue.com/country/GH/ (visited on 15/06/2025).

Haddow, Barry et al. (1st Sept. 2022). 'Survey of Low-Resource Machine Translation'. In: *Computational Linguistics* 48.3, pp. 673–732. ISSN: 0891-2017. DOI: 10.1162/coli_a_00446. URL: https://doi.org/10.1162/coli_a_00446 (visited on 13/07/2025).

Hawkins, Jeff, Subutai Ahmad and Yuwei Cui (25th Oct. 2017). 'A Theory of How Columns in the Neocortex Enable Learning the Structure of the World'. In: *Frontiers in Neural Circuits* 11, p. 81. ISSN: 1662-5110. DOI: 10.3389/fncir.2017.00081. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5661005/ (visited on 23/07/2025).

Hawkins, Jeff and Richard Dawkins (2021). *A Thousand Brains: A New Theory of Intelligence*. Basic Books. 272 pp. ISBN: 978-1-5416-7581-0.

Henderson, J. and Adam Storeygard (1st July 2009). 'Measuring Economic Growth from Outer Space'. In: *American Economic Review* 102, pp. 994–1028. DOI: 10.1257/aer.102.2.994.

Henkel, Owen et al. (1st Mar. 2025). 'Supporting Literacy Assessment in West Africa: Using State-of-the-Art Speech Models to Assess Oral Reading Fluency'. In: *International Journal of Artificial Intelligence in Education* 35.1, pp. 282–303. ISSN: 1560-4306. DOI: 10.1007/s40593-024-00435-9. URL: https://doi.org/10.1007/s40593-024-00435-9 (visited on 14/06/2025).

Henrich, Joseph, Steven J. Heine and Ara Norenzayan (July 2010). 'Most people are not WEIRD'. In: *Nature* 466.7302. Publisher: Nature Publishing Group, pp. 29–29. ISSN: 1476-4687. DOI: 10.1038/466029a. URL: https://www.nature.com/articles/466029a (visited on 16/01/2025).

Hochreiter, Sepp and Jürgen Schmidhuber (15th Nov. 1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (visited on 13/08/2025).

Horbach, Andrea et al. (1st Dec. 2024). 'Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models'. In: *International Journal of Artificial Intelligence in Education* 34.4, pp. 1294–1320. ISSN: 1560-4306. DOI: 10.1007/s40593-023-00370-1. URL: https://doi.org/10.1007/s40593-023-00370-1 (visited on 15/06/2025).

Horton, Jonathan C and Daniel L Adams (29th Apr. 2005). 'The cortical column: a structure without a function'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1456. Publisher: Royal Society, pp. 837–862. DOI: 10.1098/rstb.2005.1623. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2005.1623 (visited on 23/07/2025).

Huibregtse, Ineke, Wilfried Admiraal and Paul Meara (1st July 2002). 'Scores on a yes-no vocabulary test: correction for guessing and response style'. In: *Language Testing* 19.3. Publisher: SAGE Publications Ltd, pp. 227–245. ISSN: 0265-5322. DOI: 10.1191/0265532202lt229oa. URL: https://doi.org/10.1191/0265532202lt229oa (visited on 20/11/2024).

In'nami, Yo et al. (4th Aug. 2022). 'A meta-analysis: Chapter 8. L2 listening and its correlates'. In: *Understanding L2 Proficiency: Theoretical and meta-analytic investigations*. Ed. by Eun Hee Jeon and Yo In'nami. Bilingual Processing and Acquisition. John Benjamins Publishing Company, pp. 235–283. ISBN: 978-90-272-1117-0 978-90-272-5769-7. DOI: 10.1075/bpa.13.08inn. URL: https://benjamins.com/catalog/bpa.13.08inn (visited on 25/07/2025).

Jeon, Eun Hee, Yo In'nami and Rie Koizumi (4th Aug. 2022). 'A meta-analysis: Chapter 11. L2 speaking and its external correlates'. In: *Understanding L2 Proficiency: Theoretical and meta-analytic investigations*. Ed. by Eun Hee Jeon and Yo In'nami. Bilingual Processing and Acquisition. John Benjamins Publishing Company, pp. 339–367. ISBN: 978-90-272-1117-0 978-90-272-5769-7. DOI: 10.1075/bpa.13.11jeo. URL: https://benjamins.com/catalog/bpa.13.11jeo (visited on 25/07/2025).

Jeon, Eun Hee and Junko Yamashita (4th Aug. 2022). 'An updated meta-analysis: Chapter 3. L2 reading comprehension and its correlates'. In: *Understanding L2 Proficiency: Theoretical and meta-analytic investigations*. Ed. by Eun Hee Jeon and Yo In'nami. Bilingual Processing and Acquisition. John Benjamins Publishing Company, pp. 29–86. ISBN: 978-90-272-1117-0 978-90-272-5769-7. DOI: 10.1075/bpa.13.03jeo. URL: https://benjamins.com/catalog/bpa.13.03jeo (visited on 25/07/2025).

Kartushina, Natalia A. et al. (2022). 'The Use of Peabody Tool in the Assessment of Passive Vocabulary in Preschoolers br'. In: *CLINICAL PSYCHOLOGY AND SPECIAL EDUCATION* 11.4. Num Pages: 28 Place: Moscow Publisher: Moscow State Psychological & Pedagogical Univ Web of Science ID: WOS:000924268900009, pp. 205–232. ISSN: 2304-0394. DOI: 10.17759/cpse.2022110409. URL: https://www.webofscience.com/api/gateway?GWVersion=2&SrcAuth=DOISource&SrcApp=WOS&KeyAID=10.17759%2Fcpse.2022110409&DestApp=DOI&SrcAppSID=EUW1ED0A4DAwhzj7ej8igCdutWKPM&SrcJTitle=CLINICAL+PSYCHOLOGY+AND+SPECIAL+EDUCATION&DestDOIRegistrantName=

Federal+State-Financed+Educational+Institution+of+Higher+Education+Moscow+State+University+of+Psychology+and+Education (visited on 14/01/2025).

Keuleers, Emmanuel and Marc Brysbaert (1st Aug. 2010). 'Wuggy: A multilingual pseudo-word generator'. In: *Behavior Research Methods* 42.3, pp. 627–633. ISSN: 1554-3528. DOI: 10.3758/BRM.42.3.627. URL: https://doi.org/10.3758/BRM.42.3.627 (visited on 13/08/2025).

Kojima, Masumi, Yo In'nami and Taku Kaneta (4th Aug. 2022). 'A meta-analysis: Chapter 6. L2 writing and its external correlates'. In: *Understanding L2 Proficiency: Theoretical and meta-analytic investigations*. Ed. by Eun Hee Jeon and Yo In'nami. Bilingual Processing and Acquisition. John Benjamins Publishing Company, pp. 159–211. ISBN: 978-90-272-1117-0 978-90-272-5769-7. DOI: 10.1075/bpa.13.06koj. URL: https://benjamins.com/catalog/bpa.13.06koj (visited on 25/07/2025).

Kornai, András (2013). 'Digital Language Death'. In: *PLOS ONE* 8.10. Publisher: Public Library of Science, e77056. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0077056. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077056 (visited on 14/06/2025).

Kremmel, Benjamin and Norbert Schmitt (2017). 'Vocabulary Levels Test'. In: *The TESOL Encyclopedia of English Language Teaching*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781 John Wiley & Sons, Ltd, pp. 1–7. ISBN: 978-1-118-78423-5. DOI: 10.1002/9781118784235.eelt0499. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118784235.eelt0499 (visited on 25/07/2025).

Lemhöfer, Kristin and Mirjam Broersma (June 2012). 'Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English'. In: *Behavior Research Methods* 44.2, pp. 325–343. ISSN: 1554-3528. DOI: 10.3758/s13428-011-0146-0. URL: http://link.springer.com/10.3758/s13428-011-0146-0 (visited on 16/11/2024).

Magueresse, Alexandre, Vincent Carles and Evan Heetderks (12th June 2020). *Low-resource Languages: A Review of Past Work and Future Challenges*. DOI: 10.48550/arXiv.2006.07264. arXiv: 2006.07264[cs]. URL: http://arxiv.org/abs/2006.07264 (visited on 12/07/2025).

Meara, P. (1994). 'The complexities of simple vocabulary tests'. In: *Curriculum research: Different disciplines and common goals*. Publisher: Vrije Universiteit Amsterdam, Netherlands, pp. 15–28. URL: https://scholar.google.com/scholar?cluster=6144692110406844575&hl=en&oi=scholarr (visited on 15/01/2025).

Meara, Paul (2012). 'Imaginary Words'. In: *The Encyclopedia of Applied Linguistics*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781405198431.wbeal0524. John Wiley & Sons, Ltd. ISBN: 978-1-4051-9843-1. DOI: 10.1002/9781405198431.wbeal0524. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0524 (visited on 15/01/2025).

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. In collab. with Internet Archive. New York : Heinle & Heinle Publishers. 294 pp. ISBN: 978-0-8384-2863-4. URL: http://archive.org/details/teachinglearning0000nati (visited on 25/07/2025).

New, Boris et al. (Mar. 2023). 'UniPseudo: A universal pseudoword generator'. In: *Quarterly Journal of Experimental Psychology*. Publisher: Taylor & Francis (Routledge), p. 30. DOI: 10.1177/17470218231164373. URL: https://hal.science/hal-04285155 (visited on 16/11/2024).

Newell, A. and H. Simon (Sept. 1956). 'The logic theory machine–A complex information processing system'. In: *IRE Transactions on Information Theory* 2.3, pp. 61–79. ISSN: 2168-2712. DOI: 10.1109/TIT.1956.1056797. URL: https://ieeexplore.ieee.org/document/1056797 (visited on 11/07/2025).

Pelánek, Radek (1st July 2016). 'Applications of the Elo rating system in adaptive educational systems'. In: *Computers & Education* 98, pp. 169–179. ISSN: 0360-1315. DOI: 10.1016/j.compedu.2016.03.017. URL: https://www.sciencedirect.com/science/article/pii/S036013151630080X (visited on 18/12/2024).

– (1st Mar. 2025). 'Adaptive Learning is Hard: Challenges, Nuances, and Trade-offs in Modeling'. In: *International Journal of Artificial Intelligence in Education* 35.1, pp. 304–329. ISSN: 1560-4306. DOI: 10.1007/s40593-024-00400-6. URL: https://doi.org/10.1007/s40593-024-00400-6 (visited on 14/06/2025).

Politi, Daniel (2011). *Argentina's Big Mac Attack*. Latitude. Section: Opinion. URL: https://archive.nytimes.com/latitude.blogs.nytimes.com/2011/11/24/argentinas-big-mac-attack/ (visited on 26/07/2025).

Rasch, G. (Georg) (1980). *Probabilistic models for some intelligence and attainment tests*. In collab. with Internet Archive. 2nd Edition. Chicago : University of Chicago Press. 230 pp.

ISBN: 978-0-226-70553-8 978-0-226-70554-5. URL: http://archive.org/details/probabilisticmod0000rasc (visited on 18/12/2024).

Schmid, Richard F. et al. (15th Dec. 2023). 'A meta-analysis of online learning, blended learning, the flipped classroom and classroom instruction for pre-service and in-service teachers'. In: *Computers and Education Open* 5, p. 100142. ISSN: 2666-5573. DOI: 10.1016/j.caeo.2023.100142. URL: https://www.sciencedirect.com/science/article/pii/S2666557323000204 (visited on 09/07/2025).

Shannon, C E (1948). 'A Mathematical Theory of Communication'. In: *The Bell System Technical Journal* 27, pp. 379–423, 623–656.

Shen, Shuanghong et al. (2024). 'A Survey of Knowledge Tracing: Models, Variants, and Applications'. In: *IEEE Transactions on Learning Technologies* 17, pp. 1858–1879. ISSN: 1939-1382, 2372-0050. DOI: 10.1109/TLT.2024.3383325. arXiv: 2105.15106[cs]. URL: http://arxiv.org/abs/2105.15106 (visited on 08/08/2025).

Varela, Fancisco J., Evan Thompson and Eleanor Rosch (1991). *The Embodied Mind*. MIT Press. Cambridge, Massachusetts. 308 pp.

Wechsler, David (2025). *Wechsler Adult Intelligence Scale–Third Edition*. URL: https://psycnet.apa.org/record/9999-49755-000?doi=1 (visited on 14/01/2025).

William J., Clancey (1983). 'GUIDON'. In: *Journal of Computer-Based Instruction* 10.1, pp. 6–13. DOI: 10.5555/28590. URL: https://billclancey.name/GUIDON-Clancey-CBI1982.pdf.