Vocabulary Megastudy for non-LOL languages

2 196 words

1 Introduction

A recurrent critique levelled at cognitive sciences and psycholinguistics is their focus on WEIRD populations (Western, Educated, Industrialized, Rich, Democratic) (Henrich et al. 2010) speaking LOL languages (Literate, Official and with Lots of users) (Dahl 2015; Benítez-Burraco et al. 2024). The English nature of these puns says it all, even research on bilinguism is plaged by this bias, code-switching for instance, may pretend to do research of the 'weaker' languages, but are really an extension of the research on the influence of the 'stronger' languages. Over the years, a growing corpus of crowdsourced megastudies have attempted to gather data for an ever-larger number of languages (Brysbaert 2023) in order to reduce the risk of overgeneralization in psycholinguistics experiment.

Instead of conducing a limited and potentially biased in-lab experiment, the protocol introduced below presents the blueprint of a megastudy aiming at rating the difficulty of words for any given dictionary as well as gathering the subjects' lexical ability score and response times (RT). This experiment is based on the principle "more is better", more people, languages, time and so on will provide more solid data to validate better hypothesis. This paper focuses therefore on the technical aspects of the design. It does so by drawing largely from already established protocols in the field, but optimizing each steps upstream and during the experiment to alleviate the gap in resources most languages around the world experience compared with LOL languages (time, linguistic expertise, people to run preliminary studies etc...). However minimalist the design is made by these constraints, the test itself does not yield minimalist nor approximate results. As a matter of fact, the relative simplicity of the design also allows for evolutions and tailoring to specific needs and incremental improvements. Allowing to test skill variations through time, the dynamic aspect of its underlying framework may be able to find application behind psycholinguistics, in applied linguistics, corpora linguistics and so on.

2 Litterature Review

As studies have found vocabulary knowledge to be a good indicator of general language proficiency (Lemhöfer and Broersma 2012; Meara 1988), this section will give an account of the different approaches used to assess vocabulary mastery through history, comparing their strengths and weaknesses regarding the requirements of building a test scalable for lower-resource language while rivalizing with those used for higher-resource languages.

2.1 How to Assess Vocabulary Mastery

As pointed out by Brysbaert et al. (2016), the measure of vocabulary size will depend on the definition of what a word is (alphabetical type, lemma or word family) and the criteria used to validate that a tested word is knows. Should the word be recognized, understood, translated or described with other words? Should a mastery of all the semantic aspects of a word be displayed for that word to be truly understood? And how to deal with homonyms? This section scrutinizes the different approaches to this problem.

2.1.1 Systematic Sampling of Dictionaries

Hartmann (1946) and Goulden et al. (1990) tried to assess young adults' vocabulary size based on a systematic sampling from dictionaries. In Hartman, the testees were asked to describe the word without time limits (Brysbaert et al. 2016). In the second study, they were asked if they recognized the words. Significantly different result were found: 215 000 by Hartmann and 17 200 by Goulden. Although Goulden's study excluded proper nouns, derived words, and compounds (ibid.), the threshold for word knowledge was also arguably lower, which would indicate a small difference between the ability to recognize words and describe them, at least for these populations. Relying on self reported recognition of the

words is however trustworthy as long as the testees don't have an interest in lying. The best way to make sure that the data gathered by a test are valid is to make it impossible to cheat.

2.1.2 Vocabulary Assessment in Intelligence Tests

In psychology, different strategies have been developed to compare relative vocabulary levels as part of standardized intelligence tests, such as the Wechsler Intelligence Scale for Children (WISC) (Wechsler 1997), or specialized vocabulary tests, like the Peabody Picture Vocabulary Test (PPVT) (Eigsti 2013). Bowles & Salthouse (2008) established that these tests yield similar ranking results regardless of the task, whether it involves identification, association, or production (of definitions or synonyms). These findings are tempered by Hodapp and Gerken (1999) however, probably because productive tasks may involve other, non-linguistic abilities, this idea is supported by the fact that native English and Spanish L2 children could recognize Spanish words faster than adult native Spanish speakers in Meara (1994).
On the one hand, the choice of the task does not appear to impair the consistency of the relative distribution of the results. On the other hand, the tasks used in these tests are not easy to adapt to other language, as they rely on a minimum linguistic expertise, resources such as good synonym dictionaries, and preliminary studies for the calibration, let alone the presence of a trained psychologist to run the test. Even the PPVT, which can be thought of as being easier to translate and administer, cannot be easily adapted to other languages. Kartushina et al. (2022, 219) discovered that words/picture sets translated in Russian for preschoolers did not match their expected difficulty, as those were calibrated for English. In their study, the children spent an average of 20 minutes where they were excepted to complete the test in 5 to 7 minutes.

2.1.3 Tests Based on Lexical Decision Tasks (LDT)

Paul Meara and colleagues have been working on the question of building minimalist vocabulary tests since the 80s (Meara & Jones 1988; Meara 1994). They settled on LDT-based tests where the subjects are presented with words and pseudo-words and have to answer the question "Do you recognize this word?" for each item. This task was first introduced for the study of long-term memory (Meyer & Schvaneveldt 1971) and have shown consistent results for receptive vocabulary tests. The few problems attributed to it rather come from the way the results have been interpreted (Meara 1994), rather than the task itself. The relevance of this task is best explained by Meara himself: "What we appear to have identified is the basic skill on which all other skills depend. If you cannot even recognize that 'tree' is an English word, it is difficult to imagine that you can do anything else with it that might count as vocabulary knowledge" (ibid.). Although Meara initially thought his tests was more relevant for limited vocabulary skills (1992), a more recent study on LexTALE, another implementation of LDT test, confirms the relevance of the simple recognition task to assess advanced English L2 speakers from different backgrounds (Lemhöfer & Broersma 2011).
Another advantage of this testing methodology is its reliance on pseudo-words (Meara 2012). Sets of phonotactically valid words can easily be generated by chaining n-grams from sets of real words, provided that the words are in some form of alphabetical writing system (New et al. 2023). The only apparent limitation in regard to the requirement of this study is the way the items of the have to be selected and curated, first manually, then through preliminary studies, in order to find the most discriminant items.

2.2 Interpreting the Results of the Task

A prominant problem in adapting vocabulary tests to new languages appears to come from the calibration and the calculation of the results. Considering the requirements of this study, the calibration phase also poses a problem of available resources; running a preliminary study for each language to select the items requires time, money and an available representative sample of the speaking populations which is unrealistic for non-LOL languages. This is why the preliminary study must be done as the study goes on, using a light implementation of Item Response Theory (IRT).
Based on the same principles, but adapted to a dynamic context, where the difficulty of the tasks and the level of the subjects vary in time is the Elo Rating System (ERS), independently invented at the same period to rate chess players (Elo 1961), it can be seen as the simplest algorithmic implementation of the one-parameter model (1PL) of IRT, with a Θ of 1.
A study conducted on real data (Wauters et al. 2012) showed that IRT, the proportion of correct answers (of the items) and

the ERS all accuratly predicted the difficulty of the items. However, a study based on simulated data (Pelànek 2016) showed that the proportion of correct answers did not work as well when items are not randomly selected. This is logical, as the goal of adaptative selection is to achieve a given success ratio (ie. 50%, 80%). In the case of vocabulary testing, the large amount of word items requires an adaptative selection to obtain sensible data after a reasonable number trials, especially for beginers. Surprizingly, some degree of randomness would also benefit the ERS (ibid.).

2.3 Summary and discussion

Based on the available literature, it appears that using receptive vocabulary tests is a promising way to assess language proficiency as a whole, even accounting for other aspects of fluency such as grammatical and oral skills, although this correlation is not absolute nor constant (Hajiyeva 2015). It also appeared that the simplest task of recognizing words in LDT tests is the best suited task to effectively measure receptive vocabulary. This seems to be partly due to the fact that recognition is the first stage in any subsequent broader assimilation of the vocabulary, vocabulary around which all the subsequent verbal skills are constituted. Finally, it appeared that a modified version of the ERS is the most relevant way to rate the items (words and non-word) at the same time as the level of the examinees themselves. It also allows the study to be run continuously, without time limit, thus allowing more people to take part in the experiment, which is a non-negligible benefit when non-LOL speakers are not easily available.

3 Protocol

The study takes the form of a website on which users are asked (but not forced) to create an account and add some personal data such as their age, the language they grew up speaking in or the other languages they know. Since any interaction with the helps to calibrate the difficulty of the word items, the platform should also welcome unregistered users inputs. This strategy of accepting anyone's input was implemented for another system using the ERS (pacticeanatomy.com) (Pelánek and Rihák 2016). The few people creating an account may anyway allow the system to deduce these details about the unregistered users.

Then, the users choose which dictionary they want to test their level on. They enter a unlimited number of rounds of five-second LDT, with a front-end script measuring their RT and save it. When they quit the experiment, they can access their Elo rating for that dictionary. Samples of the data can be published periodically to allow scientist to test various hypotheses.

3.1 Generating the Pseudo-Word

Different languages use different letters and follow different phonotactic rules. Every time a dictionary is added to the system, an equivalent dictionary of non-words is generated. A solution by New et al. (2023) proposes to make Markov chains of n-grams of words of the same length. Using 3-grams holds more promising results, yet, as Meara points out (2012) some phonotactically legal words may still be highly unlikely. This is why this protocol adds a layer sorting the pseudo-words by their average Levenshtein distance (Levenshtein 1965) from the sets of words they were generated from. Instead of using a Markov chain, simply chaining all the combinations of 3-grams extracted from sets of words allows more choices to select from. Applying this technique on 600 hundred Welsh words from a Hunspell dictionary, one can expect more than 6000 strings after removing the actual words. Artefacts from the web-sourced dictionary can then be corrected by selecting the 600 "most Welsh" six-characters-long pseudo-words.

3.2 Elo Implementation

One can't know less than nothing of a language, so similarly to the rating implementation in chess, the score of the users and the real words is kept above zero, but the non-words are allowed to go lower, as they only work a penalizes, they also don't bring the subjects rating up when not recognized. There is also a gamification aspect to this. As people can visualize their vocabulary skills, they might want to engage in more reading or linguistically diverse tasks and try to test themselves more often.

Separate scales (rating scales of the items) should be implemented for different kinds of bilingualism to offset the

influence of cognates in the results (Meara et al. 1994).

The adaptability of the items' selection decreases as the user rating increases, balancing efficiency in the beginning and accuracy at a higher level (Pelànek 2016, fig. 3). Another 'Elo trick' would be to initialize the subjects score to zero, but the words and non-words items to 400, to use the initial testing sessions as a collaborative filtering, bringing down to the beginners only words already recognized.

4 Applications and Perspectives

A periodical sampling of the data could be used to categorize groups of speaker/learners in the data using IRT multiparameter models and Bayesian inference. Numerous psycholinguistic hypotheses can be tested from the as they allow to identify L2 learner's progression and stagnation. An interesting research question would be to see if a faster progression in vocabulary score follows the order outlined in the entrenchment hypothesis (Brysbaert 2017) or if, on the contrary, it is learning words "in the wrong order" that allows a faster vocabulary acquisition.

5 Limitations

As the task is dependent on reading skill, while most languages in the world are hardly ever written, a vocal version of the test could be added, by crowdsourcing the voices of the best rated users for a given dictionary. This would also prove valuable for dyslexic and blind users. Another risk is to see people trying to optimize their vocabulary recognition skills, people could start reading books... a way to handle this would be to make it a marker of the test's success.

References

Benítez-Burraco, A., Chen, S., Gil, D., 2024. Typology and the cognitive science of non-WEIRD languages: The role of memory types. A research project. https://doi.org/10.31234/osf.io/b8un7

Bowles, R.P., Salthouse, T.A., 2008. Vocabulary test format and differential relations to age. Psychology and Aging 23, 366–376. https://doi.org/10.1037/0882-7974.23.2.366

Brysbaert, M., 2023. Word megastudy data and eye movement corpora available [WWW Document]. URL https://web.archive.org/web/20230208141005/http://crr.ugent.be/programs-data/megastudy-data-available (accessed 1.16.25).

Brysbaert, M., Lagrou, E., Stevens, M., 2017. Visual word recognition in a second language: A test of the lexical entrenchment hypothesis with lexical decision times. Bilingualism: Language and Cognition 20, 530–548. https://doi.org/10.1017/S1366728916000353

Brysbaert, M., Stevens, M., Mandera, P., Keuleers, E., 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. Front Psychol 7, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Dahl, Ö., 2015. How WEIRD are WALS languages? Presented at the Diversity Linguistics: Retrospect and Prospect - Closing conference of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, Leipzig, p. 22.

Eigsti, I.-M., 2013. Peabody Picture Vocabulary Test, in: Volkmar, F.R. (Ed.), Encyclopedia of Autism Spectrum Disorders. Springer, New York, NY, pp. 2143–2146. https://doi.org/10.1007/978-1-4419-1698-3_531

Elo, A., 1961. The USCF Rating System - A Scientific Achievement. Chess Life XVI, 160–161.

Hajiyeva, K., 2015. Exploring the Relationship between Receptive and Productive Vocabulary Sizes and Their Increased Use by Azerbaijani English Majors. English Language Teaching 8, p31. https://doi.org/10.5539/elt.v8n8p31

Hartmann, G.W., 1946. Further evidence on the unexpected large size of recognition vocabularies among college students. Journal of Educational Psychology 37, 436–439. https://doi.org/10.1037/h0056310

Henrich, J., Heine, S.J., Norenzayan, A., 2010. Most people are not WEIRD. Nature 466, 29–29. https://doi.org/10.1038/466029a

Hodapp, Gerken, 1999. Correlations between Scores for Peabody Picture Vocabulary Test—III and the Wechsler Intelligence Scale for Children—III. https://doi.org/10.2466/pr0.1999.84.3c.1139

Huibregtse, I., Admiraal, W., Meara, P., 2002. Scores on a yes-no vocabulary test: correction for guessing and response style. Language Testing 19, 227–245. https://doi.org/10.1191/0265532202lt229oa

Kartushina, N.A., Oshchepkova, E.S., Almazova, O.V., Bukhalenkova, D.A., 2022. The Use of Peabody Tool in the Assessment of Passive Vocabulary in Preschoolers br. Clin. Psychol. Spec. Educ. 11, 205–232. https://doi.org/10.17759/cpse.2022110409

Lemhöfer, K., Broersma, M., 2012. Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. Behav Res 44, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levenshtein, V., 1965. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady.

Lord, F.M., 1980. Applications of Item Response Theory To Practical Testing Problems, 1st edition. ed. Routledge, Hillsdale, N.J.

Meara, P., 2012. Imaginary Words, in: The Encyclopedia of Applied Linguistics. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405198431.wbeal0524

Meara, P., 1994. The complexities of simple vocabulary tests. Curriculum research: Different disciplines and common goals 15–28.

Meara, P., Jones, G., 1988. Vocabulary size as a placement indicator.

Meara, P., Lightbown, P.M., Halter, R.H., 1994. The Effect of Cognates on the Applicability of YES/NO Vocabulary Tests. The Canadian Modern Language Review 50, 296–311. https://doi.org/10.3138/cmlr.50.2.296

Meyer, D., Schvaneveldt, R., 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of experimental psychology 90, 227–34. https://doi.org/10.1037/h0031564

New, B., Bourgin, J., Barra, J., Pallier, C., 2023. UniPseudo: A universal pseudoword generator. Quarterly Journal of Experimental Psychology 30. https://doi.org/10.1177/17470218231164373

Pelánek, R., 2016. Applications of the Elo rating system in adaptive educational systems. Computers & Education 98, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

Pelánek, R., Rihák, J., 2016. Properties and Applications of Wrong Answers in Online Educational Systems. International Educational Data Mining Society.

Rasch, G. (Georg), 1980. Probabilistic models for some intelligence and attainment tests, 2nd Edition. ed. Chicago: University of Chicago Press.

Steinberg, J., 2000. Frederic Lord, Who Devised Testing Yardstick, Dies at 87. The New York Times.

Wauters, K., Desmet, P., Van Den Noortgate, W., 2012. Item difficulty estimation: An auspicious collaboration between data and judgment. Computers & Education 58, 1183–1193. https://doi.org/10.1016/j.compedu.2011.11.020

Wechsler, D., n.d. Wechsler Adult Intelligence Scale--Third Edition [WWW Document]. URL https://psycnet.apa.org/record/9999-49755-000?doi=1 (accessed 1.14.25).