Language-learning personalized. Engage with real-world examples. Explore culture. Enter at the appropriate difficulty.

EPIC Learning Engineering Fellowship Capstone Project
Danielle Cruz, Kasra Lekan, Sid Rastogi, Hayden Johnson
August 7, 2021

Introduction

Slogan

Your guide to more meaningful language-learning

Mission statement

With a name that comes from the Portuguese word for “guide,” Guia aims to serve as a learner’s guide to more meaningful language-learning. Where other products focus on the what or how of learning a language, Guia zeros in on the why — allowing learners to do exactly what they want with the language: engage with it in context and in culture.

The Guia team

Danielle Cruz

Kasra Lekan

Sid Rastogi

Hayden Johnson

Table of Contents

Background

Language-learning tools are incomplete

Lack of real-world content. While engaging with a target language through synthetic inputs like lessons in a Duolingo module or worksheets in a textbook do help learners gain exposure to the language, they can often seem disconnected from the context needed to actively engage in conversations. Without authentic input, learning materials can often feel artificial or unrelated to a learner's personal interests or to the current events and cultural contexts where the target language (Portuguese) is spoken. For these reasons, it is rare that just taking a language course or playing language games lead to active fluency. In fact, "less than 1 percent of people are actually proficient in a language they studied in a U.S. classroom, even though 93 percent of U.S. high schools were offering foreign language courses as of 2008."[1]

Lack of grammatical context. Most language-learning apps (Duolingo, Babbel) and most language-learning extensions (Toucan, Fluent.co) place their emphasis on memorizing new vocabulary words or mastering basic phrases. For instance, Toucan — a popular Chrome extension — works by randomly substituting words in English with words in the target language via a 1:1 translation (Figure 1).

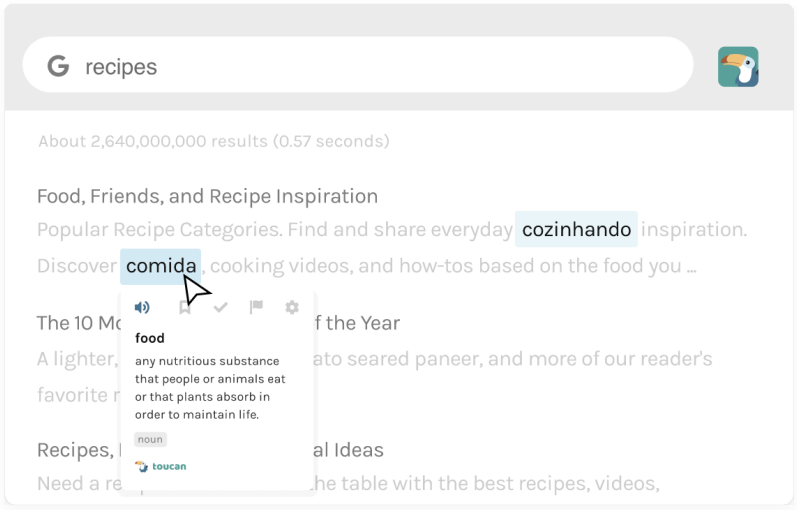

Figure 1. An example of Toucan's word substitution in action.

[1] "America's Lacking Language Skills." https://www.theatlantic.com/education/archive/2015/05/filling-americas-language-education-potholes/392876/.

However, this approach of learning words in isolation doesn't provide learners with the sense of grammatical structure and context to learn what Portuguese looks like in practice — that is, how native speakers employ the language in everyday settings. Without seeing Portuguese in context, learners miss the opportunity to study how to build sentences from scratch.

Loss of intrinsic motivation. According to a study interviewing students in the UK who were learning Swedish, motivation is at the core of successful language education. "How well the learners develop and what kind of progress they make usually depend on their motivation, i.e. how motivated they are to learn a second language and how much time they are willing to spend on learning it." At the same time, "the opportunity to communicate with native speakers of the target language" was one of the most central keys to successful second-language studies for most students.[2]

However, without visible progress in conversational fluency due to the two aforementioned problems with language-learning today, it's no surprise that students of a second language often lose their motivation to continue. Sources based on self-reported data at the Duolingo Incubator and Duolingo Unofficial Golden Owl Hall of Fame suggest that a mere 0.12% of Duolingo users who are studying Portuguese actually finish the entire module.[3]

A new approach to language-learning

Keeping these challenges in mind, we can turn our attention to exploring new approaches to language-learning tools. What if there was a way to de-emphasize synthetic inputs and rote memorization and instead allow users to engage with Portuguese via more authentic and real-world content? Even more, what if we could generate recommendations for this content and use AI / NLP to curate personal collections of articles, stories, and videos that are personalized to a user's proficiency level and interest? With Guia, we venture to explore what a more meaningful and more personalized language-learning experience could look like.

Introducing Guia

Our mission

Guia is a Google Chrome extension that is designed to create a more immersive and personalized language-learning experience by guiding learners through real-world texts at the appropriate level. Guia ultimately aims to empower users with the ability to take language-learning into their own hands and freely engage with Portuguese more authentically.

---

[2] "Students' motivation and attitudes towards learning a second language." http://lnu.diva-portal.org/smash/record.jsf?pid=diva2%3A206523&dswid=2425.

[3] Duolingo Incubator. https://incubator.duolingo.com/.

Figure 2. Guia analyzes webpages as you normally browse the web, providing context about its difficulty level and relevant topics.

Our product

Guia's core product is a Chrome extension interface that can be used to gauge the difficulty of online web pages/articles, as well as recommend articles based on the desired difficulty — all while browsing the web (Figure 2). Currently, the product only supports language learning in Portuguese, but we plan to extend functionality to other world languages as the product develops.

Features

Difficulty estimation. Employing NLP techniques of readability analysis, Guia estimates the difficulty of a Portuguese text before the user even reads it.

After installing Guia and indicating their current level of Portuguese proficiency, users can browse the web like normal and look to Guia to determine whether a Portuguese text might be an appropriate challenge at their given level. Guia employs the Common European Framework of Reference for Languages (CEFR), which ranges anywhere from beginner (A1) to mastery (C2) and helps learners discern whether an article may be at or beyond their proficiency level.[4] Difficulty estimation helps users make more informed decisions about their learning, saving them from potential discouragement from trying to tackle a far too difficult text (Figure 3).

[4] Common European Framework of Reference for Languages. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions.

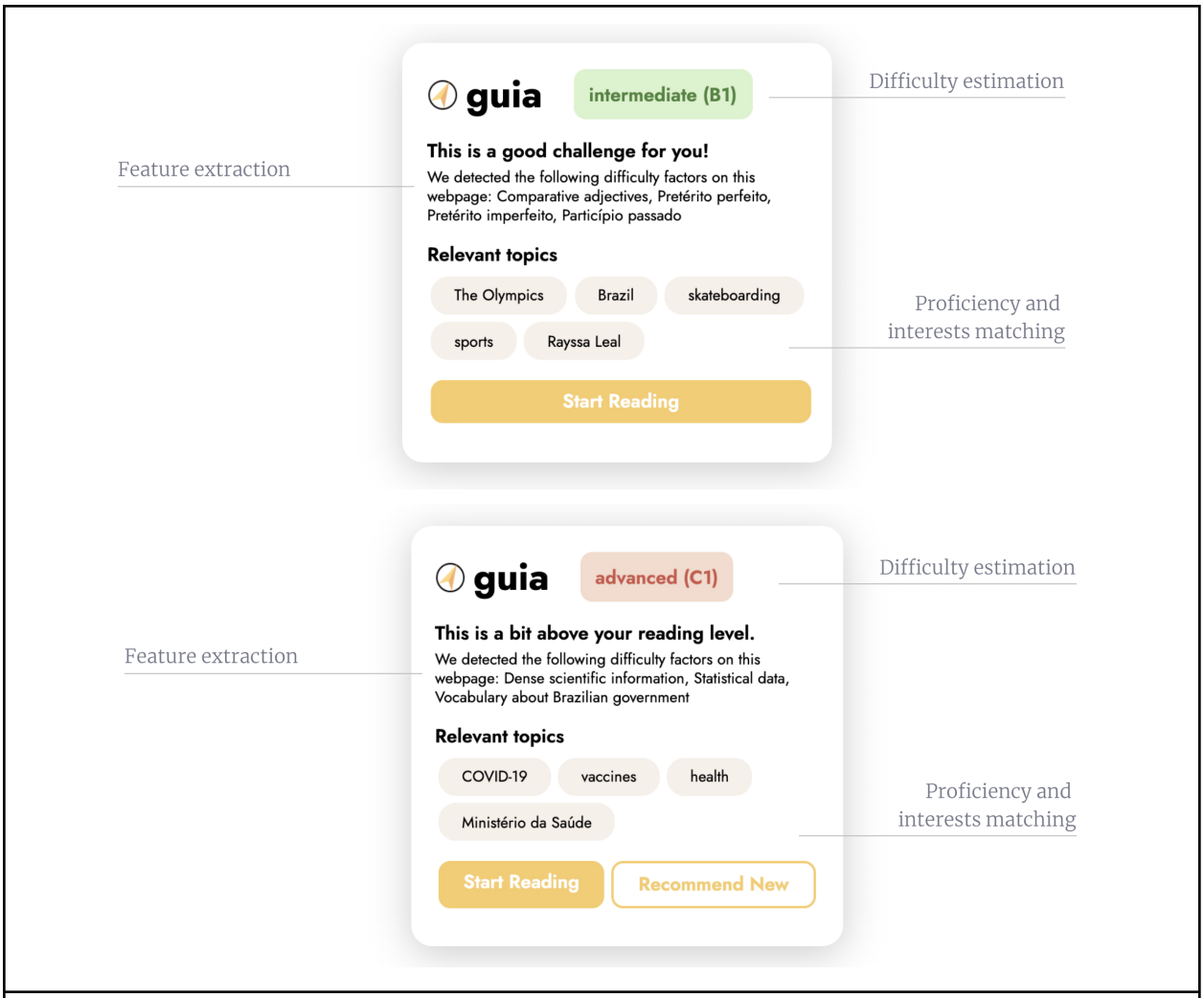

Figure 3. A closer look at the Guia feature set.

Feature extraction. In addition to estimating difficulty, Guia further augments the learning experience by specifically highlighting potential areas of growth or areas of difficulty.

Guia parses the text of the webpage, looking for certain grammatical structures or verb conjugations that might be of particular interest to a user at a given proficiency level (Figure 3). For instance, highlighting past tense verbs could be useful for an intermediate learner while highlighting complex subjunctive verbs could be more useful for an advanced one. By specifically extracting these features of language, Guia helps the user invest more time into actually learning from these areas of growth and less time searching for them.

Proficiency and interests matching. Prioritizing the enjoyment of language-learning in addition to effectiveness, Guia personalizes text recommendations not just by proficiency level, but by interests as well.

After first installing Guia, users can specify what sorts of topics they'd be interested in reading more about — whether they're related to Portuguese or not (Figure 4). Then, using this data, Guia highlights specific Portuguese texts which match these interests, even curating and recommending a collection of texts all tailored to meet these preferences. With Guia, users don't have to go out of their way searching for interesting Portuguese articles or settling for dry textbook content; Guia creates a more personalized experience and helps users integrate language-learning into their everyday browsing on the web. This approach more closely models our lived experiences, where we've observed that most proficient learners of a second language often cite less explicitly educational methods like watching TV shows or reading books in target languages as ways that they improved their fluency — that is, by exploring the interests they cared about.

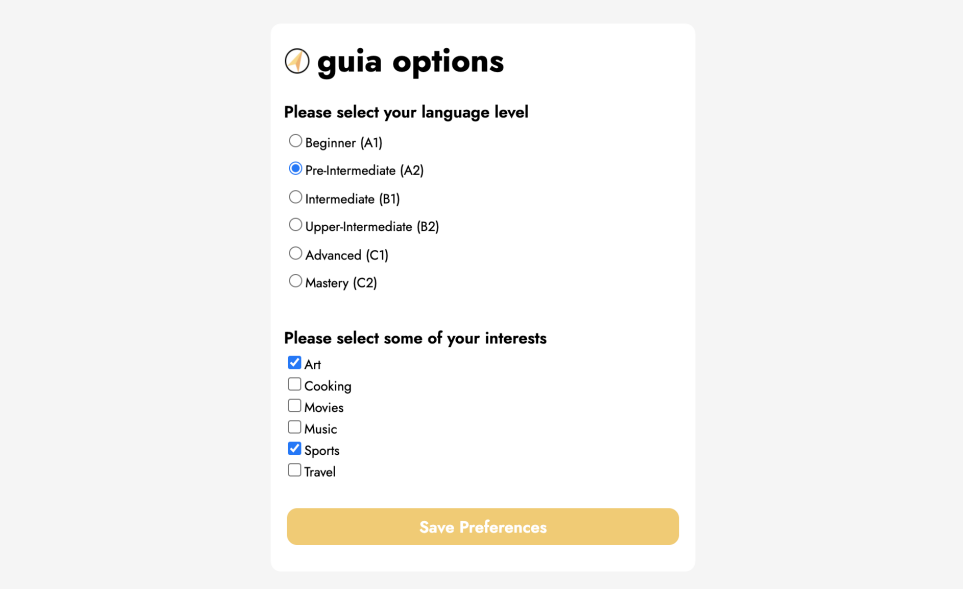

Figure 4. Options modal where users can specify their proficiency level and interests.

Learning benefits

Guia proposes an ideological shift away from learning vocabulary in isolation and toward in-context and in-culture comprehension and learning. In the field of second-language acquisition, this method — known as learning through authentic materials — has shown to be effective not only in improving learner's fluency, but also in improving their motivation and desire to engage with their target language.[5] A number of factors seem to be responsible for this effect.

[5] *Cultural Studies in Foreign Language Education*. https://books.google.com/books?hl=en&lr=&id=rUaz-565duUC&oi=fnd&pg=PP9&dq=language+education&ots=Yf7O7IRLbG&sig=r03xrzEFgCNbuARH8evGXBUnvu8#v=onepage&q&f=false

- Authentic materials contain a number of linguistic features that improve fluency but are not found in synthetic language learning texts (e.g., slang, contractions, cultural phrases).
- The vocabulary used in synthetic media place emphasis on vocabulary that is not relevant to the learner or commonly used in authentic contexts.
- Authentic materials can provide a sense of connection and cultural context that is lacking in synthetic media.[5]

Additionally, it is important to note that sharing authentic materials beyond language barriers has social benefits beyond the learners' language fluency. "The ability to situate texts within the context of real-world cultures/communities has positive social implications."[6] In this sense, Guia is not only a tool for language learning, but for cultural education as well.

Tech overview

Guia's front end is implemented in JavaScript while its difficulty estimation and recommendation functionality is handled via a Python Flask web app. After performing some web scraping and data parsing on the text, we leveraged an open source library called Pylinguistics to perform readability assessments on the Portuguese texts.

Pylinguistics has been shown in computational linguistic studies[7] to be a highly effective NLP tool that is able to contextualize and characterize stylistic elements of real world texts with high accuracy. Under the hood, Pylinguistics uses a Support Vector Machine (SVM) model with a variety of linguistic features to perform these classification tasks. Its models were tested on corpora across different genres spanning a wide range of vocabularies and textual structures, and can thus be classified into different levels of complexity. Three feature selection algorithms commonly used for text categorization (Information Gain, Gain Ratio, and Chi-square) were also used to assess the predictive power of the model's features, and even with a small number features, the model was able to reach an accuracy of up to 97% — all while maintaining a low generalization error to avoid over-fitting. Due to these results and the fact that these models work well in our Portuguese-English use case, we decided to use this package to implement our core functionality.

[6] "Considering the Efficacy of Authentic Materials in Foreign Language Teaching." https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1370&context=lpsc_facpub#page=95

[7] "Pylinguistics : an open source library for readability assessment of texts written in Portuguese." https://www.lume.ufrgs.br/handle/10183/147640

Business plan

Market research

The digital language learning market was valued at $6.3 billion in 2019 and is projected to reach $17.3 billion by 2027; it is expected to grow at a compound annual growth rate (CAGR) of 13.7% from 2020 to 2027.[8] The language learning market as a whole saw a stark decline during the COVID-19 pandemic as people became less able and less interested to travel abroad where they would use new languages. This lull in the market is an optimal moment for Guia to break in.

Using 2019 numbers, the Total Available Market is the entire digital language learning market ($6.3 billion in 2019), the Serviceable Available Market is comprised of all Portuguese 2nd-language learners ($242.9 million), and the Serviceable Obtainable Market is comprised of online Portuguese learners ($100 million).

Target customers

The target customers for Guia are primarily independent language-learners. We are primarily targeting personal language learners rather than professional ones. Professional language learners most often are learning English and use more traditional methods. On average, this group skews younger towards the 18-25 and 26-39 age brackets.

Competitive landscape

Guia has three primary competitors for market share (Figure 5).

1. Gamified language-learning apps (Duolingo, Babbel, Memrise, Busuu)
2. Page translators (Toucan, Fluent.co)
3. Traditional approaches to language learning (classroom, textbooks, courses)

Competitive advantage

Relative to gamified language-learning apps like Duolingo. Instead of playing games to learn vocabulary and basic phrases, Guia offers a more culture-based approach, where users engage with more authentic and real-world inputs. This, in turn, leads to more positive learning outcomes in terms of comprehension, motivation, and socio-cultural understanding, as learners can engage not just with vocabulary and grammar, but also with culture and conversation.

[8] Digital Language Learning Market Forecast to 2027.
https://www.theinsightpartners.com/reports/digital-language-learning-market

Relative to page-translating apps like Toucan and Fluent.co. The benefit of Guia's difficulty estimating and recommending framework is that learning actually takes place in the context of the target language. This differs from sites like Toucan or Fluent.co, where users only see a word from the target language in the context of an English sentence, ultimately preventing learners from understanding what complete sentences look like in practice. Guia's framework eliminates these difficulties by fully situating users in the grammatical context of the target language itself.

Relative to traditional approaches like courses and textbooks. Whereas courses and textbooks often follow a predefined curriculum, Guia affords users a more personalized learning experience where they can more freely engage with the language in the ways they're actually interested in. Through this, Guia helps promote more learner-driven study, where users can more easily integrate language-learning in their everyday web-browsing.

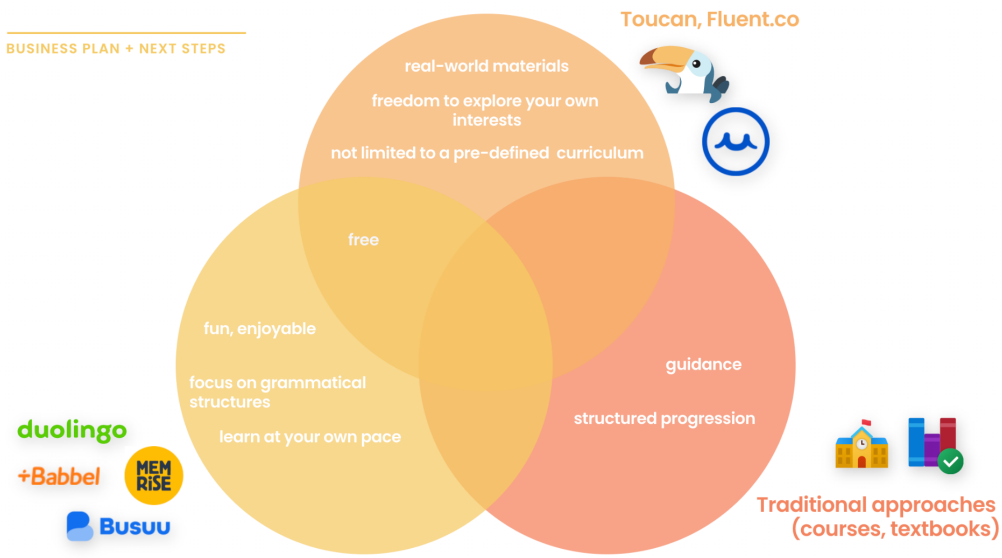

Figure 5A. Venn diagram of Guia's competitors in the language-learning space.

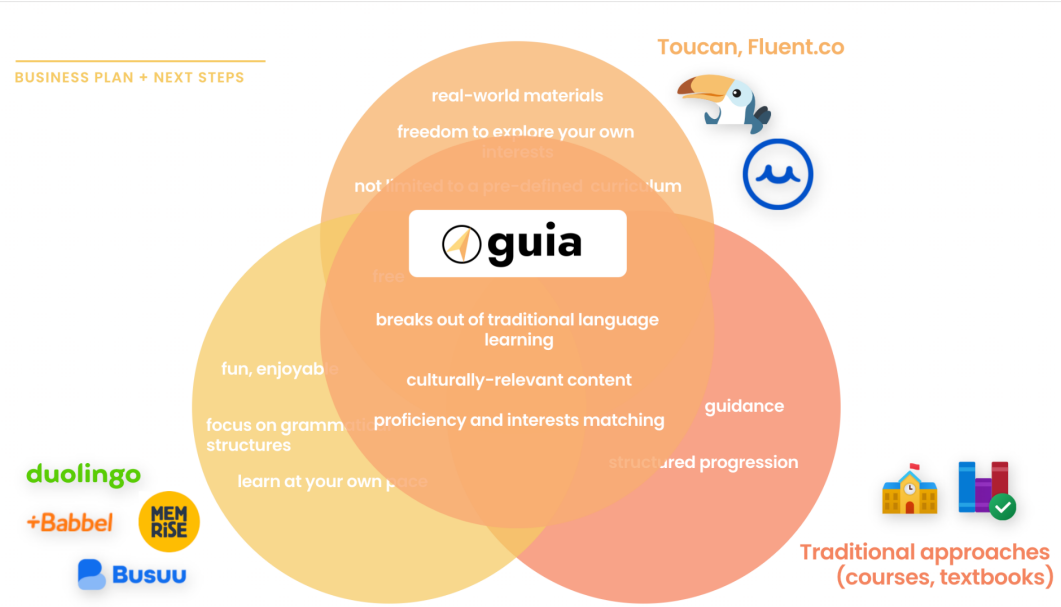

Figure 5B. Venn diagram illustrating how Guia combines the guided, yet real-world aspects of our competitors with added value propositions.

Go-to-market roadmap

1. Develop MVP with scope focused on Portuguese-English learning
2. Conduct market research and user interviews
    a. Determine which features are most important to users
    b. Find how authentically and effectively users are able to engage with Guia
3. Refine the Portuguese version of Guia under a Freemium pricing model
4. Expand to other languages

Freemium model. Revenue for the Guia app will be generated through subscription payments by pro users with access to premium features. This model was chosen over other common revenue models such as voluntary contributions, flat fees, and advertising / data sales because it best suits Guia's potential user base: lifelong language-learners who can provide consistent revenue. Additionally, the app is well suited to the freemium model due to its gambit of features and rapid development, as opposed to a sporadic release of infrequent updates.

As language learners ourselves, we know that not every language learning tool works for all learners, and we want our users to have a chance to try some of Guia's features for free before committing any money.

Alternative monetization strategies

In addition to the Freemium model, we considered a number of alternative strategies for our go-to-market plan. Below are some of the pricing models we discussed, as well as a list of the challenges that we foresee.

Model	**Challenges**
Charging download feeds	Lack of consistent revenue, may leave users disengaged or unmotivated after a while.
Asking for voluntary contributions (e.g. AdBlock)	Needs a large and willing user base to be effective.
Licensing to Duolingo, Babbel, Rosetta Stone, or other language learning platforms looking to build a larger web presence	Need to generate interest within a larger company; however, most of these platforms are more focused on mobile experiences.
Advertise directly on the extension, particularly in industries that have a connection to content (e.g. dining, travel, hospitality)	Would likely clutter and negatively affect the user experience.
Collect user data and share with advertisers	Data privacy concerns

Risk assessment

Lack of name recognition in a crowded space. Online language learning is dominated by big names like Duolingo, and the primary form of language learning in the U.S. is still school or university courses. Additionally, other companies like Toucan already offer immersive language learning opportunities via a Chrome extension.

Portuguese-English language learners are a small market that may provide insufficient growth opportunities. There are only about 25-50 million people who speak Portuguese as a second language, and the number of active new learners is a fraction of that. However, the relatively small size of the market also means there are far fewer competitors for Guia, which currently specializes in Portuguese-English learning. We believe the market is sufficient to provide a springboard for Guia's growth into the Ed-tech scene.

The human desire for easy solutions...and not always effective ones. Guia offers an inherently more difficult approach to learning languages, albeit a more meaningful and effective one via

authentic textual engagement with the target language. This model risks loss of motivation and interest when compared to easier language learning methods such as gamification and direct translation. However, we believe that our method is more beneficial to users' overall learning experience, and intend to do further user research to confirm that this is, in fact, the case.

Need to expand language capabilities gradually as NLP methods advance for foreign languages. NLP research in languages other than English is slow to develop. This could stunt growth opportunities in the future, as we expand to supporting other languages on the platform. However, there has been a lot of increased interest in this area in recent years, so we expect this to become less of a challenge moving forward.

Next steps

Moving forward, there are a number of improvements that we plan to make to future iterations of the product. Most notably, we aim to expand Guia to many other languages in order to broaden our market scope beyond Portuguese-English learners. We also intend to expand from supporting just text-based media like articles and webpages to including other forms of media, like audio, video, and interactive speaking.

Potential future features

- Identifying new vocabulary words and highlighting new verb conjugations in the texts themselves like a personal "helper."
- Asking comprehension questions at the end to see how much you learned from that text and adjusting the difficulty level accordingly.
- Forming online communities of similar-level learners and allowing learners to interact/share media that they find interesting and useful.

Appendix

Estimating the difficulty of texts (readability)

1. "How to Evaluate Text Readability with NLP." https://medium.com/glose-team/how-to-evaluate-text-readability-with-nlp-9c04bd3f46a2.
2. "Text Readability and Intuitive Simplification: A Comparison of Readability Formulas." https://eric.ed.gov/?id=EJ926371.
3. "Automatic Text Difficulty Classifier Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching." https://www.scitepress.org/papers/2015/54283/54283.pdf.
4. "Assisting European Portuguese Teaching: Linguistic Features Extraction and Automatic Readability Classifier." https://link.springer.com/chapter/10.1007/978-3-319-29585-5_5.

NLP work in Portuguese

5. "Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks." https://arxiv.org/abs/1708.06025.
6. "BERTimbau: Pretrained BERT Models for Brazilian Portuguese." https://link.springer.com/chapter/10.1007/978-3-030-61377-8_28 (Source: https://github.com/neuralmind-ai/portuguese-bert.)
7. "Pylinguistics : an open source library for readability assessment of texts written in Portuguese." https://www.lume.ufrgs.br/handle/10183/147640 (Source: https://github.com/vwoloszyn/pylinguistics.)

Language education / language-learning apps

8. *The Notion of "Context" in Language Education.* https://books.google.com/books?hl=en&lr=&id=1cVHAAAAQBAJ&oi=fnd&pg=PA1&dq=language+education&ots=CGSCq76X2E&sig=a3FHXKMaRe5PSYh7P4DPuKVAv7M#v=onepage&q=language%20education&f=false
9. *Cultural Studies in Foreign Language Education.* https://books.google.com/books?hl=en&lr=&id=rUaz-565duUC&oi=fnd&pg=PP9&dq=language+education&ots=Yf7O7IRLbG&sig=r03xrzEFgCNbuARH8evGXBUnvu8#v=onepage&q&f=false.
10. "Authentic materials and authenticity in foreign language learning." https://www.cambridge.org/core/journals/language-teaching/article/abs/authentic-materials-and-authenticity-in-foreign-language-learning/1AE0DE71E691F3D738A8FC9825C07607.
11. "America's Lacking Language Skills." https://www.theatlantic.com/education/archive/2015/05/filling-americas-language-education-potholes/392876/.
12. "Why Don't Americans Know More Foreign Languages?" https://www.babbel.com/en/magazine/american-foreign-language-education.

13. “Half the World is Bilingual. What’s Our Problem?” https://www.washingtonpost.com/local/education/half-the-world-is-bilingual-whats-our-problem/2019/04/24/1c2b0cc2-6625-11e9-a1b6-b29b90efa879_story.html.
14. “Students’ motivation and attitudes towards learning a second language.” http://lnu.diva-portal.org/smash/record.jsf?pid=diva2%3A206523&dswid=2425.
15. “Considering the Efficacy of Authentic Materials in Foreign Language Teaching.” https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=1370&context=lpsc_facpub#page=95.
16. Digital Language Learning Market Forecast to 2027. https://www.theinsightpartners.com/reports/digital-language-learning-market.