

Improving second language reading through visual attention cues to corpus-based patterns

Kate Challis¹, Tom Drusa

¹ Iowa State University, Ames, Iowa, USA

E-mail: kchallis@iastate.edu, t.drusa@gmail.com

Abstract

The patterns inherent to written text often remain opaque to second language learners due to the considerable cognitive demands that reading places on working memory. Learners must attend to the meaning of unknown words, the grammatical structure of sentences, and the meaning of the text as a whole – and this all simultaneously. One solution for helping learners to better attend to existing form, function, and frequency patterns within texts is through systematic visual attention cues, which may offload some of the burden on working memory. Lex-See is a Chrome browser extension that highlights words within a user-supplied text in a variety of shades and colors based on underlying corpus-based data about frequency and word class, and also provides further information about forms, definitions, and phonetic similarity, on mouse-over. Currently Lex-See is optimized for Czech, a less-commonly taught, morphologically rich language with a clear need for easily accessible corpus-informed language learning tools, but it is designed to work with any language for which lemma frequency, form, dictionary, and phonetic data can be supplied.

Keywords: second language acquisition; computer-assisted language learning; corpus-informed software; vocabulary; data driven learning

1. Introduction

The purpose of this design-based research study is to build a Chrome browser extension that provides second language (L2) readers with the visual attention cues to corpus-based information that can improve their attention to top-down reading strategies by offsetting the burden on working memory. In this section, we present key theoretical concepts explored in prior theoretical research related to L2 reading, visual attention cues, and data-driven vocabulary learning.

1.1 Cognitive demands of second language reading

Prior research has shown that the awareness and use of top-down, i.e. global/holistic reading strategies accounts for 52% of the total variance in L2 reading ability (Song, 1999), suggesting that tasks during reading which help readers to attend to information at the discourse level are beneficial. These activities include having a global view of the reading process, making guesses, taking risks, concentrating on the main idea rather than getting sidetracked by trivialities, reading to confirm/refine/reject hypotheses made about the meaning of the text as a whole, summarizing the main ideas, and focusing less on graphophonic and syntactic accuracy than on accurate global understanding. In other words, proficient readers employ strategies that enable them to attend to meaning at the discourse level. Top-down strategies are consistently found

to be better for L2 reading than bottom-up strategies (Brantmeier 2002).

However, L2 reading places a high cognitive burden on working memory, vocabulary recall, and discourse synthesis strategies of all readers, in particular for those who lack reading proficiency in their first language (L1) and those who are at the novice- or beginning-level in their L2 (Kupermann et al., 2022). This cognitive burden makes it especially difficult for learners to attend to top-down learning strategies, even when explicitly trying to do so. One explanation for the high cognitive load which learners experience while reading is an inability to distinguish between information that is important and that which is redundant and unnecessary for learning (Kalyuga and Sweller, 2005). Additionally, the so-called ‘redundancy effect’ occurs when information is presented through multiple simultaneous modalities without allowing for the learner to attend to prioritization of information; researchers have shown that the redundancy effect hinders learning (Mayer et al., 2001; Diao and Sweller, 2007; Liao et al., 2020). However, since reading is a visual task, visual attention cues provided by an outside stimuli can potentially be used to offload some of the burden that L2 reading places on working memory, in particular in self-paced reading tasks where learning outcomes are closely correlated to time spent looking at written text (Schmidt-Weigand et al., 2010).

1.2 Visual attention cues, working memory, and second language reading

Visual attention is a key component of reading because it allows the brain to identify orthographic units during lexical processing. The visual attention span (VAS), which is the maximum number of distinct visual elements that the brain can process simultaneously at a glance (Bosse et al., 2007; Bosse & Valdois, 2009) has been linked to reading performance in both L1 and L2 (Awadh et al., 2016; Lobier, Peyrin, Le Bas, & Valdois, 2012), especially when the stimuli are alphanumeric (Verhallen & Bus, 2011). This research also suggests that readers attend to visual cues while reading.

The connectionist Multi-Trace-Memory reading model (Ans et al., 1998) suggests that there is a correlation between visual attention capacity and reading performance, and that a reduction in VAS is detrimental to familiar word processing (Adelman, Marquis, & Sabatos de Vito, 2010; Grainger et al., 2016). In other words, visual attention “seems to be modulated by the amount of attentional resources available” (Frey & Bosse, 2018; Lobier et al., 2013), suggesting that L2 readers benefit from any strategy that can be used to shift attentional resources towards visual processing.

Some researchers disagree with the idea that visual attention is directly connected to reading performance (Gori et al., 2014, Gori and Facoetti, 2014, Lorusso et al., 2011, Facoetti et al., 2006), while other researchers have found additional non-VAS based evidence to confirm this connection, for example by measuring attentional blink, visual search, and visuospatial attention (Cirino et al., 2022).

1.3 Corpus-based vocabulary learning

Data-driven learning (DDL) is an effective pedagogical approach in which learners are encouraged to independently analyze and explore corpora. Independent, self-motivated reading from authentic texts causes target vocabulary items to become more salient (Chapelle, 2003), but can be further enriched when empirical, corpus-based word frequency and dispersion data are made easily transparent to learners. In a sea of unfamiliar words, it is difficult for learners to make intelligent decisions about which words to prioritize and which to leave for later, and all texts—authentic or contrived—are composed of words. In essence, DDL approaches to L2 reading are implicitly connected to L2 vocabulary building.

Receptive vocabulary refers to the words that a person can understand when encountered in a context, but may not necessarily be able to actively produce in writing or speech independently. Prior research indicates that a receptive vocabulary of approximately 6–9k word families (in English) is needed to achieve 98% text coverage, the amount considered by many researchers to represent an amount of unknown vocabulary that avoids cognitive overload during L2 reading (Nation, 2006; Hu & Nation, 2000; van Zeeland & Schmitt, 2013).

Useful words for L2 learners to prioritize in their learning are those which occur with high frequency and wide dispersion in the language (Gardner & Davies, 2014; Lei & Liu, 2016) since “actual frequency of occurrence is a more reliable indicator of usefulness than pure intuition” (Garnier & Schmitt, 2015). Although language variation depends on its situational context (Gray & Egbert, 2019), there is evidence that the bulk of a language’s highest frequency words, i.e. its function words, are important for expressing information regardless of the subject matter (Matthews & Cheng, 2015). Other research suggests that a small number of words can account for a large number of possible ideas which learners would be likely to either express or encounter (Laufer, 2013; Agernäs, 2015).

Another feature of word “coreness” is its dispersion, referring to how evenly a word is distributed within a certain text or text type. In corpus-based research, data about lexical dispersion is normally accounted for by using an index of dispersion and a predetermined threshold that must be reached in order for a word to be included (Burch, Egbert & Biber, 2016). There is currently no consensus on the best formula for measuring lexical dispersion within a corpus, and this remains a topic of open debate within the field of corpus linguistics (Burch, Egbert & Biber, 2016). A word’s dispersion across a range of different registers and modalities is usually obtained indirectly by designing a corpus to include texts from a range of different registers and modalities (Davies, 2005; Davies & Gardner, 2010; Brezina & Gablasova, 2015). This is particularly important to attend to when the corpus, a sample of language data, is intended to represent a larger language domain. It is well established that linguistic features of texts, including word choice, differ across registers (Biber, 1989), therefore a corpus aiming to serve as a language model must contain individual texts that are

representative of that variety (Sinclair, 1991; Atkins, Clear, & Ostler, 1992; Biber, 1993; Egbert, Biber and Gray, 2022).

It should be noted that successful vocabulary learning via DDL seems to depend greatly on the individual learner (Lee, Warschauer, & Lee, 2020). Researchers suggest that DDL approaches should make learners aware of both the general characteristics of the corpus being used (i.e. what register does the corpus purport to represent) as well as the underlying text processing methods (Gardner, 2007).

1.4 Research Questions

This design-based research study was motivated by the following research questions:

1.4.1 Research Question 1

How can a Chrome browser extension help L2 learners attend to core vocabulary items while reading authentic texts?

1.4.2 Research Question 2

How can a Chrome browser extension help facilitate data-driven learning for L2 learners?

2. Methods

The current study addresses the above research questions by exploring the specific use case of L2 Czech, hence this section presents the corpus-based Czech resources, such as the CGSL (Challis, 2022), Majka (Šmerk, 2007), Wiktionary (Wiktionary), and Euphonometer (Plecháč, 2017) which supplied the Chrome browser extension Lex-See with its underlying data. It should also be noted that certain design principles of Lex-See were specifically informed by linguistic characteristics of Czech, such as the need to create a bank of word forms for each lemmas, and a disregard for the concept of ‘word family’, which would have comprised so many word forms in Czech as to render this concept mostly useless. However, in principle, the methods outlined here could be applied to any language, limited primarily by corpus availability.

2.1 The Czech General Service List (CGSL) for frequency data

The Czech General Service List (CGSL) (Challis, 2022) is a frequency-ranked list of Czech lemmas (lemma + part-of-speech) with high frequency and wide dispersion across written and spoken Czech. It was built following the quantitative methodology developed for the new-GSL by Brezina & Gablasova (2015) through comparing the first 10 000 most highly ranked (by normalized average reduced frequency) lemmas of five different corpora of written Czech, namely: SYN2020, Koditex, csTenTen17, ORALv1, and ORTOFONv2. These corpora were purposefully chosen for their differences in modality, size, and design in an effort to minimize biases implicit to the design of any

single corpus and account for dispersion of words within the language. It should be noted that these five corpora shared many (but not all) of the underlying text processing methods (Hajič et al., 2007; Jelínek, 2008; Straková, Straka & Hajič, 2014; Suchomel, 2018; Kopřivová et al., 2017). Crucially, most corpora in this study use a very similar underlying tagset. It is currently unknown which of these tools or manual editing processes exerted the most influence on the final outcome.

The overlap of the first 10k ranked items of these five corpora were compared pairwise, and as expected, there was a high percent overlap and rank correlation between items from corpora with the same underlying modality, i.e. lemposes from csTenTen17 were more similar in order and rank correlations to those in SYN2020 and Koditex than to ORALv1 and ORTOFONv2. The final CGSL is the union of the intersection of lemposes common to the written corpora and the intersection of lemposes common to the spoken corpora. Final rank assignments on the CGSL were made by 1) ensuring that each lemos in this union had a rank value assigned to it for each list (missing rank values were assigned an arbitrary value of 10,001 as a penalty for not being common to multiple corpora), 2) combining all items on the CGSL-common, CGSL-written, and CGSL-spoken, as illustrated in Figure 1.

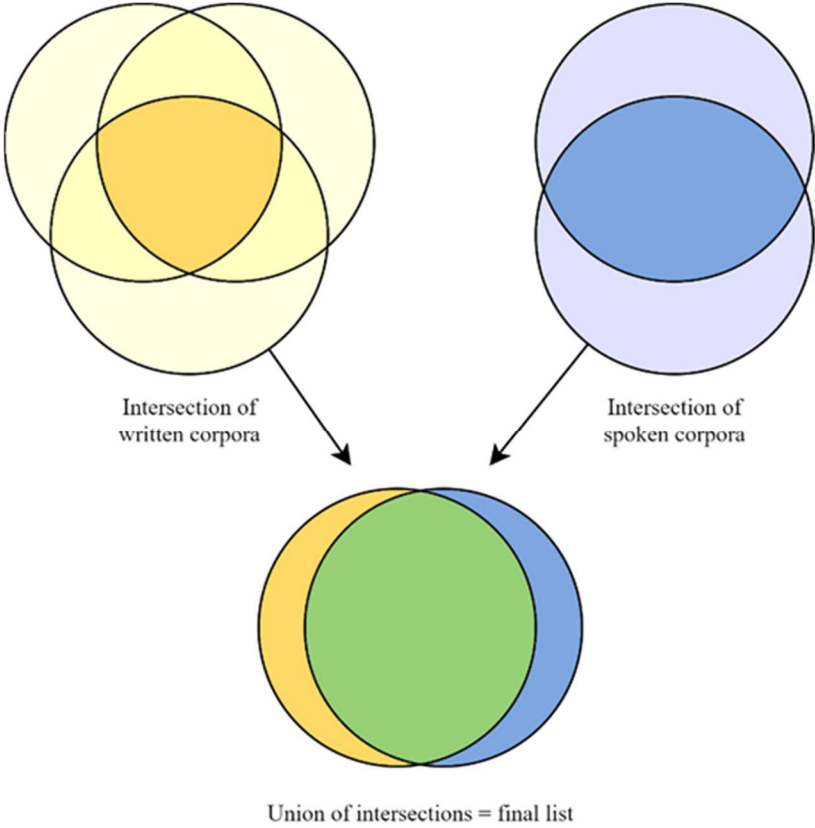


Figure 1. Illustration of CGSL design

Lempos ranks were determined by ordering according to the median, minimum, and product of the ranks across the CGSL-common, CGSL-written, and CGSL-spoken. The median value was useful as a measure of central tendency in the data, but in cases where lemposes shared the same median value, the lempos with the lower minimum value (representing a higher rank, i.e. more frequently occurring) took precedence. Even with both of these measures, there were still a few instances of lemposes “tying” in rank, especially among the highest frequency lemposes. The product of all the scores thus served as a final tiebreaker, since this measure is able to capture effects from the extreme values.

Each item on the CGSL which only occurred in the CGSL-written or CGSL-spoken was labeled as ‘written’ or ‘spoken’, respectively. Thus the final version of the CGSL consists of three main parts: 1) the common lexical core (4,903 lemposes), 2) the lemposes representing spoken Czech (3,048), and 3) the lemposes representing written Czech (2,654). Before the CGSL was compiled, each lempos was manually checked for consistency by a L1 Czech speaker.

2.2 Majka for word form data

Majka (Šmerk, 2007) is a morphological analyzer, a program that can map between the lemma and its associated word forms as well as each of their respective morphological tags. This free tool was designed as a language-agnostic solution to morphological parsing, and is currently available for 15 languages, including Czech, for which it was originally developed. Majka is designed to maximize speed, effectively traversing data precompiled in the form of a finite state automaton – it is therefore language-agnostic, the language and tagset specific data being kept in separate database files.

Lex-See was built by querying Majka’s Czech database to build a list of possible forms for the CGSL lemmata. Of the 10 605 entries, 529 were missing all data, the noun *hospoda* (Eng. ‘pub’) being one of the more curious missing entries, considering that this is a regular, high-frequency word. Apart from several other similarly inexplicable examples, missing entries were generally due to the same issues encountered when building CGSL, which included differences in decisions about the granularity of lemmata, colloquialisms, vulgarisms, interjections, etc. Since the volume was manageable, we were able to fill in the missing forms manually following the patterns and extent produced by Majka based on L1 knowledge.

2.3 Wiktionary for word meaning data

Wiktionary (Wiktionary) is a multilingual crowd-sourced web dictionary of terms, run alongside the well-known Wikipedia encyclopedia. Its openness and semi-structuredness make it suitable for use in various natural language processing tasks, as bots and an application programming interface (API) can be used to read, cross-check, or add data. Entries can typically contain etymology, part-of-speech, word forms depending on

grammatical categories, phonetic transcription, meaning, examples of use, semantically-related terms and translations.

We scraped Wiktionary for the existing entries for CGSL lemmata so that we could provide the translation and possibly an example of its use. Missing data were handled similarly to Majka missing data, i.e. via manual entry by the L1 Czech researcher.

2.4 Euphonometer for pronunciation data

While phonetic data about the base form of lemmas could have also been scraped from Czech Wiktionary, we found that it contained inconsistencies in data formatting and availability. Instead, we were able to use a tool for quantifying euphony of Czech and Slovak texts called Euphonometer (Plecháč, 2017), which features a handy phonetic transcription mode.

In addition to providing this information to the user, we then compiled similar-sounding lemmas using Levenshtein distance as a metric of phonetic similarity. Thus when the user views a lemma, we can present them with a list of the closest possible sound-alikes to be aware of.

While one might consider working with the phonetics of individual forms of the lemma, that would increase the search space by orders of magnitude. Therefore we decided against it, also because similarities between words derived from the same lemma are not especially surprising; we suspect that similarities between forms will also translate into similarities between their lemmas, as these follow a regular pattern.

3. Results

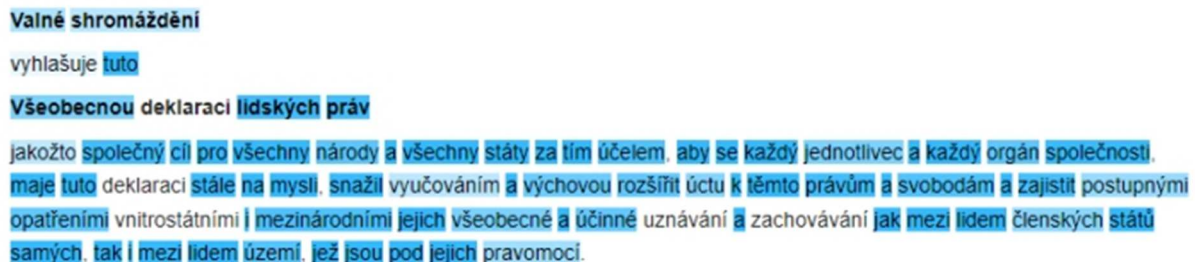
3.1 Lex-See highlighting

The primary design feature of Lex-See is that users have the ability to specify how the background color of a word on a webpage appears, aka its ‘highlighting’. The features which can specify highlighting are whether or not the word is on the CGSL, where (rank-wise) a word falls on the CGSL, what the part-of-speech (POS) of a word is associated with the lowest-rank (i.e. most frequently occurring) item on the CGSL, and if a word is part of CGSL-common, CGSL-written, or CGSL-spoken. This section will now discuss and provide examples for each of these features.

3.1.1 User-defined highlighting

The Lex-See options menu provides the ability for users to specify the color of word highlighting. This can either be a static coloring, or users can specify a range of colors for the lowest and highest rank ends of the scale, which causes words in the middle ranks of the list to appear as gradient shades between the two. For pages with a white background, if blue is chosen as the color to highlight words with the lowest rank (i.e.

the most common words), and white is chosen as the color to highlight words with the highest rank (i.e. the least common words), then all CGSL words appear highlighted on the page in a range of shades of blue, as seen in Figure 2.



Valné shromáždění
vyhláší tuto
Všeobecnou deklaraci lidských práv
jakožto společný cíl pro všechny národy a všechny státy za tím účelem, aby se každý jednotlivec a každý orgán společnosti, máje tuto deklaraci stále na mysli, snažil vyučováním a výchovou rozšířit úctu k těmto právům a svobodám a zajistit postupnými opatřeními vnitrostátními, mezinárodními jejich všeobecné a účinné uznávání a zachovávání jak mezi lidem členských států samých, tak i mezi lidem území, jež jsou pod jejich pravomocí.

Figure 2. Simple highlighting of text

The darker shade of blue is a visual attention cue that intuitively signals to readers which of the words in the text have a relatively stronger importance, which was determined by the frequency and dispersion from the underlying data. Users can choose whether the color distribution follows a linear or logarithmic function, of which the latter differentiates between relative rank differences of words more strongly.

However, if all the high frequency words on a page are highlighted, even in a range of shades, it is almost as ineffective as none being highlighted, since this does not meet the goal of providing differential visual attention cues to words with higher relative importance. Additionally, if every word is highlighted, the redundancy effect is likely to hinder learning. Lex-See solves this problem by allowing users to define the thresholds of the CGSL to either highlight or ignore certain words. For example, if a learner estimates that they already know approximately the first 2k most frequent words on the CGSL, and therefore do not need visual attention cues associated with these words, he or she can specify for highlighting to occur on just the words on the CGSL with rank 2001 or higher.

3.1.2 Part-of-speech highlighting

With lemos as the underlying unit of analysis, items on the CGSL contain at least a small measure of function information, namely the word class, or POS associated with a word. Lex-See allows users to specify highlighting rules based on a word's POS in the CGSL, and in cases where the form can belong to multiple lemmata, the default POS selection is the one associated with the lowest-ranked (i.e. most frequently occurring) lemma. However, duplicate entries are quite infrequent; the CGSL is composed of lemposes, but if we consider plain lemmata, only 73 of them contain multiple POS entries, for example rád_A (Engl: 'happy', adjective) and rád_D (Engl: 'happily', adverb). This means that 99.3% of the items on the CGSL have non-overlapping POS

tags, likely making this feature of particular benefit to L2 learners. An average word then can belong to just 1.012 lemma.

The ability to highlight words based on their most likely word class allows visual attention to be directed differently between function words and lexical words. Lex-See allows users to add multiple layers of highlighting rules based on word class, with the ability to group multiple word classes into the same rule; nouns, verbs, adjectives, and adverbs (which are typically lexical, or open-class words) can be highlighted according to one set of user-specified color, rank threshold, and scaling criterion, while numerals, prepositions, conjunctions, particles/unknown, and interjections (which are typically function, or closed-class words) can be highlighted according to a different set of criteria.

When function words are highlighted in, for example, the same static shade of yellow, it becomes a visual attention cue to L2 readers that helps differentiate them from lexical words. While we do not have sufficient empirical evidence about the difference between how L2 learners perceive, acquire, and use function words, we believe that since these kinds of words are less information-dense and occur with different frequency distributions than lexical words, it makes intuitive sense that visual attention cues can help L2 learners differentiate between these categories. Anecdotally, we have found this feature to be of particular benefit in L2 Czech reading thanks to the variety of function words present in Czech, particularly in written modality.

3.1.3 Highlighting of non-CGSL words

One of the user-defined features for Lex-See highlighting is whether or not the word or any of its associated word forms appears on the CGSL at all. Users can specify highlighting of words that do not occur on the CGSL, however these will always appear in a static color shade due to a lack of ranked frequency information. It turns out that infrequent words end up being so-called ‘keywords’, and typically include named entities, register-specific vocabulary, foreign words, and irregular forms of words, such as archaisms and diminutives (which are abundant in Czech literary texts).

It is useful for L2 readers to have distinct visual attention cues for keywords, since these are the main words which provide the ‘aboutness’ of a text. Anecdotally, it seems that top-down reading strategies are easier for L2 readers to apply to keywords than to unknown high-frequency lexical words. Perhaps this is due to the fact that keywords themselves convey information beyond the word-level; peculiar word choice seems to provide information about an author’s broader stance and message that the choice of common, high-frequency words does not.

3.1.4 Modality-based highlighting

Finally, since the CGSL also contains information about whether a word is common to written, spoken, or both registers of Czech, Lex-See is able to highlight words based on

this feature. This can be especially useful to inform how users can create their own lists.

3.2 Organizing words by meaning and sound

The definitions gathered from Czech Wiktionary (Wiktionary) and phonetic information gathered from Euphonometer (Plecháč, 2017) allow Lex-See users to quickly identify information about word meaning and sound during the process of L2 reading. For all words which occur on the CGSL, a bubble with a word definition appears upon mouse-over, saving users considerable time and effort in dictionary lookup. Additionally, Lex-See allows users to inspect specific words in greater detail via a dialog box containing the example sentences scraped from Wiktionary, as well as concordance lines of all the examples in the target text.

Another visualization feature of Lex-See is the ability to view a bar graph illustrating the counts of all word forms within the target text.

This information is especially useful for L2 Czech learners, who lack intuitions about the form frequency of certain words. When verb conjugations and noun declensions are presented in table form, as is typical in L2 Czech textbooks, it is difficult to prioritize learning one form over another and the redundancy effect takes full force, since low frequency word forms are not as salient as high frequency forms. The purpose of allowing users to explore form frequency distributions through visualization is to quickly convey information about which forms are more likely to be important.

One of the most useful features of Lex-See is the ability for users to explore other high frequency words that sound similar to a target word. This information, based on the Levenshtein distances calculated from the Euphonometer data is also displayed in the word inspection window in order of most to least similar.

3.3 List building, filtering and exporting

3.3.1 Building lists

Perhaps the feature that most closely aligns with principles of DDL is Lex-See's list-building functionality. Users can add any word, whether or not it occurs within the CGSL, to one or more lists. User lists are stored locally and persist between reading sessions, with the maximum number of lists based on compute limits. Users have the ability to name each list as well as to add a note in a text box field for each list item. Furthermore, they can define a combination of modifier keys for individual lists, which can then be used to add words to the particular list when combined with a mouse click. The beauty of being able to build a list by clicking directly on the written text is that the reader is able to minimize the shift in visual attention (i.e. distraction) caused by the act of building a list.

3.3.2 Filtering with lists

Once users have built their own Lex-See wordlists, they can then use them to define highlighting rules in addition to the other criteria. This means that it is possible to use a list of words that deserve extra attention, or the opposite, i.e. a list of words that are not necessary to highlight. Learners can use this feature to build a list that approximates their own personal receptive vocabulary of words not to highlight, which we suspect will be more useful than estimating an arbitrary rank threshold of CGSL words to avoid highlighting.

3.3.3 Exporting lists

Finally, Lex-See allows users to export personal wordlists in .csv, .tsv, and .pdf format, including all corresponding data from Lex-See as well as user created notes. This facilitates easy reuse with, for example, third-party flashcard applications.

3.4 Qualitative user data

This project was originally conceptualized as a way to solve a problem one of the researchers experienced first-hand as an L2 Czech learner. After many persistent attempts to read authentic Czech texts, which were often extremely challenging, the L2 Czech learner decided to turn to a translation of text familiar to her in English, *Harry Potter and the Sorcerer's Stone* (Rowling, 1999). While reading aloud with her L1 Czech collaborator, she was observed to have difficulty in differentiating between which new (to her) words deserved attention and which were relatively unimportant. For example, within a single chapter, the L2 Czech learner ascribed equal importance to learning *naráz*, *čest*, *šum*, *síň*, and *palec* (Engl: simultaneously, honor, noise, hall, inch) as the words *jiskrnýma*, *lektvary*, *zmodrat*, and *škrobeně* (Engl: sparkling, potions, to turn blue, starchily); in a world with limited time and attention capacity, the former set of words would be more beneficial to prioritize because they are more frequent and less specific to the content of *Harry Potter*. This real-world observation provided the original impetus to build both the CGSL and Lex-See.

The next book that the researchers read together was *Dášeňka čili život štěněte* [‘Dášeňka, or The Life of a Puppy’] (Čapek, 1935). This was done by means of the earliest versions of Lex-See, and the process informed many of the design features described in this text. For example, it wasn’t until actually using the tool that the researchers understood the need for the user to be able to specify a threshold of high-frequency words to prevent from being highlighted, and thus avoid the redundancy effect.

The researchers continue to explore L2 Czech reading, most recently with a relatively unknown text called *Valchař se směje aneb tutlanci a pozorníci* [‘The Miller Laughs, or Smugglers and Watchmen’] (Četyna, 1958). Anecdotally, the main character in this text is a fictionalized version of one of the L2 Czech learner’s 18th century ancestors.

Figure 3 illustrates Lex-See highlighting on an excerpt of this text, illustrating how visual attention cues can be used to help distinguish between different categories of words; in this example, function words are yellow, non-CGSL (i.e. ‘keywords’) are red, and CGSL words are shades of blue on a logarithmic scale.

["They're already **in** the pub," guessed the tall one.
They both quickly stood up **and** tried **to make out** **the gable** of the building
which was concealed **by** the trees.
"That's odd." The skinny **man** **shook** his head.
"What's odd?"
"**That** trees **tend to** grow where they shouldn't."
"You're right."]

Figure 3. Screenshot of highlighted words from *Valchař se směje*

One of the more humorous experiences of reading this text was seeing how Lex-See handled the glossary of archaisms found at the end of the book, shown in Figure 4, in which words not on the CGSL are highlighted in red.

práchno — **troud**
připučit — **přimáčknout**
položnica — **šestinedělka**
suchotnica — **vřes**
světadlo — **buková louč**

Figure 4. Screenshot from a glossary of archaisms.

Although the highlighting of this particular set of words did not help with L2 word prioritization in any meaningful way, the L1 Czech reader could still intuit major usage and register differences.

4. Discussion

4.1 Limitations

In this section we present limitations to the current study as well as avenues for future research.

4.1.1 Limitations in the underlying data

A computational tool is only as good as its underlying data, and there is clearly much room for improvement in all the sources of data used to fuel Lex-See. Perhaps most important to note is that it is not yet known the extent to which the items on the CGSL are actually useful to L2 Czech learners. It is assumed based on prior research in L2 vocabulary acquisition that words with high frequency and wide dispersion in a language will be useful, but this has not yet been attested and thus deserves further research.

In order to sound pleasant and make sense to L1 Czech speakers, L2 Czech learners need to be able to correctly produce names in vocative case. However, following the methodology of Brezina and Gablasova (2015), the CGSL contains no proper nouns, which means that an entire grammatical case of Czech is likely to only have limited highlighting potential in Lex-See.

The creation of the CGSL revealed inconsistencies in lemmatization and tagging in the underlying corpus data which were not immediately apparent from extensive review of the respective corpus documentation, and it is not known the extent to which variability in text processing affected the outcome of the content, rank, and modality labeling of items on the final list.

The Wiktionary definitions and examples data has not been attested for accuracy and scope of meaning, which is a known limitation. Additionally, it is not yet known the extent to which the IPA data scraped from Euphonometer reflects prototypical pronunciation of the base form of Czech words; words are known to have different pronunciations in isolation than within the context of other words in a sentence, such as connected speech. This limitation is probably not a primary concern for lower level L2 Czech learners who must first focus on building their receptive vocabulary, but it may become more problematic as proficiency levels increase.

4.1.2 Limitations in Lex-See

At the moment, the Lex-See uses lists of forms for CGSL lemmata in a plain, unoptimized format. This is somewhat ironic considering they were provided by performance-focused Majka based on an efficient storage format. While today's computers are powerful enough to pull this off, there are efficiency gains to be had in using a more specialized format. Going even further, including a morphological analyzer

directly would allow us to provide more options for words not in the CGSL that have not yet been preprocessed.

While working with live web pages can be useful to the user, it also presents many challenges. Website creators can be creative and web pages vary considerably in both structure and looks, yet the inserted user interface elements should work and blend visually with as many of these as possible. In a future version these should be rewritten to leverage modern Web Components features for better isolation.

4.2 Future Research and Conclusion

The most obvious research objective for future research is to design a user experiment to measure the extent to which Lex-See helps L2 learners to 1) attend to top-down reading strategies, and 2) improve vocabulary, grammar, and pronunciation learning. Also, while research in DDL has been shown to be effective for learning, it is not known the extent to which it is more effective than non-corpus based learning methods. Future experimental research could use Lex-See to control for variations in DDL methodology in order to gain a clearer understanding of how DDL compares to traditional classroom approaches in terms of learner outcomes.

One potential future use of Lex-See would be to use its capacity to direct a reader's visual attention to help corpus builders identify and remediate flaws in underlying data sources, such as the CGSL. While reading a text highlighted through Lex-See, we have observed multiple instances of common word forms which are incorrectly highlighted, perhaps due to inconsistencies in corpus tagging, or missing form data in Majka. In principle, Lex-See could be used to uncover similar inconsistencies in corpora of other languages, making it of value not only to L2 learners but for corpus developers and data scientists.

Lex-See could also potentially be used on texts written by L2 learners themselves to measure a variety of linguistic features, including lexical density and complexity. This information might be useful as a way to measure a learner's progress over time, and to build corpus-informed assessments. A Lex-See user study could also measure the extent to which organization strategies of user-created wordlists impacts top-down reading strategies and/or vocabulary learning.

Although adjustments can undoubtedly be made to improve Lex-See, the tool is immediately useful as a vocabulary learning tool. Visual attention cues built into Lex-See help L2 learners attend to word POS, meaning, relative coreness, modality, and patterns in form that may occur within the target text. Additionally, these cues may offset some degree of the burden on working memory during the process of L2 reading, allowing readers to more effectively apply the top-down strategies which are associated with reading proficiency and improvement. Finally, Lex-See follows the model of DDL by providing users with corpus-based information and tooling that can be applied to authentic texts, but while also allowing learners to make their own choices about how

to prioritize, organize, and explore their own L2 reading and learning experience.

In summary, Lex-See is a language agnostic Chrome browser extension tool designed to facilitate L2 reading by means of visual attention cues fueled by corpus-based data. Currently optimized for Czech, we hope to extend the scope of this tool to other languages, in particular those which lack quality corpus-based L2 learning materials.

5. References

- Adelman, J. S., Marquis, S. J., & Sabatos de Vito, S. G. (2010). Letters in words are read simultaneously, not left-to-right. *Psychological Science*, 21(12), pp. 1799–1801. doi: 10.1177/0956797610387442
- Agernäs, E. (2015). *Vocabulary size and type goals in advanced EFL and ESL classrooms. A review of research on lexical threshold, lexical coverage, reading and listening comprehension.*
- Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multi-trace memory model of polysyllabic word reading. *Psychological Review*, 105, pp. 678–723. doi: 10.1037/0033-295X.105.4.678-723
- Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*, 7(1), pp. 1–16.
- Awadh, F. H. R., Phénix, T., Antzaka, A., Lallier, M., Carreiras, M., & Valdois, S. (2016). Cross-language modulation of visual attention span: An Arabic-French-Spanish comparison in skilled adult readers. *Frontiers in Psychology*, 7, p. 307. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779959/>. doi: 10.3389/fpsyg.2016.00307
- Biber, D. (1989). A typology of English texts. *Linguistics*, 27(1).
- Biber, D. (1993). Representativeness in corpus design. *Literary and linguistic computing*, 8(4), pp. 243–257.
- Bosse, M.-L., Tainturier, M. J., & Valdois, S. (2007). Developmental dyslexia: The visual attention span deficit hypothesis. *Cognition*, 104(2), pp. 198–230. doi: 10.1016/j.cognition.2006.05.009
- Bosse, M.-L., & Valdois, S. (2009). Influence of the visual attention span on child reading performance: A cross-sectional study. *Journal of Research in Reading*, 32(2), pp. 230–253. doi: 10.1111/j.1467-9817.2008.01387.x
- Brantmeier, C. (2002). Second language reading strategy research at the secondary and university levels: Variations, disparities, and generalizability. *The Reading Matrix*, 2(3).
- Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1), pp. 1–22.
- Burch, B., Egbert, J., & Biber, D. (2016). Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 3(2), pp. 189–216.
- Čapek, K. (1935). *Dášeňka čili život štěněte.*
- Cirino, P. T., Barnes, M. A., Roberts, G., Miciak, J., & Gioia, A. (2022). Visual

- attention and reading: A test of their relation across paradigms. *Journal of Experimental Child Psychology*, p. 214, 105289.
- Challis, K. (2022). *Is there a core vocabulary for Czech? Introducing the Czech General Service List*. doi:10.13140/RG.2.2.17678.02889.
- Chapelle, C. (2003). *English language learning and technology*.
- Czerepowicka, M. (2021). The structure of a dictionary entry and grammatical properties of multi-word units. *Electronic lexicography in the 21st century (eLex 2021) Post-editing lexicography*, p. 79.
- Četyna, B. (1958). *Valchař se směje aneb tutlanci a pozorníci*. Krajské nakladatelství v Ostravě.
- Čermák, F., & Kren, M. (2011). *A frequency dictionary of Czech: core vocabulary for learners*. Routledge.
- Davies, A. (2005). *An introduction to applied linguistics*.
- Davies, M., & Gardner, D. (2010). *Word frequency list of American English*.
- Diao, Y., & Sweller, J. (2007). Redundancy in foreign language reading comprehension instruction: Concurrent written and spoken presentations. *Learning and instruction*, 17(1), pp. 78–88.
- Egbert, J., Biber, D., & Gray, B. (2022). *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge University Press.
- Facoetti, A., Zorzi, M., Cestnick, L., Lorusso, M. L., Molteni, M., Paganoni, P., ... & Mascetti, G. G. (2006). The relationship between visuo-spatial attention and nonword reading in developmental dyslexia. *Cognitive neuropsychology*, 23(6), pp. 841–855.
- Frey, A., & Bosse, M. L. (2018). Perceptual span, visual span, and visual attention span: Three potential ways to quantify limits on visual processing during reading. *Visual Cognition*, 26(6), pp. 412–429.
- Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied linguistics*, 28(2), pp. 241–265.
- Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied linguistics*, 35(3), pp. 305–327.
- Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, 19(6), pp. 645–666.
- Gori, S., & Facoetti, A. (2014). Perceptual learning as a possible new approach for remediation and prevention of developmental dyslexia. *Vision research*, 99, pp. 78–87.
- Gori, S., Cecchini, P., Bigoni, A., Molteni, M., & Facoetti, A. (2014). Magnocellular-dorsal pathway and sub-lexical route in developmental dyslexia. *Frontiers in human neuroscience*, 8, p. 460.
- Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A vision of reading. *Trends in Cognitive Sciences*, 20(3), pp. 171–179. doi: 10.1016/j.tics.2015.12.008
- Gray, B., & Egbert, J. (2019). Register and register variation. *Register Studies*, 1(1),

- pp. 1–9.
- Hajič, J., Votrubec, J., Krbec, P., & Květoň, P. (2007, June). The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the workshop on Balto-Slavonic natural language processing*, pp. 67–74.
- Hu, M., & Nation, I.S.P. (2000) Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), pp. 403-430.
- Jelínek, T. (2008). Nové značkování v Českém národním korpusu. *Naše řeč*, (1), pp. 13–20. [New Tagging in the Czech National Corpus].
- Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, 53(3), pp. 83–93.
- Karlík, P. (2012). *Příruční mluvnice češtiny*. Lidové noviny. [Handbook of Czech]
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, pp. 7–36.
- Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., Křen, M. (2017). *ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017*. Ústav Českého národního korpusu FF UK. Praha. Accessible at <http://www.korpus.cz> [ORAL: A Corpus of Informal Spoken Czech]
- Kopřivová, M., Komrsková, Z., Lukeš, D., & Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus–gramatika–axiologie*, 15, 47-67. [The ORAL Corpus: Assembly, Lemmatization and Morphological Tagging]
- Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In *International conference on analysis of images, social networks and texts*, pp. 320–332. Springer, Cham.
- Kuperman, V., Siegelman, N., Schroeder, S., Acartürk, C., Alexeeva, S., Amenta, S., ... & Usal, K. A. (2022). Text reading in English as a second language: Evidence from the Multilingual Eye-Movements Corpus. *Studies in Second Language Acquisition*, pp. 1–35.
- Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *Tesol Quarterly*, 47(4), pp. 867–872.
- Lee, H., Warschauer, M., & Lee, J. H. (2020). Toward the establishment of a data-driven learning model: Role of learner factors in corpus-based second language vocabulary learning. *The Modern Language Journal*, 104(2), pp. 345–362.
- Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for academic purposes*, 22, pp. 42–53.
- Liao, S., Kruger, J. L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, 33, pp. 70–98.
- Lobier, M., Dubois, M., & Valdois, S. (2013). The role of visual processing speed in reading speed development. *PLoS ONE*, 8, p. 4. Retrieved from

- <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058097>
- Lobier, M., Peyrin, C., Le Bas, J. F., & Valdois, S. (2012). Pre-orthographic character string processing and parietal cortex: A role for visual attention in reading? *Neuropsychologia*, 50(9), pp. 2195–2204. doi: 10.1016/j.neuropsychologia.2012.05.023
- Lorusso, M. L., Facoetti, A., & Bakker, D. J. (2011). Neuropsychological treatment of dyslexia: does type of treatment matter? *Journal of learning disabilities*, 44(2), pp. 136–149.
- Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, 52, pp. 1–13.
- Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of educational psychology*, 93(1), p. 187.
- Nation, I. (2006). How large a vocabulary is needed for reading and listening?. *Canadian modern language review*, 63(1), pp. 59–82.
- Plecháč, P. (2017). *Euphonometer 2.0*. Prague: Institute of Czech Literature, CAS. Available at: <http://versologie.cz>.
- Rowling, J. K. (1999). *Harry Potter and the Sorcerer's Stone*. Scholastic.
- Sanosi, A. B. (2018). The effect of Quizlet on vocabulary acquisition. *Asian Journal of Education and e-learning*, 6(4).
- Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learning and instruction*, 20(2), pp. 100–110.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Song, M. J. (1999). Reading strategies and second language reading ability: The magnitude of the relationship. *English Teaching*, 54(3), pp. 73–95.
- Straková, J., Straka, M., & Hajic, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 13–18.
- Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In *RASLAN*, pp. 111–123.
- Šmerk, P. (2007). Fast Morphological Analysis of Czech. In Petr Sojka and Aleš Horák *Proceedings of Third Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. Brno: Masaryk University, 2007. pp. 13–16. ISBN 978-80-210-5048-8. Available at: <https://nlp.fi.muni.cz/ma/>
- Těšitelová, M. (1987). *O češtině v číslech*. Academia.
- Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied Linguistics*, 34(4), pp. 457–479.
- Verhallen, M. J., & Bus, A. G. (2011). Young second language learners' visual attention to illustrations in storybooks. *Journal of Early Childhood Literacy*, 11(4), pp. 480–500.

- Wiktionary: The free dictionary*. Accessed at: <https://wiktionary.org>.
- Wright, B. A. (2016). Transforming vocabulary learning with Quizlet. *Transformation in language education*. Tokyo: JALT, pp. 436–440.
- Ziková, M. (2017). Supletivismus in Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), *CzechEncy - Nový encyklopedický slovník češtiny*. Accessed at: <https://www.czechency.org/slovník/SUPLETIVISMUS>