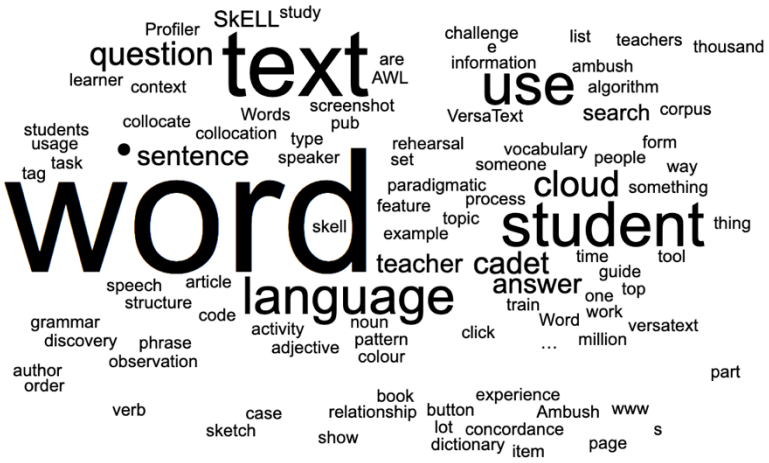
Two web-based tools for learning language from language

James Thomas

April 2020

This article introduces two online tools for learning language from language, with a special focus on observing how words and phrases combine to create meaning, and on how teachers can guide their students to learn language through such observations. The article draws on material that the author is currently writing for a book to be entitled *Grammar, Vocabulary and Everything in Between.*

Here is a word cloud containing 100 words. Being generated from a text, a Word cloud reveals the text's key ideas simply by assembling its most frequent key words.



At first glance, our eyes draw a triangle connecting *text, word* and *student*. This prompts us to assume that the text will do the same. Our triangle morphs as we take in *question, use, cloud* and *language.* There are easy relationships between some of these seven words, and the less easy ones pique our interest. We wonder why *cloud* is in the top set, and why the most prominent verb is *use*: is the text going to tell us *who uses what*?

On closer inspection, several topics weave through the cloud like trails in a forest. There are words related to language such as *vocabulary, grammar, sentence*. And

to education: *teacher, learner, study*. And to computers: *click, algorithm, code*. And there are all-purpose words like *case, set, part*.

To discover the message that these interweaving topic trails convey, it is necessary to read this article to its end. On your journey, you may be *ambushed* by *cadets*, the two most mysterious items in the word cloud.

Language learning

Nothing improves learners' competence in the four skills more than building a bigger and better vocabulary. There are many ways that teachers can help their students in this endeavour but there is a general belief that students will learn vocabulary from a few tried and tested rote procedures such as memorising bilingual lists and completing gapfills. These procedures are used regardless of the students' age, their language level, their interests, their motivation, their goals or their ability to make their own discoveries about language. Such memorising by repeating over and over is referred to as *maintenance rehearsal*. This is a very ineffective way to retain information. People cycle the target items through the working memory, never processing it more deeply, as Richard Roberts and Roger Kreuz, two researchers at MIT, have recently demonstrated (2015). They go on to claim that in order to grow a bigger and better vocabulary, students need to engage in *elaborative rehearsal*, which involves personalisation, paraphrase, focus on meaning and the relationships between words. Decades before, the renowned psychologist, Jean Piaget (1896-1980), had claimed:

> Each time one prematurely teaches a child something he could have discovered for himself, the child is kept from inventing it and consequently from understanding it completely.

In discovery learning, the teacher presents students with pieces of language in one form or another and some language questions and challenges. The teacher then guides the students towards possible answers. The students learn through processes associated with discovery and inquiry by observing, inferring, formulating hypotheses, predicting and communicating (Richards and Schmidt 2002) which are close relations of both Task-based Language Teaching (Ellis 2003) and Bloom's Taxonomy (Anderson 2001).

There are two broad categories of relationships between words in a text, namely paradigmatic and syntagmatic. Syntagmatic relationships include chunks, fixed phrases, collocation, colligation, complementation and word templates. These are considered horizontal because they are linear combinations that form texts.

Paradigmatic relationships include lexical sets, synonyms, antonyms, meronyms and hypernyms. These are considered vertical because they offer alternatives, only one of which is chosen at any moment when creating text. Every one of these linguistic features begets language tasks.

These relationships are widely described in books on semantics and lexicography from various theoretical perspectives. Some dictionaries provide phrases that words are used in, some synonyms, pronunciation and word grammar, such as the use of prepositions and complementation (e.g., + -ing, + to inf). Grammar books provide general rules about essential language structures ranging from morphology to sentence structure. The two books of grammar patterns by Francis et al. (1996 and 1998) do provide information about the complementation of words but the books are out of print and their website does not work properly. Thus, there are no readily accessible lists of words that provide these relationships. Another feature of dictionaries, grammar books and student workbooks is the sentences that exemplify or test the current point. Until recently, these were mostly made up by the books' authors and usually unremittingly bland. Given that millions of sentences are spoken and written every day in every conceivable context, it is disingenuous of these authors to choose to write their own instead of providing learners with snippets of language that were produced in authentic communicative contexts.

In addition to the lack of lists which textbook authors and teachers could use as the basis for tasks, there is also the issue of language quandaries. Whenever there are questions, there is uncertainty. And in speech, this affects fluency. Some questions have more than one right answer. There are times when we are not sure how to pronounce a word, which words typically keep company with a word, if the word has certain connotations in certain contexts, if a synonym would be better in the given context, what prefixes to use to make an opposite, what the noun form of the verb is. Here are some specific questions whose answers most native speakers would agree on.

- Do we say *Is it false or true*, or *Is it true or false*? Or both?
- Do we say *He is interesting for ...?* And if so, does it mean the same as: *He is interested in …*?
- Do we use the infinitive with to after *let*? e.g., *Let's to do this.*
- Is *thunk* an alternative past participle of *think*?
- Do we say *play good* in English?
- How are words like *whereabouts* and *underneath* used?
- What comes after, *it's as if …* in a sentence?
- Does *lecture someone* mean the same as *give a lecture*?
- Does English have a form of question tag that is not contracted, e.g., *have you not?* instead of *haven't you?*

- Is *they* normal usage in this sentence? *In order for someone to say something, they need a verb and a noun.*

These are among thousands of questions that dictionaries and grammar books cannot answer or explain fully. Neither can they provide many example sentences.

There is still a need for other sources of information about language. There is often the possibility of asking advanced non-native speakers such as teachers and translators and even native speakers. Some questions have a simple answer. But those which permit interpretation will get a range of answers that depend on the person's idiolect which may not always be representative.

An alternative is to ask a thousand native speakers, from whose thousand answers we can derive a **consensus**. In statistics, this is the *mode*. In language, these are **patterns of normal usage**, which is precisely what language learners need. The most practical and objective way to ask thousands of native speakers is to search a database of authentic language that has been sampled and stored. Such a database is a **corpus**.

SkELL

SkELL is a web-based tool that contains both a **corpus** and its own search engine. It lives in the cloud. The name stands for *Sketch Engine for Language Learning*.

A *corpus* is a body of texts, i.e., a large collection of spoken and/or written texts that were created in authentic acts of communication rather than being written for teaching purposes. A corpus is therefore a record of how people have used words, phrases and structures. And it is a record of what people have said about people, places and things.

For searches, go to skell.sketchengine.co.uk. For information about SkELL, go to www.sketchengine.co.uk/skell.

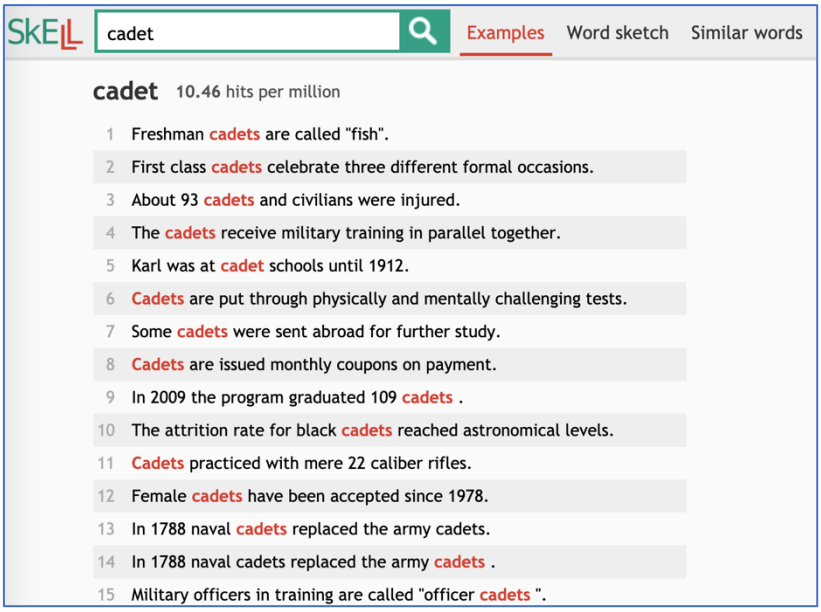
To find out how thousands of native speakers used a word or phrase, type it into the search field: see the screenshot below. Then click on one of the three formats that SkELL's search engine offers:

- Forty student-friendly **Sentences**.
- A collocation table called a **Word sketch**.
- A set of semantically and/or grammatically **Similar Words.**

Sentence examples

A search for sentences generates a page of forty sentences. A page of corpus-derived sentences is called a **concordance**, which is why corpus search engines are usually called **concordancers**. This screenshot of *cadet* shows the first fifteen.

The corpus consists of over a billion words in thousands of texts that were drawn from the internet. The sentences which result from an Examples search all come from different texts. The page, therefore, is not meant to be an aesthetic reading experience. Furthermore, it is not usually necessary to read whole sentences to answer questions such as those above where specific items are being sought. Skimming the co-text of the search item, which is red, usually answers the question.



SkELL says that *cadet* occurs 10.46 hits per million. This indicates how frequent a word is in this very large sample of English. We can calculate that there must be about 10,460 sentences containing the word *cadet.* For the sake of comparison,

- time 1598.42 hits per million
- experience 321.58 hits per million
- soldier 88.34 hits per million
- advantage 82.97 hits per million
- disadvantage 12.02 hits per million
- recruit 31.39 hits per million
- tactical 15.08 hits per million
- dictionary 14.19 hits per million
- intestines 1.51 hits per million
- tactician 0.49 hits per million

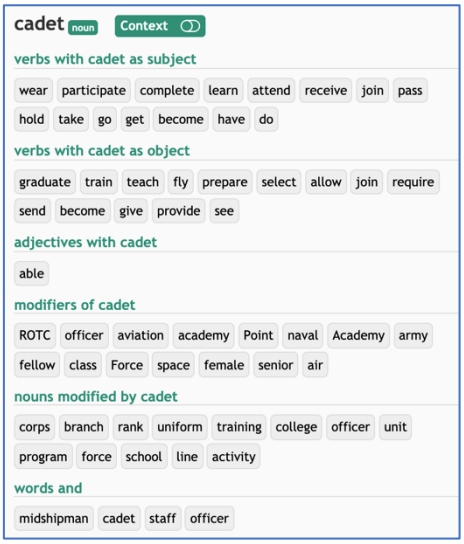
It is interesting to note that a set of sentences that all contain the same word or phrase usually has other recurring features – most often collocates and topic related words. In the *cadet* examples, the first sentence explains a nickname. There are combinations such as *cadet school, female cadets, naval cadets, officer cadets.* There are topic related words such as *first class, civilians, military training, study, program, attrition rate, rifles, army.*

This is not just interesting: all of these features beget observation tasks that teachers can set and work through with their students. They learn some specific vocabulary from context through observation. And it draws their attention to an essential paradigmatic feature of language. This is guided discovery *par excellence*.

The students need to record their observations as any data collector does, meaningfully and creatively, in their learning journals. And they need opportunities to use their findings.

Word sketch

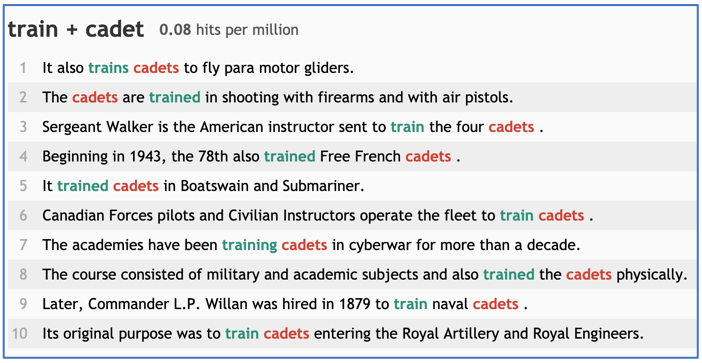
What do cadets do? Look at the verbs which have *cadet* as subject in this screenshot of the word sketch of *cadet*. What happens to cadets? Look at the verbs which have *cadet* as object. How are cadets described? Look at the adjectives. What types of cadets are there? Look at the modifiers.



A collocation is the kernel of a clause, which itself, is a chunk of information. In order for someone to say something about something, they need a verb and a noun. This is what makes word sketches such a valuable resource for vocabulary learning in the service of speaking and writing.

The kernels of clauses are surrounded by other patterns of normal usage which students can explore by clicking on the collocates. This opens the sentences page with the collocation.

Even though the search is for *train* in the context of *cadet*, SkELL searches for the **lemmas** of all the search items. In this example, we see *train, trains, trained, training*. Further down the concordance page than this screenshot shows, we see not only *cadets* but *cadet.* These are **word forms** and their set is a lemma. The lemmatizing that the software does makes possible word sketches and concordance pages such as these.



This screenshot shows the first ten of the forty sentences containing *cadet* + *train*. The words in the immediate context of the search items are referred to as co-text, and this can be read as a chunk of information. We can ask students which lines contain:

- someone trains cadets (3, 4, 5, 9, 10)
- someone trains cadets to do something (1)
- someone trains cadets in (doing) something (2, 7)
- how cadets are trained (8).

Students need to learn such paradigmatic relationships of words in order to have them in their active vocabulary. It impacts on their fluency. The guided discovery questioning and answering that leads students to their conclusions is part of a student's knowledge creation. This is very different from learning vocabulary from lists.

The word sketch of a common word contains a lot of words and many of them may be unfamiliar. Teachers can encourage students to use dictionaries and they can help them prioritise the words by asking which words in the word sketch they already know, if any of the known words are especially interesting in the context of *cadet*, and if they have seen some of the interesting words in other contexts. The students may then decide which of the new, unknown words look useful and add them to their learning journals.

Similar Words

Words can be similar in many ways. Even opposites are similar in some ways.

When we are writing and speaking, it sometimes happens that we cannot remember the right word, even though it feels like it's on the tip of our tongue. The words that come to mind are usually similar in some way. We may also have the feeling that the word we are about to use is not quite right. We feel that we know a better word. Similar Words provide a list which may well contain the sought-after word.

SkELL's sets of Similar Words make good lexical sets, which are useful in developing vocabulary. For the students to use words convincingly, it is necessary that they pronounce them well. The teacher can use a selection of words from a lexical set in a Syllable Stress Table activity (see www.versatile.pub/sst). Since SkELL creates the sets by collecting words that are in the same collocations, students can be challenged to guess verbs and adjectives that might be used with a lot of these nouns. Teachers might offer several and ask the students to indicate the nouns that they think collocate with them. Students can be tasked with choosing words from the Similar Words list that relate to training and rank, and then putting them on a cline according to stages, levels or in chronological order. The teacher could ask the students to choose the word that has most in common with *cadet* and explain

the reasons for their choice. Show the students a Similar Words page and ask them to group the words in various ways. These activities require students to interact with each other, and to use related language in the process – this makes them good speaking activities which the teacher can set up and monitor.

Caveats

SkELL processes its corpus using algorithms. If English were used only according to grammar rules, or if the corpus only contained simple, error-free texts, the algorithms could produce perfect results. However, this is not how people use language nor how linguistic software works, so we have to tolerate errors. When we see things that look wrong, we can ask if it might be right in some contexts, or we can choose to ignore it because there are plenty of other examples to work with. Furthermore, if something occurs very infrequently, it cannot be a pattern of normal usage. It is therefore not a useful model for students.

Another caveat is access to the internet in the classroom. The work described here is best done live, where the teacher can search and click and discuss things as they come up in the course of a lesson. However, preparing printouts in advance and giving copies to students that they can make notes on has its own advantages, especially when SkELL is being introduced.

SkELL conclusion

As claimed at the top of the article, nothing has a greater impact on learners' competence in the four skills than a large vocabulary. In order to use words, it is necessary to study their paradigmatic relationships for which Similar Words is a useful tool. And it is necessary to study their syntagmatic relationships for which we have Word sketches.

The abilities to ask questions, form searches, identify potential answers, arrive at a working hypothesis and create structures of related words, all under the guidance of a teacher and in collaboration with classmates involve knowledge creation, social constructivism and elaborative rehearsal. These are diametrically opposed to rote, maintenance rehearsal procedures. These processes often answer more than the original question and fulfil more than the original task. When the information that students derive is integrated into their thinking, knowledge is created.

This is a brief introduction to what SkELL is, how it works and how teachers and students can use it. Many teachers and students around the world use SkELL. As is the case with any activity or approach, a teacher's choice to use it is essentially guided by their views of language, language acquisition and pedagogy,

VersaText

VersaText is the other online tool described in this article. There will be a brief introduction to it and then a description of its three tools and how they can be used in language teaching.

Because VersaText relies on colour-coding, the images in this text are reproduced on the website where the example text can also be found: www.versatile.pub/ambush-text. It can also be opened using this QR code.

Whereas SkELL demonstrates patterns of normal usage through a huge sample of English drawn from the web, VersaText offers tools for exploring the language of a single text. The text may be one of the preloaded texts that the program provides or it may be a text that the teacher or students upload themselves. The latter is preferable. All of the sentences in SkELL are unfamiliar to the students, whereas in VersaText, the students might already be familiar with the text or the topic of the text. VersaText might be being used to introduce a text that the students will be studying. Another difference from SkELL is the accuracy of the language: algorithms should work perfectly in a deliberately chosen text.

As with SkELL, the focus of working with VersaText is the systematic nature of vocabulary. Using corpus techniques with a single text provides insights into its words and phrases, patterns, grammatical features and some aspects of word frequency. This is quite different from exploring patterns in the whole language.

Students work with texts in course books mostly by answering comprehension questions and identifying linguistic features such as the use of verb tense/aspect, the conditionals and non-defining clauses. A pedagogically and linguistically sound treatment of vocabulary is rare. A text is a goldmine for challenging, creative and motivating vocabulary tasks.

VersaText's three tools are the word cloud, a concordancer and the text profiler. To use VersaText, go to www.versatile.pub. where you will see the page shown in this screenshot. Click the VersaText button in the navigation bar.

The first thing you see is the **Input page** where you will see the message: *Paste in your own text or click on a preloaded text title below*. Do that. Click on the **Word cloud button** and you will see a word cloud of the text. Hover over a word in the word cloud and when it turns red, click on it to see it in a **concordance**. Click on the **Profiler button** to see text statistics and word lists that are descriptions of the chosen text. This is most of what you need to know about clicking in VersaText. For more specific information, click on the **Info button** that you see in the screenshot above. It contains background information about the program, its texts, its colour codes, part of speech tags, etc.

Word clouds

A word cloud is an image made up of the words from a text that you provide. Word clouds are popular in presentations, on webpages, coffee cups and are used in books and lessons as attractive assemblies of words. But they are not just decorative. The important feature of a word cloud is that the more **frequent** a word is in the text, the bigger it appears in the word cloud. This gives you a good idea of what the text is about. We saw this at the top of this article.

While the colours of words in most word clouds are for decorative purposes only, in VersaText, they indicate parts of speech.

Word clouds invite many questions related to both content and language. The teacher can lead the students to make observations that will arouse curiosity about the text itself, which is more engaging and respectful than asking students to read a text and answer comprehension questions.

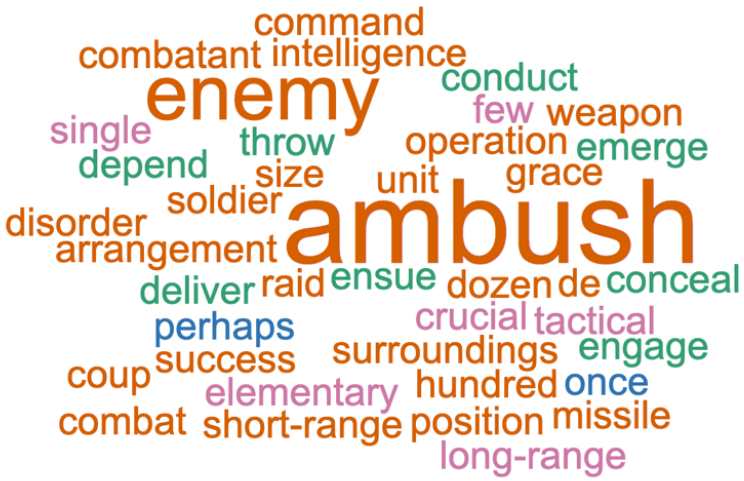
The following examples are based on the short *Ambush* text. Show the students the word cloud and ask some questions. For example:

- Based on the words in this word cloud, what do you think this text is about?
- Is it a story about someone's experience of an ambush?
- Which words are closely related to ambush?
- Is the text about more than one thing?
- Can you identify any topic trails and the words in them? See www.versatile.pub/topic-trails.

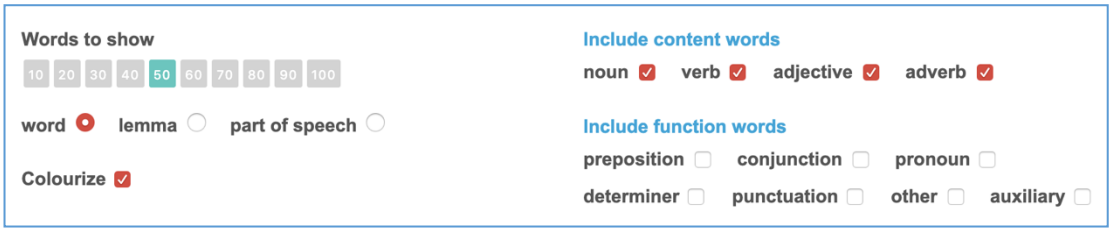
The students can also:

- Use a dictionary
- Underline the nouns in the word cloud that are used in a military context.
- Circle the verbs in the word cloud. Match them to the nouns that collocate with them.
- The pink words are adjectives. What nouns do they collocate with?

- Put some of the words in the word cloud into a syllable stress table. See www.versatile.pub/sst.



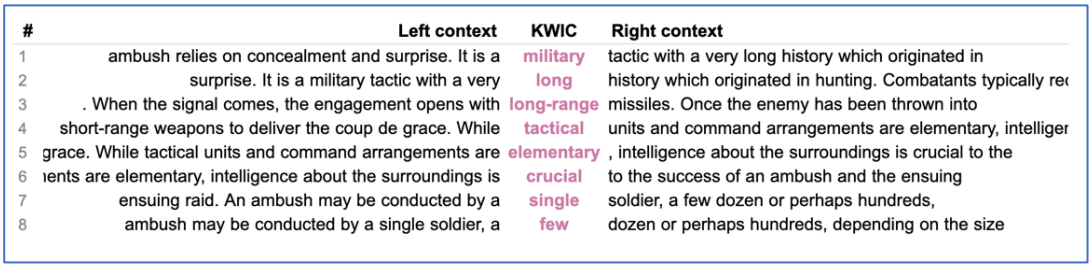
Because the texts are lemmatised and tagged for part of speech, it is possible to customise the word cloud, that is, the user can determine what it contains. Every combination of the buttons, as shown in this screenshot, produces a different word cloud. Each version of a word cloud shows something different about the text and therefore invites different questions.



In addition to words, the word cloud can show part of speech tags. A word cloud of tags shows the relative frequencies of parts of speech in the text. Tags starting with N are nouns, those starting with V are verbs, and those starting with J are adjectives. Full information is at Info.

Concordance

The second tool in VersaText is the Concordance. The target words run down the middle of the page, a format known as *Key Word in Context* (KWIC). This screenshot is the result of clicking on the adjective tag.



#	Left context	KWIC	Right context
1	ambush relies on concealment and surprise. It is a	military	tactic with a very long history which originated in
2	surprise. It is a military tactic with a very	long	history which originated in hunting. Combatants typically rec
3	. When the signal comes, the engagement opens with	long-range	missiles. Once the enemy has been thrown into
4	short-range weapons to deliver the coup de grace. While	tactical	units and command arrangements are elementary, intelliger
5	grace. While tactical units and command arrangements are	elementary	, intelligence about the surroundings is crucial to the
6	ients are elementary, intelligence about the surroundings is	crucial	to the success of an ambush and the ensuing
7	ensuing raid. An ambush may be conducted by a	single	soldier, a few dozen or perhaps hundreds,
8	ambush may be conducted by a single soldier, a	few	dozen or perhaps hundreds, depending on the size

The second concordance shows all of the instances of *ambush* in this short text.

#	Left context	KWIC ↑	Right context
1	An	**ambush**	relies on concealment and surprise. It is a
2	about the surroundings is crucial to the success of an	**ambush**	and the ensuing raid. An ambush may be
3	success of an ambush and the ensuing raid. An	**ambush**	may be conducted by a single soldier, a

These contexts allow us to observe the paradigmatic features of the key word. When observing them in the word cloud, it was only possible to surmise. Using the concordance lines, teachers can ask questions whose answers are in the concordance. Students are not only learning the answers to the questions, but they are learning what questions are worth asking.

Furthermore, a concordance reveals everything that the text says about *ambush*, except for references to it that might be made using pronouns or hypernyms.

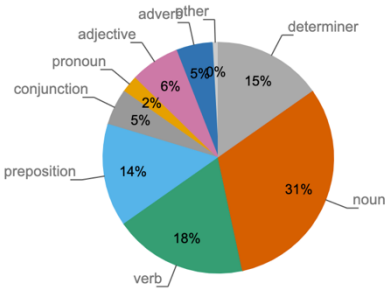
The lines in a concordance can be sorted by clicking on the three column headings. When a key word is quite frequent, patterns can be seen, such as the use of articles, prepositions, compound word forms, etc.

Profiler

The Profiler shows lexical statistics about the text. Click on the Profiler button to obtain data in this format.

	Types	Words	Percent
Top 1,000	46	74	56.49%
1,000 to 2,000	11	12	9.16%
Academic Word List	2	2	1.53%
Text-specific words	29	28	21.37%
All	90	116	100.00%

characters: **734**, lowercase types: **88**, tags: **25**, type-token ratio: **68.70%**, sentences: **6**



The Ambush text has 74 **words**. A lot of the words are repeated, especially function words. In fact, there are only 46 **types** in this text – types are the different words.

The **Top 1,000** words are the most frequent 1,000 words in English generally. This table says that 56.49% of all the words in the text are among the most frequently used 1,000 words of English and that only 9.16% are in the second thousand.

The **Academic Wordlist** (AWL) is a widely used, published list of words that are not in the top 2,000 but are commonly found in many kinds of academic texts. The AWL

does not include terminology. Less than 2% of the words in the Ambush text are AWL words. In an academic journal article, 15% is standard.

Text-specific words are the words in the text that are not in the other three groups. These are usually words related to the topic, since most topic-based words are not in the top 2,000 most frequent words in English.

Text-specific words

PoS	Items
noun	ambush concealment surprise tactic combatants reconnoitre terrain stalk signal engagement missiles disorder combat coup grace surroundings raid dozen
verb	originated hunting concealed ensuing
adjective	tactical elementary
adverb	typically

VersaText vs. SkELL

SkELL is the most student-friendly corpus software publicly and freely available. Teachers need to ask specific questions whose answers can be gleaned from skimming the example sentences, not usually needing to read them closely. The task types presented above are designed to train students to find answers to specific questions. Using corpora helps learners make decisions consciously that native speakers make subconsciously.

Many of the observations that can be made using SkELL can also be made using VersaText. However, when observing features of language from a single text, the conclusions are only valid for that text. It is, in effect, a case study. If many case studies were conducted, it would be possible to create a meta-analysis of them from which patterns of normal usage would emerge. Informally, subconsciously, this is how we gradually acquire the norms of a language.

When we study language from a single text, we cannot be sure that the findings are generalisable to the whole language. For example, in the Ambush text the word *crucial* is used once and is used in the grammar pattern, *something is crucial to something*. Such anecdotal evidence should not convince a learner to include this in their learning journal. If encountering this word in the text has motivated the student to prioritize learning it, they need to check with a thousand native speakers first: search SkELL for *crucial* and again for *be crucial to*.

It may be argued that spending so much time on one word is an obstacle to students learning acquiring a big vocabulary. It may be countered that in the process of investigating one word, the students are having multiple experiences with the language in much more focussed ways than through doing passive

activities. Roberts and Kreuz argue that even if students study fewer words through elaborative rehearsal, the ones they do study will be more meaningful, better remembered and used correctly. This is a matter of quantity vs. quality.

The advantages to working with a single text, however, are the students' closeness to the text, the fact that every word, phrase and sentence is a part of something they know, and that the teacher has chosen the text for its relevance. It is most likely correct English. While VersaText does not represent the whole language, it can be argued that the students are not learning the whole language, not at any given moment in their language learning lives.

References

Anderson, L.W. (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Longman.

Ellis, R. (2003) *Task-based Language Learning and Teaching.* Oxford University Press.

Francis, G., Hunston, S. & Manning, E. (1996) *Grammar Patterns 1: Verbs.* HarperCollins

Francis, G., Hunston S. & Manning E. (1998) *Grammar Patterns 2: Nouns and Adjectives.* HaperCollins

Piaget, J. (1972) *Some Aspects of Operations. In Play and Development.* Norton.

Richards, J., and Schmidt, R. (2002) *Longman Dictionary of Language Teaching and Applied Linguistics (3rd ed.)* Harlow: Longman.

Thomas, J. E. (*Forthcoming*) *Grammar, Vocabulary and Everything In Between.* Versatile Publisher.

Thomas, J. E. (*forthcoming*) *Practical Methodology for English Teachers in Military Institutions.*

Thomas, J. E. (2017) *Discovering English with Sketch Engine (2nd ed)* Versatile Publisher.