Discovery Learning: Teaching Languages with Corpora

by Nina Vyatkina, University of Kansas

Do you want to help your students…

- learn how to use L2 words in context?
- understand L2 grammar not from prescriptive rules but from usage examples?
- develop autonomous language learning skills?

You can do all this with the help of language corpora: large electronic collections of texts. Here are five points outlining the principles and benefits of this approach and some tips to get you started.

1. Corpora are rich resources of contextualized examples of language use

Corpora are very large collections of naturally occurring texts, so if you are tired of inventing example sentences for your students to illustrate the use of a word or an expression or to create test banks, consider using these treasure troves of language! For instance, if you enter "green energy" in the search line of the front page of the News on the Web corpus (part of COCA, the Corpus of Contemporary American English), you will find almost 20,000 usage examples. Note that many corpora group texts by register and genre (e.g., newspaper, blog, fiction, radio interview texts), so that you can find examples from the text type of your choosing. For a list of and links to open access corpora in different languages see the Appendix in this article.

2. There are many resources to help you start using corpora for language teaching

There are many corpus user guides available to help you navigate corpora, analyze corpus search results, and design corpus-based activities. A number of book-length guides written in accessible language have appeared in recent years (e.g., Bennett, 2010; Friginal, 2018; Poole, 2019). In addition to instructions for how to use corpora in language teaching, they contain suggestions for specific activities and exercises. Although written for English teachers, they provide useful advice for corpus-based teaching of any language. Especially promising is the new movement toward creating Open Educational Resources (OER) for corpus-based language teaching. An excellent example is the Corpus-Aided Platform for Language Teachers (CAP). This rich suite of open-access corpus-based teaching resources developed under the leadership of Ma Qing won the 2020 CALICO / Esperantic Studies Foundation Access To Language Education Award. For an example of corpus-based OER for German see Incorporating Corpora.

3. Corpus-based learning is an active discovery learning method

Corpora are especially suited for inductive language learning, which is reflected in another widely used name of this method – data-driven learning (DDL) – proposed by Tim Johns back in 1990. As opposed to deductive methods, where the teacher explains a rule and then illustrates it with examples, DDL usually follows an inductive approach. Learners are first presented with corpus examples, then they observe patterns in these examples, and finally, they induce rules from this analysis. This process is scaffolded by visualization techniques frequently used in the

corpus search interface. For example, if you search the German DWDS corpus, namely, its ZEIT newspaper subcorpus, for the word “Internet” (an English borrowing), you receive the result in form of the so-called concordance, or example sentences stacked under one another with the search word highlighted and centered:

Kurz darauf kommen im	**Internet**	Videos in Umlauf, die sich rasant im ganzen Land verbreiten.
ung der Kommunikationsnetze und des schnellen	**Internets**	vor allem auf dem Land, von Straßen, Brücken, Eisenbahnstrecken
ingestanzt, wer einen findet, soll den Fundort per	**Internet**	oder Handy-App der Uni melden.
Im	**Internet**	suchte der Bezirksverband Hessen-Süd jemanden für die Nachmitta
Es gibt im	**Internet**	Musterbriefe für Klageschriften.
Entscheiden nicht Computer darüber, was wir im	**Internet**	zu sehen bekommen?
rtungen der Leitunternehmen des kommerziellen	**Internets**	: Google/Alphabet schafft es auf 580 Milliarden Dollar, Apple auf sog
Was kauft ein Mensch im	**Internet**	?
, dass mit dem Gesetz die Rechtsdurchsetzung im	**Internet**	an private Firmen delegiert wird.

This format helps some visual patterns jump at you even if you don’t know any German: 1) “Internet” starts with a capital “I”; 2) it is frequently preceded by the word “im”; 3) it sometimes takes the ending “-s”. After exploring this and similar examples, even beginning learners of German can induce that 1) nouns in German are capitalized; 2) “im Internet” stands for “on the internet”; 3) this (neuter) noun takes the ending “-s” to mark the genitive case. Using this method, learners act as “language detectives” and discover language use patterns by themselves instead of being told about them by the teacher. The role of the teacher in DDL, however, remains extremely important: learners need to be guided in this inductive learning process, especially when they first start interacting with corpora or with corpus printouts.

4. Corpora show us that languages are complex, dynamic systems

The usage patterns we observe in corpora show that language rules are not written in stone and that vocabulary and grammar are inextricably intertwined. For example, while completing corpus search tasks, your learners may discover that: different near-synonyms may be preferred in different genres; words attract some words as frequent neighbors and repel other words; certain words tend to occur in certain syntactic constructions; spelling norms fluctuate depending on the time period; the popularity of certain words rises and declines over time; native speakers and respected publishers may make mistakes or intentionally flout prescriptive grammar rules! These discoveries will help learners appreciate and tolerate the complexity of the human language and feel much better about themselves as second language users.

5. DDL works!

Empirical DDL research has been conducted for over 30 years, and its findings convincingly demonstrate that corpus-based teaching and learning works with different languages, different L2 learners (beginning and advanced, young and adult), in different contexts (high school and university), and for different learning targets (vocabulary, grammar, discourse; reading, writing, speaking). For an overview, you may read this one-page accessible summary of a recent meta-analysis of DDL research or Alex Boulton’s research timeline. Most learners also like DDL as a learning method. Perhaps most importantly, after learning how to use corpora in your class, learners can continue doing so independently for life-long language learning. Last but not least, DDL helps learners develop their analytical and critical thinking skills which can be applied in all areas of employment and human activity.

Cite as: Vyatkina, N. (2020, December). Discovery learning: Teaching languages with corpora. *CALICO Infobytes*. Retrieved from http://www.calico.org/infobytes

Selected CALICO Journal articles on corpus-based language teaching and learning

Curado Fuentes, A. (2005). The use of corpora and it in evaluating oral task competence for tourism English. *CALICO Journal, 22*(1).

Garner, J. R. (2013). The use of linking adverbials in academic essays by non-native writers: How data-driven learning can help. *CALICO Journal, 30*(3).

Hegelheimer, V. (2007). Helping ESL writers through a multimodal, corpus-based, online grammar resource. *CALICO Journal, 24*(1).

Liou, H.-C., Chang, J. S., Chen, H.-J., Lin, C.-C., Liaw, M.-L., Gao, Z.-M., Jang, J.-S. R., Yeh, Y., Chuang, T. C., & You, G.-N. (2007). Corpora processing and computational scaffolding for a web-based English learning environment: The CANDLE project. *CALICO Journal, 24*(1).

Pereira, L., Manguilimotan, E., & Matsumoto, Y. (2016). Leveraging a large learner corpus for automatic suggestion of collocations for learners of Japanese as a second language. *CALICO Journal, 33*(3).

Poole, R. (2012). Concordance-based glosses for academic vocabulary acquisition. *CALICO Journal, 29*(4).