10

A Corpus-Based Pedagogy for German Vocabulary

Colleen Neary-Sundquist
Purdue University

Abstract

Contemporary foreign language textbooks used in the United States have been criticized for shortcomings both in their presentation and vocabulary exercises. The inclusion of authentic materials in the language classroom would seem to help alleviate this problem; however, the use of authentic materials at lower levels of language instruction poses its own set of practical challenges. This paper presents corpus-based exercises designed for lower-level language classes that are paper-based, thereby eliminating potential practical problems while offering students the opportunity to explore vocabulary as well as culture through authentic materials.

Introduction

Beginning in the late 1980s, corpus linguistics, or the study of language through collections of written or spoken language, experienced a renaissance due in part to the comparative ease of creating and managing large amounts of data with computers. Despite the widespread availability of increasingly large and sophisticated corpora of natural language, the application of corpus-based methods to problems in second language teaching has remained limited. Many teachers and learners are unaware of the corpus resources that exist and how they could be used to facilitate language teaching and learning.[1]

The relative lack of corpus-based pedagogical treatments is unfortunate because they offer a number of features that stand in contrast to more traditional textbook pedagogies and can therefore serve as a useful supplement to them. Corpus-based activities involve exposure to authentic language data, encourage learner autonomy, and are compatible with an inductive approach in which the learners are encouraged to make their own discoveries about the language (Chambers, 2010; Gilquin & Granger, 2010).

Problems with the current state of vocabulary in foreign language textbooks

The current presentation of vocabulary in contemporary lower-level language textbooks widely used in the United States is problematic. Previous research has shown that the vocabulary chosen for presentation fails to include the most frequently used words (Lipinski 2010). Furthermore, the activities presented concentrate to an overwhelming extent on only some sub-types of vocabulary knowledge (Brown 2011; Neary-Sundquist, in press).

A number of previous studies have compared textbook vocabulary to natural language corpora and found substantial discrepancies between the two. Many of these studies have been conducted on materials for English language teaching (Carter & McCarthy 1995, Glisan & Drescher 1993, Gilmore 2004).

Research comparing the vocabulary found in U.S. foreign language textbooks with corpora has also been conducted, although this area is not as well developed as the English corpus-based textbook studies. For Spanish, Davies and Face (2006) compared vocabulary word lists from six college Spanish textbooks with frequency data from the *Corpus del Español*. They found that "…for whatever N number of vocabulary words a textbook includes, only 10-50% of those are among the N most frequent lemma in the language. For example, as Table 4 above indicates, if a textbook presents 2000 vocabulary words, only 10-50% of those words are among the most frequently used 2000 lemma in the language." In other words, the majority of the words covered in contemporary Spanish textbooks are not the most frequent words in the language according to language corporaFor German,

Lipinski (2010) compared the frequency of vocabulary presented in German textbooks with corpora or frequency lists for German. Lipinski (2010) compared the vocabulary presented in three first-year textbooks of German with the most frequent German words as presented in the *Frequency Dictionary of German*. She found that 29-44% of the words found in the three textbooks were in the 4000 and less frequent words. Only 24-36% of the words in the three books belonged to the 1000 most frequent word group. Although Lipinski notes that frequency alone cannot be the sole factor in selecting vocabulary for textbook presentation, she characterizes the results as "disheartening" and observes that this may contribute to cognitive overload on the part of the students.

In sum, studies on various foreign and second language textbooks have found a serious discrepancy between the vocabulary presented and the vocabulary frequently used by native speakers. A majority of the vocabulary items presented in textbooks is composed of relatively low-frequency words.

The comparison of the vocabulary found in language textbooks with that found in natural language corpora is only one aspect of textbook vocabulary instruction. Brown (2011) investigated another aspect, that of the types of vocabulary knowledge that textbook activities focus on. In this analysis, Brown (2011) examined textbook vocabulary activities using Nation's (2001) framework of the various aspects of vocabulary knowledge.

Nation (2001) proposed that vocabulary knowledge is not a matter of making a simple form-meaning connection. On the contrary, he identified nine aspects of knowledge that together make up what it means to know a word. Nation distinguished three overarching aspects of vocabulary knowledge, each with three subcategories: Form (spoken form, written form, word parts), Meaning (form and meaning, concept and referents, associations), and Use (grammatical functions, collocations, constraints on use). Nation pointed out that the psychological reality of these distinctions between form, meaning, and use aspects of vocabulary knowledge is supported by previous research (Ellis 1994; 1995, Aitchison 1994).

Brown (2011) analyzed the vocabulary activities in English as a Second Language textbooks using Nation's (2001) nine aspects of vocabulary knowledge. Brown found that textbook exercises overwhelmingly focus on the aspects of form and meaning and grammatical function. Spoken form was given moderate attention, but the other six aspects of vocabulary knowledge (written form, word parts, concept and referents, associations, collocations, constraints on use) were largely neglected. A similar study of German textbooks (Neary-Sundquist, in press) found results that were largely similar to those of Brown (2011). The aspects of vocabulary knowledge that received the most attention were form and meaning and grammatical function, while collocations and constraints on use received the least attention.

Practical difficulties with integrating corpus-based exercises in the language classroom

There are a number of practical difficulties that have most likely contributed to the fact that the use of corpora in the classroom as language learning tools has not become widespread. First of all, language classes are not usually conducted in classrooms that have a computer for every student. This is possible, but requires access to a computer lab, which in turn requires advance planning and limits the amount of time available for the learning activity. Once in the computer lab, the set-up of the room may make it difficult to work on other types of activities. In other words, a teacher cannot simply work a corpus exercise into a class on an ad-hoc basis, but must more-or-less plan for an entire corpus-based lesson. Once the issues surrounding computer lab access have been dealt with, the next hurdle involves becoming proficient at using the technology. Although we may think that we are living in the digital age in which all of our students are comfortable with anything computer-based, this is not always the case. It has been my experience that students' familiarity with technology is often limited to particular programs, and that they are just as intimidated by new and unfamiliar technology as those who are not avid users of the Internet might be. They are unsure of how to do

things and fearful of pressing the wrong button. Thus using corpora in the language classroom requires training time for the students as well as troubleshooting time. This may further discourage teachers from bringing corpus-based activities into the classroom. The time needed to teach the students how to use the corpus combined with the time that the students will actually be accessing it makes this a time-consuming pedagogy. It is not surprising that teachers might choose to employ a more traditional approach; they might ask themselves if using a corpus to illustrate the difference between two words is really worth it. It is easier, simpler, and less time-consuming to simply tell the students when to use two words, such as *studieren* 'to study (a discipline)' and *lernen* 'to learn'. An inductive approach will likely take more time, and initially definitely more preparation on the part of the teacher. Teacher preparation is another issue that disfavors corpus use. As in the case of students, teachers are often not comfortable with new technology. Especially when faced with a situation in which they must become expert users and in turn teach others in a relatively short time, it is easy to understand why teachers might avoid bringing natural language corpora into the classroom.

One of the biggest challenges of making the use of language corpora more widespread, however, may simply be that the teachers lack familiarity with the resources and a lack of ideas of how to use them. The only way to solve this problem is to educate teachers through presentations and articles in order to make the entire process of accessing a corpus less intimidating and to offer suggestions and examples of how this can be integrated into their teaching. This paper aims to promote teacher awareness of the utility of integrating corpus based activities into their curriculum and offering practical suggestions for how to make their own activities that align with their own teaching units and pedagogical goals.

A paper-based, alternative approach to corpus exercises

Boulton (2010) argues persuasively that corpus exercises that are paper-based have a number of advantages. Following Kirschner, Sweller, and Clark (2006), Boulton notes that paper-based materials may be particularly appropriate for lower-level learners. A relatively free activity in which learners interact with the corpus without much supervision may be too demanding for lower-level learners and overload their working memory capacity. Paper-based corpus exercises also allow learners to get used to the idea of the corpus and how it works, serving as an entry point into corpus-based pedagogy. Boulton notes the following:

> In other words, learners may find it easier to graduate from "soft" to "hard" DDL (Gabrielatos, 2005) or from what Cresswell (2007) called "deductive DDL" (i.e., starting with teacher-led exercises) to fully "inductive DDL" (i.e., starting with the data on their own). (p.539)

The corpus: *Das Digitale Wörterbuch der deutschen Sprache (Digital Dictionary of the German Language)*

The activities for this project use corpus data and online features of the *Digitales Wörterbuch der deutschen Sprache* (Digital Dictionary of the German

Language), or DWDS. The DWDS is an online corpus project sponsored by the *Deutsche Forschungsgemeinschaft* (German Research Society) and the *Berlin-Brandenburgische Akademie der Wissenschaften* (Berlin-Brandenburg Academy of Sciences), available online at www.dwds.de. The project's main purpose is to provide an online, digital dictionary and a massive repository of searchable 20th- and 21st-century German-language texts that serve as sources for the dictionary. The DWDS is based on several dictionaries and aggregates information from the *Wörterbuch der deustchen Gegenwartssprache* (Dictionary of Contemporary German), the *Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm* (German Dictionary by Jacob Grimm and Wilhelm Grimm), and its updated edition, as well as the *Etymologisches Wörterbuch des Deutschen* (German Etymological Dictionary) by Wolfang Pfeifer. The main reasons the DWDS was used for this study are that it is one of the largest German-language corpora available online, but it is also balanced and representative, with many different sources and its interface is relatively easy for both students and teachers to learn to use with little experience using corpora.[2]

In addition to providing definitions, synonyms, etymologies, and all other lexical information about each word gathered from the various dictionaries, the DWDS provides examples and data from a large, balanced, and representative corpus of texts. These texts make up the *Kernkorpus* (Core Corpus) that was used in the design of activities for this paper. The *Kernkorpus* consists of over 125 Million words in 7 Million sentences found in 79, 830 documents from various genres and text-types written in the 20th century, including literary works, scientific texts, non-fiction and newspapers. The corpus is annotated for parts of speech and is lemmatized to allow for searches of various grammatical forms of each word.[3]

Although there are several features of the DWDS and the *Kernkorpus* that can be used in the design of classroom activities, this study focused only on one, namely, the *Wortprofil 3.0*, or Word Profile. After the user enters a word in the DWDS, the Wortprofil panel appears automatically among several DWDS panels as a default that display different aspects of the original entry and its lexical characteristics. The *Wortprofil* panel displays a word cloud, or a graphic display of words that are associated with the entry word based on co-occurrences with it in the corpus. The user can choose to display between two and 250 associated words in the word cloud; the associations are displayed with varying sizes and boldness based on the frequency with which they co-occur with the lexical entry, as in the popular word clouds generated online by sites such as www.wordle.com. When the user clicks on any associated word in the cloud in the *Wortprofil*, sentences appear in a panel below the cloud that provide examples of real examples from the corpus in which the word its associates. Other features of the *Wortprofil* include searches for various grammatical forms that occur along with the entry word, including attributive adjectives that often occur with the word or other words that often occur in coordinated constructions with the original entry. In addition, the user can enter a *Vergleichswort* (comparison word) so that the set of associated words for two entries can be displayed in the panel at the same time. Moreover,

quantitative data are available for all associations along with the strength and frequency of these associations.

Two corpus-based exercises for lower-level learners of German

The two exercises presented below (in Appendix A) were created using data from the DWDS corpus. They are entirely paper-based and could be printed out and used in the classroom as is. The only additional materials needed to work through the exercises is a dictionary of some kind, and even this is be optional if the teacher would rather translate some words for the students.

The exercises first introduce the students to the idea of a corpus as a collection of language. They are then introduced to the first word cloud, which has the fairly intuitive feature that the larger a word is, the more frequent it is used. Students are initially asked simply to find three of the larger (=most frequent) words or phrases that occur with the word *Kaffee* 'coffee'. This is a simple exercise that could be done even in the first semester of study. Similarly, the rest of the exercises also only ask the student to find words or phrases, write them down, and look up their meanings or ask their instructor as necessary.

Exercise E takes the learners a little further, asking them to try to decipher some full sentences from the corpus regarding coffee drinking habits of various nationalities. Likewise, this activity was designed to keep the burden on the learner relatively low by giving them simple true/false questions in English regarding the content. The phrasing of the true/false questions gives a clue to the content of the sentences if a learner is completely lost. If this particular exercise were judged to be too demanding for very low-level beginning students, it could of course be omitted or moved to the end of the exercises and treated as optional.

The final activity asks learners to compare the results from the DWDS search with results for the same word in a corpus of American English, the Corpus of Contemporary American English (COCA). This last step allows the learner to consider the potential cultural connotations of the word and idea of coffee in both German and American cultures. Part of the intent here is to notice the co-occurrence of *Kaffee* and *Kuchen* 'cake' in the German corpus, which could in turn lead to a discussion of this afternoon ritual. Similarly, the final question in the pizza exercise asks the students to compare collocates of pizza in the German and American corpora. Many of the words that occur most frequently with pizza in the American corpus relate to the names of chain restaurants or to words that have to do with pizza delivery. However, the goal of this part of the activity is open-ended and designed to go beyond the author's expectation. It has been my experience that students often make connections and observations that escaped me when I designed the activities. This is to be welcomed in this type of exercise.

The final comparison activity could also lead to a discussion how arguments are constructed and what constitutes evidence. This undoubtedly involves higher-order thinking skills that some might find challenging to incorporate into language classes. However, it is mentioned here as an example of how language learning can build critical thinking skills, which is of particular relevance for university language programs that are increasingly called upon to justify their existence. An

example of this type of evaluative skill would be to ask the class what it might mean that *aus Pappbechern* 'out of paper cups' is mentioned in the German but not in the English corpus. Does this indicate that Germans drink coffee out of paper cups more than Americans do? Not necessarily--it could also be the case that this phrase co-occurs with coffee in the German corpus because it is a practice that is being discussed in the media more frequently and has a particular cultural significance. In contrast, it might not be mentioned as much in the American sources because it is an accepted fact of life that is not worth remarking on. To resolve this question, the corpus itself must be consulted to see how the expression is used in context. But even if this is not done inside or outside of class, it is important to highlight that the co-occurrence of one word or phrase with another may signify different things.

These example exercises expose learners to authentic vocabulary, but they also put an emphasis on words that occur with the vocabulary word under investigation and ask the student to identify the superordinate categories to which the words belong. Both of these aspects of vocabulary knowledge were identified by Brown (2011) as receiving very little attention in textbook exercises. These activities therefore supplement the textbook focus on the form-meaning connection and grammatical use of vocabulary items.

An additional advantage found in these materials is that they allow for varying levels of interest and ability. Some students may feel comfortable doing the minimum required of filling in the blanks, while others may eagerly look up everything in the word cloud and later proceed to access the corpus itself online. The use of materials that offer something for different levels of proficiency and interest is not a trivial consideration. In classes that may contain 15-25 students, it is not possible for the teacher to target lessons for every level; they must by necessity try to reach the middle level of students with most of their planned activities. One potential solution to this problem is to include minimum and maximum levels of achievement within one exercise so that learners at either extreme do not feel either overwhelmed or bored with the activity. In the exercise presented here, a closer look at the word cloud should offer a challenge for more proficient learners.

Other advantages: Data Driven Learning and Learner Autonomy

The activities presented above were created to help correct the fact that contemporary German language textbooks often present relatively low-frequency vocabulary. These activities are designed with the every-day classroom teacher in mind, with a goal of making the incorporation of corpus-based authentic materials more accessible and less prone to practical or technological problems. But aside from utility, there are indications in the previous research that working with these types of material can increase both learner motivation and learning.

Johns (1988, 1991) first suggested what he termed data-driven learning. In this method, learners examine a set of examples of a vocabulary word (or other grammatical feature) taken from natural language by native speakers. They explore the material themselves, and discover how the language works inductively. Johns likened the learner's role to that of Sherlock Holmes; each learner is an active

language detective. I think that this approach is both appropriate to the vocabulary-learning problem outlined above as well as empowering for the students. I think that our students' natural curiosity and desire to learn is sometimes deadened by the way we present material to them in the traditional classroom environment.

The type of inductive approach to learning stands in contrast to the typically deductive approach favored by traditional pedagogy and textbooks. The switch from a teacher-led, deductive approach to a learner-centered, inductive approach has important consequences for the role of both the learner and the teacher. Sripicharn (2012) characterizes the learner's role in this type of data driven learning as a researcher, detective, or traveller, and notes that this role is particularly well-suited to corpus exploration. In this type of pedagogy, the learner has direct access to real language data without the mediation of the teacher. This may be intimidating to some learners, but it can also build confidence in their abilities and develop their critical thinking skills. Self-determination theory (Deci & Ryan 1985) argues that autonomy is a key component of learner motivation. The role of learner autonomy in language learning has received increasing attention in recent years, and has been shown to increase motivation and active participation as well as a greater sense of the learner's own responsibility for their learning (Nguyen & Gu 2013). Corpus exploration using a data driven learning approach has the potential to increase learner autonomy.

An increase in learner autonomy also affects the role of the teacher. It allows the teacher to reduce the extent to which they are seen as the authoritative and final source of knowledge about the language being taught. This role is burdensome for the teacher and has been called the "Atlas complex" in reference to the Titan Atlas, who held up the sky, literally bearing the weight of the world on his shoulders. By familiarizing learners with the use of a corpus, the teacher is able to show students another source for knowledge about the language and answers to their questions, one that they can use themselves and one that does not always give simple answers. The more complex answers found when searching real language data may sometimes make students uncomfortable, but they also reflect the complexity of language.

Conclusion

Best practices in language learning technology advocate a "pedagogy first, technology second approach", in which a pedagogical problem is identified, and then a solution is sought that may or may not involve technology. Technology is never applied simply because it is available or seems to be cutting-edge. Rather, it is used only when it is the best tool to solve a pedagogical problem.

This paper has argued that corpus-based exercises are an appropriate tool to solve the pedagogical problem of the lack of natural and frequent vocabulary in contemporary foreign language textbooks. The state of affairs in current textbooks is unlikely to change anytime soon, nor is their widespread use in language classrooms. The textbook is entrenched as a given in both secondary and university classrooms, providing an established and familiar framework for language learning.

Since textbooks are unlikely to be replaced or extensively revised anytime soon, the aim of this paper has been to raise awareness of resources that are available to augment them. A corpus-based inductive model could be brought into any classroom as a counterpoint to the traditional presentation-practice-production model. This will serve to expand both the students' and teacher's knowledge of what constitutes language and will allow them to see the textbook as one resource, rather than as the ultimate source of knowledge about the language.

Notes

1. For an overview of the use of corpora in language teaching, see Vyatkina (2012), O'Keeffe & McCarthy (2010) and Sinclair (2004).

2. Vyatkina (2013) provides useful information on using the DWDS for teaching purposes. Her project focuses on advanced-level students who use the corpus for collecting data on grammatical constructions.

3. In addition to the *Kernkorpus*, the DWDS includes several other sub-corpora that weren't used in this study. They include a journalistic corpus with articles from *Bild, Welt*, and *Die Zeit,* and several other newspapers; the *DDR-Korpus* with 9 Million words from texts written in the German Democratic Republic between 1949 and 1990; the *Wendekorpus*, which includes transcriptions of 77 interviews with East and West Berliners' experiences with the Fall of the Berlin Wall; and the *Korpus Gesprochene Sprache* (Corpus of Spoken German), or 2.5 million tokens from speeches and interviews from the 20th century. The DWDS displays information from these other corpora, but only the *Kernkorpus* was used here.

References

Aitchison, J. (1994). *Words in the mind.* Oxford: Blackwell.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language learning*, *60*(3), 534-572.

Brown, D. (2011). What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research*, *15*, 83–97.

Carter, R. & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*, 141–158.

Chambers, A. (2010). What is data-driven learning? In A. O'Keefe and M. McCarthy (Eds.) *The Routledge Handbook of Corpus Linguistics (pp.345-358).* New York: Routledge.

Cresswell, A. (2007). Getting to "know" connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 267–287). Amsterdam: Rodopi.

Davies, M. & Face, T. (2006). Vocabulary coverage in Spanish textbooks: How representative is it? In *9th Hispanic Linguistics Symposium,* Somerville, MA. Retrieved from http://www.lingref.com/cpp/hls/9/paper1373.pdf

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Ellis, N. (1994). Factors in the incidental acquisition of second language vocabulary from oral input: A review essay. *Applied Language Learning, 5,* 1-32.

Ellis, N. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics, 16,* 409-441.

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells? *Teaching English as a Second Language—Electronic Journal, 8*(4), 1–35.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, *58*(4), 363–71.

Gilquin, G. & Granger, S. (2010) How can data-driven learning be used in language teaching? In A. O'Keefe and M. McCarthy (Eds.) *The Routledge Handbook of Corpus Linguistics (pp.359-370).* New York: Routledge.

Glisan, E.W. & Drescher, V. (1993). Textbook grammar: Does it reflect native speaker speech? *The Modern Language Journal, 77*(1), 23–33.

Johns, T. (1988). Whence and whither classroom concordancing. In T. Bongaerts, P. De Haan, S. Lobbe & H. Wekker (Eds.), *Computer applications in language learning* (pp. 9–27). Dordrecht: Foris.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing* (pp. 1-16). Birmingham: Centre for English Language Studies, University of Birmingham.

Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86.

Lipinski, S. (2010). A frequency analysis of vocabulary in three first-year textbooks of German. *Die Unterrichtspraxis, 43*(2), 167-174.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.

Neary-Sundquist, C. A. (in press). Aspects of Vocabulary Knowledge in German Textbooks. *Foreign Language Annals*.

Nguyen, L T. C. & Gu, Y. (2013). Strategy-based instruction: A learner-focused approach to developing learner autonomy. *Language Teaching Research, 17*(1), 9-30.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. New York: Routledge.

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research*, *11*(4), 393-412.

Sinclair, J. McH. (2004). (Ed.). *How to use corpora in language teaching*. Amsterdam and Philadelphia: John Benjamins.

Sripichan, P. (2012). How can we prepare learners for using language corpora? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 371-384). New York: Routledge.

Vyatkina, N. (2013). Discovery learning and teaching with electronic corpora in an advanced German grammar course. *Die Unterrichtspraxis, 46*(1), 44-60.

Appendices

Appendix A: Sample Exercises

Example 1: *Kaffee*

Below you will find a word cloud for the word *Kaffee* in German. A word cloud shows the words that commonly occur with the word *Kaffee*. The larger the word in the box is, the more often it occurs with *Kaffee*. The data used to make this list comes from a collection of German language, Das Digitale Wörterbuch der deutschen Sprache. The DWDS is composed of over 1.8 billion words. A large collection of natural language like this is referred to as a corpus.

Now we will do some exercises to learn more about the word Kaffee in German.

A. Look at the words in the box. Write three of the biggest (=most frequent) words or phrases below:

_____ _____ _____

Do you know what these words or phrases mean? If not, take a moment to look them up or ask your instructor.

B. Look at the box again. Find three other food words (not drinks) that are mentioned. If you are not sure whether something is a food word, you may need to look it up.

_____ _____ _____

C. Some of the words above are containers for holding coffee. Can you find three?

_____ _____ _____

D. Some of the phrases in the box start with the word *wie (=like)*. These expressions often indicate a category to which *Kaffee* belongs. Find three of them and write them below. What do they mean?

_____ _____ _____

E. Look at the top left corner of the word cloud, and you will see the following expressions: *als Bier* and *als Tee*. What does *als* mean, and why do you think these expressions commonly occur with the word *Kaffee*? What kinds of sentences might they be a part of?

Here are some sample sentences that show how the expressions *als Bier* and *als Tee* are used with *Kaffee* in the corpus. Take your time and see if you can figure out what the sentences are saying about the consumption of coffee versus other beverages in Germany, the U.S. and the U.K., then answer the questions below.

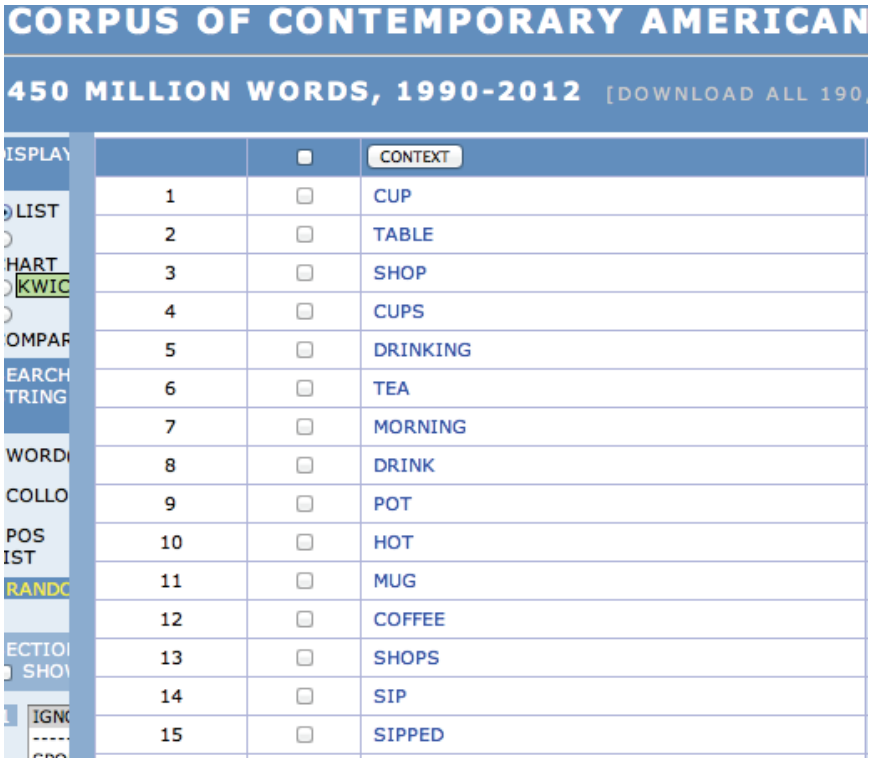
Mit durchschnittlichen 160 Litern im Jahr trinkt der Deutsche mehr ***Kaffee als Bier*** *und Mineralwasser.*

Zum ersten Mal in der Geschichte des Vereinigten Königreichs wird mehr ***Kaffee als Tee*** *konsumiert, schrieb kürzlich der Guardian.*

Zwar wird heute in den USA mehr ***Kaffee als Tee*** *getrunken, aber ganz vergessen die Amerikaner den Tee sicher nicht.*

Germans drink more coffee than beer.	True	False
Americans drink more coffee than tea.	True	False
In Great Britain, they drink more tea than coffee.	True	False

E. Below is a box that shows similar information for the word coffee in American English. This data comes from the Corpus of Contemporary American English (COCA), which consists of 450 million words. In this sample, the words are ranked rather than in a word cloud. This sample shows the 15 most frequent words that occur with coffee in English. What similarities and differences do you notice between the German word cloud and the English word list. What could this suggest about the differences between how and when coffee is consumed in German vs. American culture?



Example 2: Pizza

A. Several of the words and phrases above refer to other kinds of "fast food". Can you find three of them?

_____ _____ _____

B. Another group of words that occur with *Pizza* in German are foods that are not of German origin. Can you find three of them?

_____ _____ _____

C. Several of the expressions that occur frequently with *Pizza* refer to how the pizza is baked, *aus Holzofen* and *aus Steinofen* (in the upper left area of the word cloud). Do you see a word you recognize in either of these words? Can you guess what they mean? If you do not know, look them up.

aus Holzofen =
aus Steinofen=

D. Towards the middle of the word cloud, you can see two expressions that start with *mit* (=with), mit Champignons and mit Schinken. What do these expressions mean, and why do you think that they occur frequently with Pizza?

E. Can you find three verbs that occur commonly with pizza? What do they mean?

_____ _____ _____

F. Below is a sample from the COCA corpus of American English for the 15 words that occur most frequently with the word pizza. How many of the words relate to ordering pizza for delivery? Can you find the phrases above in the German word cloud that relate to pizza delivery? What are the most common toppings in the American corpus, and how does this compare to the German results? Can you make any guesses about differences in how Americans and Germans consume pizza?

		CONTEXT
1		HUT
2		PIZZA
3		DOUGH
4		CRUST
5		CHEESE
6		SLICE
7		PARLOR
8		DELIVERY
9		ORDER
10		EAT
11		PEPPERONI
12		DOMINO
13		OVEN
14		PASTA
15		BOX

You can search the corpora yourself. Here are the sources:

DWDS is available online at: http://www.dwds.de
COCA corpus is available online at: http://corpus.byu.edu/coca/