

Using Corpora in Teaching Languages to Enhance Competency in English as a Foreign Language with Focus on English for Business and Economics

Michelle Duruttya

Abstract: The necessity to be competent in understanding and interpreting written as well as spoken English language is vital in our society. There are many methods and approaches set to confront this particular demand, one of which corpus linguistics addresses in a highly efficient way. Corpora can be used to study purposes with which words are used in a text and the circumstances under which they are used. This work considers the advantages of using corpora, carefully and meticulously designed collections of written texts, in teaching a foreign language, thus promoting cross cultural competency. Recent development in communication and the requirements for a growing intercultural cooperation bring demands on today's society to be able to communicate in an effective way. Languages are the most prominent part of one's culture, that is why it is imperative to master them as well as possible to avoid ambiguities and misunderstandings. This paper shows an introduction of a possible approach to accomplish this task, demonstrating methods and views upon a tiny fraction of features of language utilizing available data concentrating on some aspects of using an online corpus, namely the British National Corpus, to provide a communicative context in language learning, focusing on English as a second language with special emphasis on English for business and economics.

Key words: corpus linguistics; language learning; business English; sketch engine; concordance; collocation; language methodology

Introduction

English language achieved a genuinely global status because it formed a distinct role that is recognized in every country. The role of the English language is most evidently manifested in countries, where it has obtained a priority in a country's foreign language teaching, even though this language has no official status. It becomes the language which children are most likely to be taught when they enroll in school, and the one most available to adults, who are seeking to learn a foreign language. (Crystal, 2003)

Countless materials and resources written in English or translated from English are being used today in various areas. The quantities of texts to be translated because of the integration into the European Union or the NATO are constantly increasing, requiring a considerable amount of English language materials and documentation. The laws, bank materials, international companies' documentations, ministries as well as educational and cultural institutions use translated materials and translate their own resources. The administration used to be conducted in one

language, but this has changed into an administration in two or more languages. This is the case of calls for proposals, research plans, biographies, recommendations, business plans, spending, records, reports, and many other documents of everyday life.

English language examinations and certificates, such as IELTS or Cambridge English Qualifications are accepted worldwide as proof of English proficiency for higher education, private and public sector employers, and global migration. Furthermore, English being the language of the media industry gives access to entertainment, since many books, movies, TV shows, music, as well as other types of media, are published and produced in English. It is the language of international communication, as well as global business and it makes traveling more convenient.

Indisputably, the necessity to be competent in understanding and interpreting written as well as spoken English language is vital. There are many methods and approaches set to confront this particular demand, one of which corpus linguistics addresses in a most efficient way.

This article considers the advantages of using corpora, carefully designed collections of written texts stored on computers or available online, in teaching of English as a foreign language. With the internet and computers widely accessible in classrooms, using corpora in language methodology is a functional and useful method to improve and enhance the interpretative and productive skills of learners as well as their ability to learn. This article focuses on some aspects of using an online corpus, namely the British National Corpus, to provide a communicative context in English language learning.

What is a corpus?

According to the Corpus Encoding Standard (Ide), a corpus is a large collection of texts of any type, including prose, newspapers, poetry, drama, word lists, dictionaries etc. Corpora can be used to study purposes with which words are used in a text and the circumstances under which they are used. We can investigate how speakers and writers exploit the resources of their language; the language is being looked at in naturally occurring circumstances and cultural settings, looking at ways of similar structures serving distinct functions in various contexts.

Since online corpora and tools for analyzing of corpora are becoming increasingly accessible, corpus-based research as well as studies have become progressively more common as are their uses in teaching languages and language pedagogy.

Written and spoken varieties can be investigated regarding their preference determining patterns in language use, individual style of given authors or speakers

depending on social background or different situations, registers. Monolingual as well as multilingual dictionaries are a good collection of information on meanings that can be expressed by a word and on its translations.

However, dictionaries cannot list all possible contexts, in which a word can be used, and which can influence the translation of the given word into another language. In teaching English as a foreign language, it is advisable to adopt a communicative approach, in which words are not treated as isolated units or references to concepts, but as means of communication between the speaker and the hearer, concentrating on the meaning and usage of the given word or phrase. As Howatt (Howatt, 2004) puts it, language is experienced as a communicative behavior, which is why the main focus in English language teaching is put on the communicative competence of learners.

In this sense, I am not going to focus on the logic-based theories of meaning according to which a word has a meaning as an entry in a dictionary, but rather on the approach assuming that the meaning of a word is the function of its purposeful use in communication, the exchange of information and meanings between the speaker and hearer or the producer and receiver.

Corpora are a useful instrument for both teachers and learners to develop communicative, linguistic as well as cultural competence and at the same time improving the learning skills and productive and interpretative competence (Aston, 2001). We must take into consideration that learning a language is a matter of learning about how the language is used and learning how to use it, develop learning skills.

That is why corpora are a valuable tool in developing these skills leading students toward greater autonomy, providing teachers as well as learners with information about the culture of a given language as a helpful complement to other teaching aids.

Contextualizing the Learning Process with Help of the British National Corpus

As an example, let us look at one of the largest available corpora, the British National Corpus (The British National Corpus, n.d.), which was originally created by Oxford University Press containing 100 million words.

According to the website of the British National Corpus (The British National Corpus, n.d.), the written part of the BNC (90%) includes, for instance, extracts from regional and national newspapers, professional magazines and journals for all ages and interests, academic books, and popular fiction, published and unpublished letters and communications, school and university compositions, along

with numerous other kinds of text. The spoken part of the British National Corpus (The British National Corpus, n.d.) (10%) consists of orthographic transcriptions of unscripted informal exchanges (recorded by volunteers chosen from different age, region, and social classes in a demographically balanced way) as well as spoken language composed in different contexts, varying from formal business or government meetings to radio shows and phone-ins.

The examined sample of investigated words has been chosen based on a brief survey conducted over a period of time among learners of English as a second language from various European as well as non-European countries, focusing on business English, which produced a sample of words of interest in addition to words causing problems in case of translating them from English to their respective languages correctly. Due to the space and time constraints, only a few words had been chosen. One of the recurring statements of the enquiry respondents was the occurrence of words of English origin used in their respective first languages in business situations, e.g., startup, venture, hub etc. and related words which generally caused confusion and struggle in determining their exact meanings, using them in a correct way and translating them or more importantly, understanding their meanings in other languages accurately. These are the investigated words in this study.

The Corpus Processing Tools Used for the Purposes of this Study

The Sketch Engine (www.the.sketchengine.co.uk), is a practical tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in the language and what is rare, unusual, or emerging usage. The Sketch Engine as part of the British National Corpus (BNC), is for anyone wanting to research how words behave. It is a Corpus Query System integrating concordances, word sketches (corpus-derived summaries of a word's grammatical and collocational behavior), distributional thesaurus, collocations, and many other useful features. As Kilgarriff and Tugwell (Kilgarriff) state, this information can be obtained from any large corpus, provided there are part-of-speech taggers (i.e. software that reads text in some language and assigns parts of speech to each word), lemmatisers (i.e. software which removes inflectional endings to return the base or dictionary form of a word, the lemma) and grammars available for the language of the corpus. Word Sketches build on recent developments in lexicography, corpus linguistics and Natural Language Processing to provide an improved way for lexicographers, terminologists, teachers, and learners to find out what the corpus has to say about a word. (Kilgarriff, p. 136)

In Sketch Engine, the British National Corpus has always been one of the principal corpora, both in terms of usage and demonstration of analytical tools and

technologies in Sketch Engine, including concordance search, word lists, collocation extraction, word sketches and distributional thesaurus among its other features. That is why it is used in this study. The advantages of using British National Corpus are in the fact that the aforesaid corpus is (1) monolingual: deals with modern English, not any other languages although some foreign language words may occur in the corpus; (2) synchronic: it comprises British English of the late twentieth century; (3) wide-ranging: it includes different styles and varieties, does not focus on any particular subject area, genre or register, also it contains spoken as well as written styles. (<http://www.natcorp.ox.ac.uk/corpus/>, www.the.sketchengine.co.uk)

Words in Context based on the British National Corpus

The starting point for the following search, is a word cloud of business English terms generated by the Sketch Engine for Language Learning distribution thesaurus. (<https://skell.sketchengine.eu>)

SkELL (Sketch Engine for Language Learning) is a simple tool for students and teachers of English to easily check whether or how a particular phrase or a word is used by real speakers of English. All examples, collocations and synonyms presented by SkELL are identified automatically by complex algorithms and advanced software analyzing large text corpora.



Fig. 1: Word cloud of "startup" in the SKELL Thesaurus

We will choose a word of interest, in this case the word "startup" (one of the words identified above, belonging to the group of business English related words with a high frequency), and look at its occurrence in the corpus. Concordance is a very useful way to investigate large text collections (or text corpora) via results

of various search queries put in context. Each concordance line shows one occurrence of a search query, usually in the middle and in red color, typically referred to as KWIC (key word in context). (www.the.sketchengine.co.uk)

For illustration purposes, the following is a shuffled sample concordance of the word “startup” from the British National Corpus, as represented by the BNC interface (with bibliographic sources highlighted on the left side column, logged in abbreviations, all fully expandable for further insight):

9	HAC	W_pop_jore	A	B	C	little diamond shaped dot, but it produces a beep quite happily. # The startup group contains the programs you want each time you open Windows # To make the
10	FA3	W_ac_polit_law_edu	A	B	C	literature on the academic profession in this country (see Halsey and Trow 1971; Startup 1979; Whitburn et al. 1976), but again this deals only indirectly
11	CSG	W_non_ac_tech_engin	A	B	C	other officers of the company who said in some instances Informix could provide backing for startup firms, such as those which may have developed important tech
12	CTJ	W_non_ac_tech_engin	A	B	C	is shut down unexpectedly, the file system is not damaged. It also reduces startup time to under 40 seconds. For the software developers, there is a new
13	CTP	W_non_ac_tech_engin	A	B	C	AG's print manager and NetLabs Inc's NetLabs/Manager. Add backup and restore, startup and shutdown and software installation to that for starters. Whilst DCE get
14	K8W	W_commerce	A	B	C	of regulation are front-end loaded, with the majority of costs being met with the startup of the SROs, subsequent running costs being minimal compared with the tu
15	HB2	W_misc	A	B	C	their well,' Smith notes.' The installation of platform equipment and the startup of the well went smoothly.' Broussard says other opportunities to tie-in deep water
16	HBK	W_misc	A	B	C	compared against identified criteria. This testing and results were the corner stone of the startup case which allowed the Unit to begin generating before the end of

Fig. 2: Concordance and context of “startup” in the BNC

The advantage of corpus concordancing is in its demonstration of the investigated word in a context. A concordance is the fundamental tool for any researcher working with a corpus showing them what is in the corpus. In case of a more in-depth investigation, it is possible to expand a particular sentence containing the word of interest, as follows:

Learning a language involves gradual adoption of the target language as well as its culture, using corpora as a useful aid can be well utilized in order to provide learners with the most contexts possible and frequency of occurrence, i.e., surrounding them with the language learnt. A corpus is one of the means of achieving this goal. Compared with the dictionary’s broad generalizations, a corpus provides a wider range of patterns, examples thus revealing more specific aspects on the use of a particular word.

In order to discover the meanings of words, or narrow down their implications, the corpus can be used to show collocations, i.e., the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance as well as word combinations. Collocational and grammatical behavior of a particular word and its grammatical relations, i.e., words serving as subjects or objects, modifiers and adjectives etc. are a significant help in discovering meanings of problematic words and corpus is a reliable and flexible tool used to show collocations at a click of a button as follows:

The collocation candidates are sorted and arranged according to their logDice score, a logarithmic variant of the dice coefficient (a statistic used for comparing the similarity of two samples), which differentiates between salient candidate pairs and a sheer coincidence of some co-occurring ones, but it is crucial to note

Source information:

Date	(1985-1994)
Title	UK financial institutions and markets. Pawlet, Michael; Bentley, Patrick; Winstone, David. London: Macmillan Press Ltd, 1991, pp. 105-261. 2043 s-units.

Expanded context:

“(1987) argues that most regulation is simply added on to existing frameworks (often as the result of specific frauds or crises) rather than building up new systems, and that the efficiency of regulation suffers as a result and costs escalate. The direct costs of financial regulation are estimated at over 100 million per year in running costs (Lomax 1987), and some (e.g., Goodhart 1987) would regard this as a conservative estimate. It is often argued that the costs of regulation are front-end loaded, with the majority of costs being met with the **startup** of the SROs, subsequent running costs being minimal compared with the turnover of most financial institutions. It would seem however that the costs hit small financial firms far harder than larger ones, and may create a powerful barrier to entry to new firms. Independent advisers regulated under FIMBRA probably feel the costs of regulation the hardest, as they are generally small in size. Concerning costs, it has been reported in the financial press that FIMBRA merely had to take the word of Dunsdale that it was investing in...”

(The British National Corpus, n.d.)

Fig. 3: Expanded context in the BNC

startup noun Show context

verbs with startup as subject	verbs with startup as object	adjectives with startup	modifiers of startup
1. launch startup launched	1. fund funded startups	1. such startups such as	1. tech tech startups
2. fail startups fail .	2. launch launching a startup		2. early-stage early-stage startups
3. call a startup called	3. found startup founded by		3. Silicon a Silicon Valley startup
4. need startups need to	4. help help startups		4. lean the lean startup
5. look startup looking	5. base startup based in		5. high-tech high-tech startups .
6. get startups get	6. support support startups		6. technology technology startups
7. use startups using	7. build building a startup		7. promising promising startups
8. do startup does	8. call startup called		8. Valley a Silicon Valley startup

Fig. 4: Collocation candidates and Word Sketch of “startup” in the BNC via Sketch Engine

and observe all the association scores necessary to take into consideration. The association scores are, however, not the topic of this paper, that is why they are not going to be dealt with in detail.

The co-occurrence score shows the frequency of two terms occurring along with each other in the corpus in a particular order; it is an indicator of semantic

proximity or an idiomatic expression and adopts the interdependency of the two terms.

The candidate count shows the number of occurrences of the particular word in the investigated corpus.

Upon further examination of the corpus one can learn that certain items appear in certain types of context, the tendency of an item to occur in contexts which have a particular function, in a particular structural or lexical environment or in a particular lexical field. Clicking on (or simply hovering over) select words given by the Word Sketch as co-occurring with the investigated word, learners can discover wider contexts, for example, that “Ordinary investors will soon have the opportunity to invest in *early-stage startups* in exchange for company stock”, “Websites are an internet showcase for a *startup company*”, “Evonik plans to invest 100 million in *promising startups* with break-through technologies and leading specialist venture capital funds”, “The tool was developed by a San Francisco-based *startup* called Entelo” and so forth. This is another effective approach in contextualization of the language learnt, the communicative way making the understanding and learning easier and more straightforward for the students.

While examining the words “startup” and its collocations, the investigator can narrow the search further and make the query more interesting and noteworthy as well as communicative for the students by looking for expanded context, where any two particular co-occurring words are used to make sure students discover and appraise the exact meaning of the word in context. The students probably have their own ideas connected to the uses of the featured words and it is very useful to let them discover those types of contexts, where the examined words might occur.

Looking at the word sketch of “startup”, students can click on any word that catches their interest and the need to examine it and its relationship and context further. The words listed in first places are usually the first and most frequent candidates for investigating the context in more detail.

Here is an example of extended context featuring the words “startup” and “fund”, where “fund” is a frequently occurring verb with “startup” as object:

Here is an example of extended context featuring the words “startup” and “tech”, where “tech” is a frequently occurring modifier of “startup”:

Here is yet another example of extended context featuring the words “startup” and “incubator”, where “incubator” is a frequently occurring noun modified by “startup”:

fund + startup 0.02 hits per million

1. Now they're doing something new : planning to **fund** biotech **startups** .
2. The same trend continued during the first six months of 2013 with only three **startups funded** .
3. And heavily **funded startups** such as Actifio and DataGravity are trying to make data-storage systems smarter.
4. And we haven't had much help when it comes to **funding** the **startup** of ObamaCare.
5. This work was fruitful and advanced the technology to a level at which venture capital **funded startups** could develop commercial products.
6. With the papers in hand, Pavele and Moses approached several venture capital firms that showed interest in **funding** a Macsyma **startup** .
7. Investors like me would **fund startups** , and not expect to see the payoff until five to 10 years down the line.
8. Silicon Valley venture capitalists have **funded** several food-related **startups** in the past year, but Hampton Creek has gathered the most momentum.
9. Evernote is a well **funded startup** , with large technical talent so a breach of this scale would require a some skill.
10. However, this year there is a significant upward trend in new semiconductor startup funding with almost 20 new **startups funded** from January through June.

Fig. 5: Expanded context in the BNC via Sketch Engine

tech + startup 0.12 hits per million

1. Running a **tech startup** is incredibly different from running a large company.
2. PMs are a vital part of a **tech startup** .
3. So you might have heard of this little **tech startup** called Google.
4. It 's little surprise then that investment in clean **tech startups** is shrinking.
5. The will cater to **tech startups** interested in flexible workstations or dedicated office space.
6. Advertisement It's not unusual for a **tech startup** to raise \$18 million.
7. For now , many of these **tech startups** attract a niche group of borrowers.
8. The new website features an interactive map of New York City 's **tech startups** .
9. She also took part in several discussion panels involving **tech startups** over the past few months .
10. Unfortunately for them , **tech startups** usually don't take the side of petulant children.

Fig. 6: Expanded context in the BNC via Sketch Engine

Learners can discover relationships between words of interest and even prove or disprove their own ideas and opinions as to the occurrence as well as meaning of the investigated words in certain contexts. The examples shown above give the learners the expected context, namely a situation, where the particular words naturally occur, and they automatically come to mind as first when investigating a certain combination of words. This procedure can be done at class with many different groups of words making the learning and understanding easier and more straightforward. In order to reach a communicative competence, the use of expanded context is vastly effective. As Meyer (Meyer, 2002) states, corpora consist of texts (or parts of texts), thus they enable researchers, teachers, or learners to

1. Such support infrastructure as clusters, co-working spaces, **startup incubators** and accelerators have been emerging.
2. Perhaps most intriguingly, some individuals associated with Y Combinator, the **startup incubator** that hatched Reddit in 2005, are also in the mix.
3. Combining the best from an innovation lab, learning community and **startup incubator**, Impact Hub Zürich provides a collaborative space to prototype the future of business.
4. **Startup incubators** are sprouting left and right, eager to jump on promising startups such as AVA.ph, a premium fashion retailer based in Manila.
5. Esri, the world leader in geospatial technology and mapping software, welcomed 1776, a **startup incubator** in Washington, D.C, today as the newest addition to the Esri Emerging Business Program.
6. Last week Y Combinator, the well-known and very successful technology **startup incubator**, announced that it was going to begin experimenting with biotech startups.
7. He serves as a legal mentor to companies in the StartX program, formerly SSE Labs, a **startup incubator** affiliated with Stanford University, according to Cooley.
8. The first location will be in the Galvanize **startup incubator** in San Francisco's Sout of Market neighborhood, which is home to more startups per square foot than anywhere else in the world.
9. There are hundreds of **startup incubators**, accelerators and innovation conferences and thousands of "pitch events" to meet many of the 100,000 active startups – with more being founded every day.
10. Upstart Labs, an early stage fund and **startup incubator** in Portland, is continuing to churn with added funding, new investments, and an expanded team.

Fig. 7: Expanded context in the BNC via Sketch Engine

contextualize language and can facilitate functional discussions of language as well as understanding certain cultural situations e.g., what kind of language is used in a given cultural and social context.

Considering the word “startup”, students may investigate further looking upon the word sketch, where “startup” may lead to discovering different words of interest in their context with “startup” being the starting point for further research. The above-mentioned word “incubator” may spark interest in learners of languages by questioning its meaning in their first language, wherein the particular word might not be associated with a “startup”. This can easily lead to the investigation of the word “incubator” itself, leading to further results as follows:

Corpora may prove to be a very useful tool of linguistic research for teachers as well as students. In order to highlight and identify specific lexical or grammatical features or patterns, it is helpful to use concordancers, i.e., computer programs which automatically construct concordances for a more effective use of the corpora. The Sketch Engine for Language Learning was used for the purposes of this paper, but there are many other interfaces that students as well as teachers can use, which provide straightforward and easy to read examples and contexts, that can be used at a click of a mouse to discover vast amounts of reliable data in a communicative framework.

The distinguishing advantage of using a corpus over an internet search engine like e.g., Google, is the fact that a corpus, like the British National Corpus or a similar body of text, is built by linguists and professionals, it is meticulously compiled and tested, granting it considerable relevance and applicability.

Conclusion

This article seeks to briefly present arguments in favor of the use of corpora in English language teaching since it can provide learners with information and knowledge in a communicative form and readily available. Corpus use can pro-

incubator 1.72 hits per million

1. This effect is further reinforced through the virtual **incubator** network.
2. Several private companies operated feed mills and **incubators** .
3. **Incubator** space has been set aside for startups with big potential.
4. Tech **incubators** and accelerator programs kept popping up.
5. God offers a lot of coral reef **incubators** .
6. Having an **incubator** with oxygen support is recommended.
7. How much more room will those **incubators** take?
8. The **incubator** space remains intact and will continue.
9. An industrial **incubator** will completed its action.
10. The boardroom is an **incubator** for evil .

Fig. 8: Concordance and context sample in Sketch Engine for Language Learning

vide opportunities for learning in many areas, it can complement the immersive teaching process in a motivating way, promoting language use and improving the understanding by wider available contexts.

Corpus linguistics being a relatively new scientific domain is growing fast and the materials developed for research are constantly being improved and enhanced. The great advantage of working with corpora is a range of almost endless possibilities for research and also a very fast and effective response in using computer tools for corpus processing. That is why it is a fairly easy task to gather substantial amounts of data to be able to process and extract relevant and applicable information for conclusive results.

It is evident, as mentioned in this work, English is a global language, it is the language of international business and politics and it is here to stay. That is why it is crucial to utilize and attain the most effective and valuable means of studying it and managing it, not just discovering efficient ways to make it understandable for the ever-expanding audience and users, but also means of finding the best practices in interpreting and understanding it, while keeping in mind that more

and more non-native speakers use English as the main language of their communication. One of these means is offered by corpus linguistics.

Recent development in communication and the requirements for a growing intercultural cooperation bring demands on today's general public to be able to communicate in an effective way. Languages are the most prominent part of a society's culture, that is why it is imperative to master them as well as strive to avoid ambiguities and misunderstandings. This work shows just the introduction of a possible approach to fulfil this task, demonstrating methods and views upon a tiny fraction of features of language utilizing available data.

References

- ASTON, G. (2001). *Learning with Corpora*. Houston: Athelstan.
- BAKER, M. (1993). Corpus Linguistics and Translation Studies – Implications and Applications. In M. B. (eds), *Text and Technology. In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins Publishing.
- CRYSTAL, D. (2003). *English as a Global Language*. Cambridge: Cambridge University Press.
- HOWATT, A. P. (2004). *A History of English Language Teaching*. Oxford: Oxford University Press.
- IDE, N.,-D. (n.d.). *Corpus Encoding Standard*. Retrieved 02 11, 2011, from <https://www.cs.vassar.edu>
- KILGARRIFF, A. T. (n.d.). *Sketching Words*. Retrieved 02 1, 2016, from www.kilgarriff.co.uk/Publications
- Lexical Computing. (2021, January 25). *SKELL*. Retrieved from SKELL: <https://skell.sketchengine.eu/#result?f=wordsketch&lang=en&query=startup>
- MEYER, C. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- The British National Corpus*. (n.d.). Retrieved 03 22, 2017, from <http://corpus.byu.edu/bnc/>

Author

Michelle Duruttya, M.A., Ph.D., e-mail: michaela.duruttya@vse.cz, Prague University of Economics and Business

She is an English language teacher, translator and linguist based in Prague, Czech Republic. She cooperates with a number of Czech as well as international educational institutions as teacher trainer and teacher of English language. She has taught English language in Hungary, Slovakia, UK and China for more than 15 years, while also being an Oral Examiner of all the main suite Cambridge Examinations for the British Council. Currently, she works for the Prague University of Economics and Business and cooperates with the Czech Academy of Sciences as well as several language schools as teacher of English.