Designing a Corpus Workbook for Students of Czech as a Foreign Language

Adrian Jan Zasina

1 INTRODUCTION

Language corpora are increasingly utilised in language learning around the world. The interest in using corpus methods for learning Czech appeared at the beginning of the 21st century. In 2005, a team of scientists from the Institute of the Czech National Corpus (Čermák et al., 2005) published the first study guide on how to use the Czech corpus and recommended it as a supplementary source for students from primary and secondary schools and universities. Two years later, Šulc (2007) described in a short guide for schools how to work step by step with corpus data; he explained the possible functions and usage of the Czech corpus. Over the next four years, the first works concerning corpus use in teaching Czech as a foreign language (CFL) started to appear (Vališová, 2009; Lukšija, 2010; Osolsobě, 2010). Since then, several studies have appeared on corpora use in CFL teaching (cf. Lukšija, 2012; Vališová, 2012, 2016; Kočařová, 2013; Konečná & Zasina, 2014; Zasina, 2018b). The last few years have seen publications aimed at Czech teachers and students of teacher education for primary and secondary schools; these works (Šmejkalová & Kopřivová, 2019; Šormová et al., 2019) attempt to incorporate corpus methods developed for teaching native speakers of Czech. Zasina's (2019) comprehensive study describes the most prominent problematic areas of non-native Czech based on a learner corpus and offers a coherent methodological framework for creating corpus-based exercises.

Despite a great deal of interest in corpus approaches to teaching, no substantial collection of corpus-based exercises has been created. Teachers are reluctant to use corpora in the classroom as it is a time-consuming task requiring computer skills and knowledge of the potential of corpus tools that might not be user-friendly (Bennett, 2010, pp. 206–207; Tribble, 2015, pp. 57–59; Leńko-Szymańska, 2017, p. 217). Thus, it is especially important to provide teachers with suitable learning materials that they could easily use. Giving them ready-to-use materials could help them to actively apply corpora in teaching.

This paper aims to familiarise teachers with the method of data-driven learning, describe the design of the corpus workbook, and recommend examples of corpus-based exercises.

2 DATA DRIVEN LEARNING: PROS AND CONS OF USING CONCORDANCE IN TEACHING

Using a corpus in foreign language learning was termed *data-driven learning* (DDL) by Tim Johns (1991). In his seminal work, DDL is defined as an:

> approach to foreign language learning that takes seriously the notion that the task of the learner is to "discover" the foreign language, and that the task of the language teacher is to provide a context in which the learner can develop strategies for discovery — strategies through which he or she can "learn how to learn". (*ibid.*, p. 1)

As proposed by Johns (1991, p. 4), it is an inductive approach to language learning built on three pillars: identification — classification — generalisation. Students identify a language problem they would like to solve; then they observe a concordance with examples to follow language features and classify them; finally, they generalise rules about how a language phenomenon works. It means that students do not obtain information, for example, about the function and formation of the accusative case, but instead must derive it based on concordance observation: following concordance lines, students identify that the accusative case has its objective function and specified endings for each grammatical gender.

The usage of concordance in teaching is often underestimated and criticised. According to Boulton and Cobb (2017, p. 351), many teachers and students criticised the use of concordances due to an unfavourable attitude towards approach to working with computers in general; for others, results displayed in the form of a classical concordance are disturbing as it truncates sentences to centre the key word, which goes against most students' and teachers' experience of promoting meaning analysis; corpora represent an authentic language that may be beyond the scope of the learner's language knowledge; DDL activities require at least basic skills in corpus usage, and are time-consuming because learners formulate their own rules based on data observation.

In my opinion, these misgivings may come from fear of the unknown, uncertainty concerning the advantages of corpus methods in language learning, and non-user-friendly corpus tools (cf. Leńko-Szymańska, 2017). The benefits of concordance displays lie in their focus on a particular language phenomenon; learners are not distracted by having to read whole sentences, and they may restrict their observation to the target phenomenon. Reading the whole sentence is a habit rather than a necessity; we may often limit ourselves to the closest context to understand the meaning. In addition, vertical reading helps to find more repeating patterns (such as collocations) for a search word (cf. Boulton, 2009; Ballance, 2017). This approach also forces learners to think about what they are learning, which has a positive influence on retention of new knowledge in memory (Bernardini, 2004; Römer, 2011). However, it must be admitted that authentic language in a corpus can be demotivating, especially for beginners. Therefore, it is crucial to adjust corpus-based exercises to target the learner group and provide ready-to-use activities that do not always require direct work with a corpus and too much information. Thanks to that, each user can easily choose whether to work with a computer or to do on-paper exercises. Concordance is a key element here, as was already postulated in an early stage of DDL (Johns, 1991, p. 2) and is still actively used in many studies (Bennett, 2010; Boulton, 2010; Liontou, 2020; Szudarski, 2020). However, nowadays, concordance-based exercises are not the only solution; we may also work with frequency lists

(cf. Nation, 2016), collocations (cf. Vyatkina, 2016), and many other functions provided by corpus browsers and tools (cf. Zasina, 2018a).

The strong point of DDL lies in the authenticity of language usage and the inductive approach that engages the learner in individual work and develops their independent thinking. Moreover, this method makes it possible to adapt classroom materials to the specific purposes for any group of learners. Therefore, the workbook presented tries to deliver easy-to-use materials for teachers and students who can, without much knowledge of corpus linguistics, simply start to use corpora in teaching and learning.

3 DESIGN OF THE WORKBOOK

The workbook is suitable for university students of CFL. It was set based on exercises that were prepared within an academic course on *Grammatical and Lexical Corpus-based Exercises* at Charles University. The aim of the course was to teach students how to use language data in language learning — DDL. The same procedure was adopted in the workbook, where specific exercises guide students through three steps (identification, classification, generalisation).

The workbook consists of hands-on and hands-off exercises (Boulton, 2012). Hands-off exercises are suitable in a classroom without access to a computer. This makes it possible to use corpus-based exercises in facilities without access to a computer study room, thereby making the corpus approach more accessible. Hands-on exercises cover direct corpus activities with corpus tools and classical gap filling exercises that use corpus data displayed in the form of frequency lists or collocations. There are also supplementary non-corpus exercises. Each exercise is marked by a symbol indicating the type of exercise (direct corpus, indirect corpus, non-corpus). Exercises that involve direct computer work are linked with a specific online tool that is required to solve the exercises.

The workbook is designed to help teachers and students to choose topics that are of interest. There is no need to proceed chapter by chapter. It is up to the user to plan their own study programme. The workbook consists of a short introduction and more than 100 exercises with a key. It is a collection of corpus-based exercises; therefore, there are no detailed instructions on how to work with corpus data and corpus tools. However, the introductory part features a short description of corpus tools with links to manuals which are available on-line. There are two reasons to not include manuals in the corpus workbook. First, they are already described and easily accessible in Czech and English on the website https://wiki.korpus.cz/doku.php/manualy. Second, corpus tools are under continuous development to improve their functionalities, and it is better to get acquainted with the most current version. In addition, brand-new tools can appear as well. Thus, each user of the workbook should first become familiar with basic work with corpus tools. Fortunately, the direct exercises in the workbook use user-friendly tools that are mostly intuitive.

Due to the authenticity of the corpus material, the workbook is recommended for higher intermediate students starting from level B1, although some exercises could

be introduced to beginners as well; in that case, the teacher's assistance is required. Individual teachers know their students best and may select suitable corpus exercises for their classes. It is also possible to modify a task to adjust its form to the students' needs. Therefore, all exercises offered should be taken as a source for inspiration.

3.1 THEMATIC AREAS

The workbook is divided into 14 chapters. Each chapter represents a specific linguistic issue focused on the following thematic areas: difference in meaning (word meaning based on context), grammatical gender, declension, nominalised adjectives, hard and soft adjectives, pronouns, prepositions, prefixes (*s-* vs. *z-*), past participle, collocability, idioms, stylistic variants, the competing endings *-a* and *-u* in the genitive singular, and vowel length. The topics were organised around the content of the academic course *Grammatical and Lexical Corpus-based Exercises*, previous in-depth analysis of the most problematic areas of foreign Czech in learner corpus (Zasina, 2019), experiences with university teaching abroad (Zasina, 2022), and on my personal experience with teachers' needs.[1]

3.2 EXERCISE EXAMPLES

This subsection zooms in on exercise examples included in the corpus workbook. The following tasks present two corpus-based exercises concerning the dative case, and collocability. The exercises are on-computer activities (hands-on). They make use of data from the SYN2020 corpus (Křen et al., 2020).

Exercise 1
This exercise concerning the dative case represents the chapter on declension. The dative case is (along with the vocative) one of the less frequent cases in Czech; its most common uses are to express the addressee as indirect object or a single object governed by a verb requiring a dative construction (Cvrček et al., 2010, p. 139). There are also several typical prepositions used in this task that require this case. Students work with a provided list of the five most frequent dative prepositions as identified in a corpus of contemporary written Czech. Their task is to search in the corpus for the prepositions and find two example sentences for each of them. Through observing the concordances, students have the opportunity to familiarise themselves with each preposition and its meaning. Then, students share their ideas in pairs. In the final step, the teacher discusses examples with the whole group to systematise the new material.

Task. *Podívejte se na pět nejčastějších dativních předložek v níže uvedené tabulce. Zkuste je vyhledat v korpusu současné češtiny a vypsat dvě příkladové věty pro každou z nich.*
'Take a look at the five most common dative prepositions in the table below. Try to find them in a corpus of contemporary Czech and write two example sentences for each of them.'

1 Special thanks to Barbora Hrabalová for her comments and recommendations.

No.	*Předložka* 'Preposition'	*Frekvence* 'Frequency'
1	*k* 'to'	545,155
2	*proti* 'against'	53,821
3	*kvůli* 'due to'	42,560
4	*díky* 'thanks to'	40,362
5	*vůči* 'against'	12,564

TABLE 1: Seznam nejčastějších dativních předložek na základě korpusu SYN2020 'List of the most frequent dative prepositions based on the SYN2020 corpus'

Exercise 2
This exercise presents an example from a chapter dealing with collocability. All exercises work with the co-occurrence of two words, mainly adjectives with corresponding nouns. In this chapter, students can work with the Word at a Glance tool (Machálek, 2020b), a word profile aggregator displaying collocations and many other details. Alternatively, in the exercise below, students can search for collocations using simple frequency lists or the collocation candidates function of the KonText corpus browser (Machálek, 2020a). The following task exemplifies the use of the collocation function. First, the students search the corpus for the nouns given in the exercise. Second, they choose the "collocations > custom" option from the list in the KonText browser to identify the most likely adjectival collocates. They can specify the parameters of the collocation search; students must ensure that *lemma* is set as an attribute, the collocation windows span is <–5, 5>, and results are sorted using the logDice collocation measure. Thanks to this setup, they will be able to identify the most usual collocations. The task helps students to understand the collocational behaviour of the word and teaches them autonomy in learning. Teacher assistance is not required; therefore, this exercise may be done as homework.

Task. *Na základě dat z korpusu s použitím funkce KOLOKACE zjistěte, která níže uvedená adjektiva se pojí s jednotlivými substantivy:*
'Using the COLLOCATIONS function, please use corpus data to identify which adjectives co-occur with individual nouns:'

historický, žhavý, starý, širý, nutný, lidský, modrý, odborný, přirozený, dopravní

..... *nebe,* ..... *podmínka,* ..... *škola,* ..... *jádro,* ..... *prostředí,*

..... *muž,* ..... *nehoda,* ..... *oči,* ..... *novinka,* ..... *život*

4 CONCLUSION

Integrating corpus methods into regular teaching is a challenge, as it is a new approach that requires a change in thinking, not only for teachers, but also for learners.

Teachers and learners need to know exactly what kind of advantages the DDL method brings. Several studies (Vališová, 2012, 2016; Zasina, 2018b, 2019) have already explained the role of DDL in CFL learning and proposed concrete corpus-based activities. However, it is still necessary to promote this approach and to meet the requirements and expectations of potential users. Furthermore, the possibility that computers might not be available during regular classes should also be taken into account; on-paper exercises could then be a useful introduction to DDL.

As Boulton and Cobb's (2017, p. 386) meta-analysis concludes, "DDL works pretty well in almost any context where it has been extensively tried", therefore a corpus workbook with a broad range of materials is an opportunity to take a first step into the use of corpus in teaching and learning Czech. The workbook delivers ready-to-use comprehensive study materials for CFL students and their teachers. Although previous publications (Šmejkalová & Kopřivová, 2019; Šormová et al., 2019) tried to motivate teachers to use corpus methods in language teaching, they only focused on teaching native speakers of Czech, and they did not address issues in CFL teaching. This publication pays attention to the most prominent obstacles in the Czech language for foreigners, which are different in many respects from the mistakes made by native users. The rich offer of corpus-based exercises can provide a basis for university course content dealing with the DDL method. It can also serve as an inspiration for teachers to prepare their own corpus-based exercises meeting the requirements of a specific group of learners. Last but not least, all corpus exercises are primarily intended for CFL students to help them resolve difficulties and understand the principles of the Czech language. This material is a starting point that needs continuing development; therefore, I do believe it will provide an impulse that contributes to popularising the DDL method in CFL teaching and learning.

Acknowledgements

This paper was supported by the project International mobility of research, technical and administrative staff at the Charles University no CZ.02.2.69/0.0/0.0/18_053/0016976. I would also like to extend my gratitude to Neil Bermel for helpful comments and proofreading.

REFERENCES

Ballance, O. J. (2017). Pedagogical models of concordance use: Correlations between concordance user preferences. *Computer Assisted Language Learning*, *30*(3–4), 259–283.

Bennett, G. R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. University of Michigan Press.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. McH. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 15–36). John Benjamins.

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, *21*(1), 37–54.

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

Boulton, A. (2012). Hands-on / hands-off: Alternative approaches to data-driven learning. In J. Thomas & A. Boulton (Eds.),

Input, Process and Product: Developments in Teaching and Language Corpora (pp. 152–168). Masaryk University Press.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348–393. DOI

Čermák, F., Blatná, R., Klímová, J., Kopřivová, M., Kučera, K., Petkevič, V., Schmiedtová, V., & Šulc, M. (2005). *Jak využívat Český národní korpus*. Nakladatelství Lidové noviny.

Cvrček, V., Kodýtek, V., Kopřivová, M., Kováříková, D., Sgall, P., Šulc, M., Táborský, J., Volín, J., & Waclawičová, M. (2010). *Mluvnice současné češtiny 1: Jak se píše a jak se mluví*. Karolinum.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*, *4*, 1–16.

Kočařová, B. (2013). *Korpus a jmenný rod ve výuce češtiny jako cizího jazyka* [Bachelor's thesis]. Masaryk Universisty, Faculty of Arts. Available at https://is.muni.cz/th/fvi7f/.

Konečná, H., & Zasina, A. J. (2014). Studium českého jazyka a internet. In E. Rusinová (Ed.), *Přednášky a besedy ze XLVII. běhu LŠSS* (pp. 104–112). Masaryk University.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: Representative corpus of written Czech*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. Available at www.korpus.cz.

Leńko-Szymańska, A. (2017). Training teachers in data driven learning: Tackling the challenge. *Language Learning & Technology*, *21*(3), 217–241.

Liontou, T. (2020). The effect of data-driven learning activities on young EFL learners' processing of English idioms. In P. Crosthwaite (Ed.), *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners* (pp. 208–225). Routledge.

Lukšija, M. (2010). *Korpus jako zdroj dat při prezentaci předložek do/na s místním směrovým významem ve výuce češtiny pro cizince* [Bachelor's thesis]. Masaryk University, Faculty of Arts. Available at https://is.muni.cz/th/217240/ff_b/.

Lukšija, M. (2012). *Korpusy a česká deklinace ve výuce češtiny jako cizího jazyka* [Master's thesis]. Masaryk University, Faculty of Arts. Available at https://is.muni.cz/th/217240/ff_m/.

Machálek, T. (2020a). Kontext: Advanced and flexible corpus query interface. In N. Calzolari et al. (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 7003–7008). European Language Resources Association.

Machálek, T. (2020b). Word at a Glance: Modular word profile aggregator. In N. Calzolari et al. (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 7009–7014). European Language Resources Association.

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. John Benjamins Publishing Company.

Osolsobě, K. (2010). Jak se učit česky s korpusem. In E. Rusinová (Ed.), *Přednášky a besedy z XLIII. běhu LŠSS* (pp. 112–119). Masaryk University.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225. DOI

Šmejkalová, M., & Kopřivová, M. (2019). *Jazykové korpusy a elektronické zdroje ve výuce českého jazyka*. Charles University, Faculty of Education. Available at https://futurebooks.cz/books/pedfa_esf_5/?/obalka/.

Šormová, K., Šebesta, K., Gráf, T., Hejhalová, V., & Kukrechtová, B. (2019). *Korpusy v jazykovém vyučování*. Charles University, Faculty of Arts.

Šulc, M. (2007). *Experimentujeme s češtinou. Jak pracovat s korpusem českého jazyka ve školách i mimo ně*. Nakladatelství Lidové noviny.

Szudarski, P. (2020). Effects of data-driven learning on enhancing the phraseological knowledge of secondary school learners of L2 English. In P. Crosthwaite (Ed.), *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners* (pp. 133–149). Routledge.

Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey.

In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning* (pp. 37–62). John Benjamins.

Vališová, P. (2009). *Korpus jako zdroj dat systémového popisu české konjugace při výuce češtiny jako cizího jazyka* [Master's thesis]. Masaryk University, Faculty of Arts. Available at https://is.muni.cz/th/v98z0/.

Vališová, P. (2012). Data-driven learning a výuka češtiny jako cizího jazyka. *CASALC Review*, *2*, 22–39.

Vališová, P. (2016). Korpus ve výuce češtiny jako cizího jazyka — typy cvičení. In I. Starý Kořánová & T. Vučka (Eds.), *Čeština jako cizí jazyk VIII* (pp. 128–141). Charles University.

Vyatkina, N. (2016). Data-driven learning of collocations: Learner performance, proficiency, and perceptions. *Language Learning & Technology*, *20*(3), 159–179. DOI

Zasina, A. J. (2018a). Korpusy językowe w nauczaniu języków obcych — Metoda, narzędzia, praktyka. In M. Jedynak (Ed.), *Specyficzne potrzeby studentów szkół wyższych a nauczanie języków obcych. Tom II. Praktyczne narzędzia* (pp. 110–123). Studium Praktycznej Nauki Języków Obcych Uniwersytetu Wrocławskiego.

Zasina, A. J. (2018b). O problémech nerodilých mluvčích s kvantitou na základě analýzy korpusových dat. In S. Škodová & M. Hrdlička (Eds.), *Čeština jako cizí jazyk v průsečíku pohledů* (pp. 281–298). Charles University, Faculty of Arts.

Zasina, A. J. (2019). *Korpusový přístup ve výuce češtiny jako cizího jazyka*. [Dissertation]. Charles University, Faculty of Arts. Available at https://dspace.cuni.cz/handle/20.500.11956/115540.

Zasina, A. J. (2022). Corpus approach to teaching Czech as a foreign language in university courses. *Bohemistyka*, *XXII*(3), 435–462. DOI

Adrian Jan Zasina | Institute of the Czech National Corpus, Faculty of Arts, Charles University
<adrian.zasina@ff.cuni.cz>
School of Languages and Cultures, University of Sheffield
<a.zasina@sheffield.ac.uk>