



INSTITUTE OF CZECH AND DEAF STUDIES

Faculty of Arts
Charles University



FACULTY OF ARTS
Charles University

Institute of Czech and Deaf Studies

*Typology of corpus-based
exercises and their
automatic generation
using AI*

Adrian Jan Zasina

Prague, 23.05.2025

Acknowledgment

This work was created with the support of the project "Human-centered AI for a sustainable and adaptable society", reg. no.: CZ.02.01.01/00/23_025/0008691, co-financed by the European Union and the project Multilingual Lens: Investigating Large Text Corpora from Different Methodological Perspectives.



Multilingual Lens

Investigating Large Text Corpora
from Different Methodological
Perspectives

Content

1. Data-driven learning
2. Corpus-based exercises
3. Typology of corpus-based exercises
4. Generating corpus-based exercises
5. Conclusion

Data-driven learnig

Data-driven learning

- Tim Johns (1991) – popularised as data-driven learning (DDL)
- the first mentions of the method in 2005 (Čermák et al., 2005)

Three main stages:

- identification of the language problem
- categorization of the language problem
- generalization of rules

Typology of DDL studies

- theoretical underpinnings
- descriptive papers
- surveys and syntheses
- empirical evaluations:
 - learners' attitudes and representations about corpus use
 - learners' behaviour and processes in corpus use
 - effects of corpus use as a learning aid
 - effects of corpus use as a reference resource

DDL in the Czech Republic

- **predominantly descriptive work** (Konečná & Zasina, 2014; Maturová, 2016; Osolsobě, 2010; Šindelářová & Škodová, 2013; Šormová, 2017; Vališová, 2011, 2012)
- **empirical evaluation of students' attitude towards the corpus approach in teaching** (Vališová, 2011; Vališová & Osolsobě, 2012)
- **evaluating the effects of the corpus as a teaching aid** (Zasina, 2022a)
- **missing syntheses summarizing previous work in this area**

Corpus-based exercises

Corpus-based exercises

Corpus-based exercises:

- friendly corpus-based activities that guide users through corpus data and lead them to findings about linguistic patterns,
- teaching materials that serve as language learning materials for students to improve language skills,
- aim to meet users' expectations and enhance their everyday work with the corpus.

Corpus-based exercises – examples

English

- *Concordances in the classroom: A resource guide for teachers* (Tribble & Jones, 1997)
- *Discovering English with Sketch Engine* (Thomas, 2016)
- *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom* (Le Foll, 2021)

Corpus-based exercises – examples

German

- *Incorporating corpora: Using corpora to teach German to English-speaking learners* (Vyatkina, 2020)

French

- *Webové korpusy Aranea. Učebnica pre učiteľov cudzích jazykov, prekladateľov, tlmočníkov, filológov a študentov filologických odborov* (Benko et al., 2019)

Czech

- portál *Pro školy* (Lukešová, 2017)
- *Korpusová cvičebnice pro studenty češtiny jako cizího jazyka* (Zasina, 2023)

Typology of corpus-based exercises

Typologie korpusových cvičení

9 types of corpus exercises can be distinguished:

- observing,
- filling in,
- producing (oral and written),
- comparing,
- searching,
- matching,
- classifying,
- identifying,
- analysing.

(Lukešová et al., 2017; Vališová, 2016; Zasina, 2023)

Generating corpus exercises

AI and exercise typology

- tool: **Chat-GPT** (OpenAI, 2025)
- models: GPT-4o, GPT-4o mini, o3-mini
- *Corpus Linguist* model (Milička, 2024)

AI and exercise typology

Prompt: *Můžeš mi napsat, jaké typy cvičení pro cizince jsi schopen generovat?*
Can you write what types of exercises for foreigners you are able to generate?

- A. Vocabulary exercises
- B. Grammar exercises
- C. Listening exercises
- D. Reading exercises
- E. Writing exercises
- F. Speaking exercises
- G. Pronunciation exercises
- H. Translation exercises
- I. Advanced tasks and activities

Exercises types by *Corpus Linguists*

Prompt: *Mohl bys vyjmenovat typy cvičení, které jsi schopen vygenerovat?*
Could you list the types of exercises you are able to generate?

A. Filling in missing words

- Cloze test (filling in missing words in sentences)
- Filling in words based on context (frequent collocations, conjunctions)
- Filling in prepositions / conjunctions / pronouns

B. Matching exercises

- Matching a word with its definition / synonym / antonym
- Matching a word to a picture
- Matching a sentence with a missing word / phrase

C. Rearranging words into the correct order

- Word order in a Czech sentence (using authentic data)
- Reconstructing broken sentences
- Placing commas in sentences

D. Error Correction

- Finding and correcting typical errors (based on real mistakes made by foreigners)
- Correcting errors in inflection / tense
- Choosing the correct variant from multiple options

E. Exercises on synonymy and collocations

- Replacing words with synonyms according to the corpus
- Choosing the correct collocation („udělat chybu“ vs. „dopustit se chyby“)
- Assigning words to the correct context

F. Exercises on grammatical phenomena

- Declension and conjugation according to context
- Completing the correct form of the word
- Recognizing parts of speech in sentences

G. Exercises on text comprehension

- Questions on authentic sentences / texts from the corpus
- Working with the occurrence of keywords in real texts
- Searching for specific grammatical phenomena in the text

Analysis

- comparison of two types of exercises
- 2 target levels (A2 and B2)
- 2 target groups with regard to language (Poles and Koreans)
- 4 exercise variants:
 - a) exercise for Polish native speaker at level A2
 - b) exercise for Korean native speaker at level A2
 - c) exercise for Polish native speaker at level B2
 - d) exercise for Korean native speaker at level B2

Analysis

- comparison of two types of exercises
- 2 target levels (A2 and B2)
- 2 target groups with regard to language (Poles and Koreans)
- 4 exercise variants:
 - a) exercise for Polish native speaker at level A2
 - b) exercise for Korean native speaker at level A2
 - c) exercise for Polish native speaker at level B2
 - d) exercise for Korean native speaker at level B2

Analysis

Prompt: *Potřebuji typ cvičení „Výběr správné kolokace“ pro polského/korejského rodilého mluvčího na úrovni B2.*

I need a type of exercise "Choosing correct collocation" for a Polish/Korean native speaker at level B2.

Polish learners

 Plán cvičení: 1. Najdeme časté kolokace vhodné pro úroveň B2 (slova, která se přirozeně pojí). 2. Vytvoříme cvičení, kde studenti vyberou správnou kolokaci. 3. Zaměříme se na typické chyby polských mluvčích, například: • Nesprávné slovesné kolokace (např. <i>vzít rozhodnutí</i> místo <i>učinit rozhodnutí</i>). • Falešní přátelé (např. <i>realizovat projekt</i> vs. <i>provést projekt</i>). • Nesprávné předložky (<i>záviset od</i> místo <i>záviset na</i>). ✦ Nejprve najdu časté kolokace pro slovo „rozhodnutí“, protože polští studenti často používají nesprávné sloveso. 
--

Korean learners

 Plán cvičení: 1. Najdeme časté kolokace vhodné pro úroveň B2 (slova, která se přirozeně pojí). 2. Vytvoříme cvičení, kde studenti vyberou správnou kolokaci. 3. Zaměříme se na typické chyby korejských mluvčích, například: • Nesprávné slovesné kolokace (např. <i>vzít rozhodnutí</i> místo <i>učinit rozhodnutí</i>). • Chybné předložky (<i>záviset něčem</i> místo <i>záviset na něčem</i>). • Nesprávná slovesa v ustálených vazbách (<i>mít otázku</i> místo <i>položit otázku</i>). ✦ Nejprve najdu časté kolokace pro slovo „vliv“, protože korejští studenti mají potíže se správným slovesem (např. <i>mít vliv</i> místo <i>ovlivnit</i>). 

Analysis

Working with corpus data

Using the SYN2020 Corpus with Collocation function and LogDice for lemma search

Polish learners

 Časté kolokace pro „rozhodnutí“ na základě korpusu:

- padnout rozhodnutí (např. Rozhodnutí padlo včera.)
- konečné rozhodnutí
- zásadní rozhodnutí
- správné rozhodnutí

 Polští studenti často nesprávně překládají „vzít rozhodnutí“ místo „učinit rozhodnutí“.

Korean learners

 Časté kolokace pro „vliv“ na základě korpusu:

- negativní vliv
- pozitivní vliv
- pod vlivem (např. pod vlivem alkoholu)
- příznivý vliv
- zásadní vliv

 Korejští studenti často nesprávně překládají „mít vliv“ místo „ovlivnit“.

Analysis

Polish learners

Ukázkové cvičení:

♦ Vyberte správnou možnost:

1. Vláda dnes _____ důležité rozhodnutí.

a) vzala

b) učinila 

2. Soud včera _____ konečné rozhodnutí.

a) vydal 

b) udělal

3. Po dlouhém zvažování nakonec _____ zásadní rozhodnutí.

a) padlo 

b) spadlo

 Správné odpovědi: 1b, 2a, 3a

Korean learners

Ukázkové cvičení:

♦ Vyberte správnou možnost:

1. Tento lék má _____ na lidské tělo.

a) příznivý vliv 

b) dobrý vliv

2. Alkohol _____ lidské rozhodování.

a) ovlivňuje 

b) má

3. Špatné počasí mělo _____ na dopravu.

a) pod vliv 

b) vliv

 Správné odpovědi: 1a, 2a, 3b

Conclusion

Corpus and generative AI

Evaluation of automatically generated corpus exercises revealed several lacks:

- inconsistent use of corpus data,
- drawing conclusions about common problems of foreigners based on sources other than corpus data,
- providing incorrect examples in sample exercises.

Korpus vs. generativní AI

GenAI

- user-friendliness
- rapidity
- generative AI surpasses corpus tools used in teaching in many technical solutions

Corpus

- replicability and multimodality of corpus results
- secure data handling
- credibility of data
- active student involvement in inductive learning

To not be overshadowed by GenAI but cope with it!

Yet, the field of corpus linguistics is at a crossroads. Despite our best efforts, our field risks being overshadowed by GenAI researchers who are essentially just doing what we as corpus linguists already do, but in a way that has finally captured the imagination of the public.

Thank you for your attention!

Adrian Jan Zasina

adrian.zasina@ff.cuni.cz



INSTITUTE OF CZECH
AND DEAF STUDIES
Faculty of Arts
Charles University

References 1

- Benko, V., Butašová, A., Lalinská, M., Pařová, M., Puchovská, Z., Segretain, A., & Zeleňáková, M. (2019). *Webové korpusy Aranea. Učebnica pre učiteľov cudzích jazykov, prekladateľov, tlmočníkov, filológov a študentov filologických odborov*. Univerzita Komenského v Bratislave.
- Boulton, A. (2017). Research timeline: Corpora in language teaching and learning. *Language Teaching*, 50(4), 483–506.
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), 1–4. <https://doi.org/10.1016/j.acorp.2023.100066>
- 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.
- Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*, 4, 1–16.
- Konečná, H., & Zasina, A. J. (2014). Studium českého jazyka a internet. In E. Rusinová (Ed.), *Přednášky a besedy ze XLVII. běhu LŠSS* (s. 104–112). Masarykova univerzita.
- Le Foll, E. (Ed.). (2021). *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom. A step-by-step guide for (trainee) teachers using online resources*. Open Educational Resource. <https://pressbooks.pub/elenlefol/>
- Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics*, 3(3), 100065. <https://doi.org/10.1016/j.acorp.2023.100065>
- Lin, P. (in press). Generative AI Chatbots and Data-Driven Learning (DDL). In L. McCallum & D. Tafazoli (Ed.), *Encyclopedia of Computer-Assisted Language Learning*. Springer.
- Lukešová, L. (Ed.) (2017). *Pro školy—Repozitář korpusových cvičení* [Software]. FF UK. <https://korpus.cz/proskoly>

References 2

- Maturová, M. (2016). Využití nástroje SyD ve výuce češtiny. *Český jazyk a literatura*, 67(4), 183–187.
- Milička, J. (2024). Corpus Linguist (model for Chat-GPT). [Software]. <https://chatgpt.com/>
- OpenAI. (2025). *Chat-GPT* (Verze January 17, 2025) [Software]. <https://chatgpt.com/>
- Osolsobě, K. (2010). Jak se učit česky s korpusem. In E. Rusinová (Ed.), *Přednášky a besedy z XLIII. běhu LŠSS* (s. 112–119). Masarykova univerzita.
- Šindelářová, J., & Škodová, S. (2013). *Práce s korpusy ve výuce žáků-cizinců*. Metodický portál rvp.cz. <https://clanky.rvp.cz/clanek/c/Z/17481/prace-s-korpusy-ve-vyuce-zaku-cizincu.html>
- Šormová, K. (2017). Akviziční korpusy češtiny a jejich využití ve výuce. *Jazyk – literatura – komunikace*, 1, 76–84.
- Thomas, J. E. (2016). *Discovering English with sketch engine*. Versatile.
- Tribble, C., & Jones, G. (1997). *Concordances in the classroom: A resource guide for teachers*. Athelstan.
- Vališová, P. (2011). Výukové materiály založené na korpusu. In F. Čermák (Ed.), *Korpusová lingvistika Praha 2011. 2 Výzkum a výstavba korpusu* (s. 313–323). Nakladatelství Lidové noviny.
- 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 (Ed.), *Čeština jako cizí jazyk VIII* (s. 128–141). Univerzita Karlova, Filozofická fakulta.

References 3

- Vališová, P., & Osolsobě, K. (2012). Using data-driven method in teaching Czech as a foreign language. In J. Thomas & A. Boulton (Ed.), *Input, Process and Product: Developments in Teaching and Language Corpora* (s. 183–194). Masaryk University Press.
- Vyatkina, N. (Ed.). (2020). *Incorporating corpora: Using corpora to teach German to English-speaking learners [Online instructional materials]*. University of Kansas Open Language Resource Center. <https://opentext.ku.edu/corpora>
- Zasina, A. J. (2022a). Corpus approach to teaching Czech as a foreign language in university courses. *Bohemistika*, XXII(3), 435–462. <https://doi.org/10.14746/bo.2022.3.7>
- Zasina, A. J. (2022b). Designing a Corpus Workbook for Students of Czech as a Foreign Language. *Studie z aplikované lingvistiky - Studies in Applied Linguistics*, 13(2), 125–132.
- Zasina, A. J. (2023). *Korpusová cvičebnice pro studenty češtiny jako cizího jazyka*. Karolinum.
- Zasina, (forthcoming). Typologie korpusových cvičení a jejich automatické generování pomocí generativní AI. In *Studie z Korpusové lingvistiky*. Nakladatelství Lidové novin.