CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

# Extracting Multi-word Expressions for the Czech Academic Phrase List

Dominika Kováříková, Oleg Kovářík & Lucie Lukešová

Charles University

# OUTLINE

1. Briefly about academic language
2. Our data and methodology in 4 steps
3. Analyzing the core list of academic phrases
4. Possible classification of the core
   - POS-grams, semantic groups, formal groups...
5. Introducing a web-based search tool for APL

# Introduction

- Inspired by English academic language resources
- Our research is corpus-driven, based on n-grams
- Academic texts generally include:

**TERMINOLOGY**

- higher frequency in academic texts
- used predominantly in a **small number of disciplines**
- e.g. *sulphuric acid*

**ACADEMIC VOCABULARY**

- higher frequency in academic texts
- distributed evenly and shared among **many disciplines**
- e.g. *carry out experiment*

# Introduction

- Inspired by English academic language resources
- Our research is corpus-driven, based on n-grams
- Academic texts generally include:

**TERMINOLOGY**

- higher frequency in academic texts
- used predominantly in a **small number of disciplines**
- e.g. *sulphuric acid*

**ACADEMIC VOCABULARY**

- higher frequency in academic texts
- distributed evenly and shared among **many disciplines**
- e.g. *carry out experiment*

- Main criterion: distribution in academic disciplines (cf. Kováříková 2017)

# ACADEMIC LANGUAGE

# What is academic language?

- consise definition by Nagy & Townsend, 2012

"*the specialized language, both oral and written, of academic settings that facilitates communication and thinking about disciplinary content*"

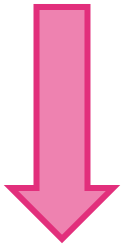
- used in academic journals, scientific papers, monographs as well as university textbooks and students' theses
- there are no native speakers of academic language

# Academic word/phrase lists

- Especially popular in English
  - *The Academic Word List* (Coxhead 2000)
  - *Academic Keyword List* (Granger & Paquot 2010)
  - *A New Academic Vocabulary List* (Gardner & Davies 2013)
  - *Academic Phrasebank* (University of Manchester 2015)
- Often organised according to the main sections of a research paper or dissertation
- Useful resources for
  - University students (both L1 and L2) as well as teachers
  - Young researchers
  - Also translators

# Academic Czech

- No such resources available for Czech



- Our main objectives:
  - To provide a corpus-driven list of academic phrases
  - To contribute to a knowledge base for teaching academic skills at universities

# DATA AND METHODOLOGY

# Our source of data

- Our specialized corpus of academic texts - 13 mil. tokens
- subcorpus of SYN2015
  - 13 mil. tokens
    - Monographs
    - Scientific papers
    - University textbooks
    - Reference books
- Our reference corpus:
  - 80 mil. tokens
  - Fiction, newspapers and magazines

# Disciplines in our SCI corpus

| | | |
|---|---|---|
| BIO biology | THE theatre | ANT anthropology |
| MED medicine | HIS history | PHY physics |
| PHI philosophy | MUS music | CHE chemistry |
| EDU education | ART arts | TEC technology |
| LAN philology | GEO geography | ECO economy |
| POL politics | REC sports | ITD interdisciplinary |
| LAW law | INF library science | AGR agriculture |
| SOC sociology | PSY psychology | ICT IT |

# Methodology

- Step 1: Automatic extraction of 1-6-grams
- Step 2: Choosing criteria for selection of valid phrases
- Step 3: Setting thresholds
- Step 4: Manual sorting and classification of the core

# Methodology

Step 1: **Automatic extraction of 1-6-grams**

- Several issues to resolve:
  - lemmas v. word forms?
  - punctuation excluded, but conjunctions included
  - word order variability

# Methodology

Step 2: Choosing **criteria for selection** of valid phrases

- **Ratio of relative frequencies in the corpus of academic texts and the reference corpus**
- **Distribution in the disciplines**
- Dispersion in the disciplines (another type of distribution)
- Association measures for 2-grams (MI) and 3-grams (PMI)

# Methodology

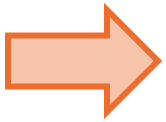
## Step 3: Setting **thresholds**

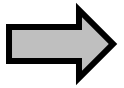
- Question: Where to cut it?
- no perfect answer > users have different needs
- focus on the most universal expressions, i.e. used in all disciplines

- **Our thresholds**:

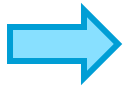
Distribution = 24

Ratio of FQ > 2



1-grams: 989  future plan

2-grams: 793

3-grams: 99  basis for the **CORE**

# Methodology

Step 4: Manual **sorting and classification** of the core

- Different needs and conceptions > difficult to choose the final set
- Inter-rater agreement low – only 54 % items unequivocally classified as academic phrases, the rest subject to negotiation

- **Final selection** includes:

2-grams: 793

3-grams: 99

2-grams: 346

3-grams: 60

# ANALYZING THE CORE LIST

# POS characteristics: bigrams

Most common POS-grams:

preposition – noun (RN): 45

*za účelem, s výjimkou, na základě, na přelomu, pod pojmem, v rámci...*

verb – preposition (VR): 41

*považovat za, vyplývat z, záviset na, přispívat k, poukazovat na...*

noun – preposition (NR): 37

*rozdíl mezi, informace o, základ pro, práce s, přístup k, pokus o...*

verb – verb (VV): 37

*lze říci, lze nalézt, být označován, být využit, moci nalézt...*

pronoun – verb (PV): 31

*se nazývat, se vztahovat, se nacházet, se vyskytovat, který tvoří*

# POS characteristics: trigrams

Most common POS-grams:

preposition – noun – preposition (RNR): 10

*v souladu s, ve srovnání s, ve vztahu k, s ohledem na, na rozdíl od...*

preposition – pronoun – noun (RPN): 9

*do jaké míry, z tohoto důvodu, v této souvislosti, v této oblasti...*

verb – verb – preposition (VVR): 6

*moci dojít k, být považován za, být založen na, moci vést k...*

pronoun – verb – preposition (PVR): 4

*se jednat o, se setkávat se, který vychází z, který vede k*

preposition – adjective – noun (RAN): 3

*do jisté míry, do určité míry, ve velké míře*

# Formal groups

= groups manually classified based on their form

- Reflexive: verbs with a mandatory reflexive particle
  - *lišit se, nacházet se, vztahovat se…*
- Syntax: patterns pointing to a specific syntactic/syntagmatic structure
  - Clause endings/beginnings: *platit že*, *otázka zda*, *být zřejmé že...*
  - Passive voice: *být uveden, být popsán, být rozdělen*
- Phrase: a broad category of set phrases, multiword prepositions etc.
  - *v důsledku, tato problematika, definovat jako, výše uvedený, pod pojmem, na principu, přispívat k, vyplývat z, být důsledkem...*
  - *do jisté míry, bez ohledu na, jedním ze základních, jedním z nejdůležitějších, v této souvislosti, v závislosti na...*
- Incomplete phrase: a part of a longer n-gram > excluded

# Semantic groups

= groups manually classified based on their meaning and function

- Non-specific terminology: terms related to research and science in general
  - *obecná teorie, přesně definovat, vědecký výzkum, teoretické východisko…*
- Formulas: typical academic phrases, incl. multiword prepositions ~ formal Phrase
  - *v rámci, a to i, v souvislosti s, do jisté míry, může vést k, z tohoto hlediska…*
- Mathematical and statistical expressions, used in all disciplines:
  - *statistické údaje, kvantitativní metoda, naměřené hodnoty, průměrná hodnota…*
- Text organizing phrases:
  - *jak již bylo zmíněno, v následující kapitole, výše uvedený…*
- References to literature:
  - *v odborné literatuře, seznam literatury, často citovaný…*
- Time and space references:
  - *v první polovině 20. století, v Čechách a na Moravě…*

# Looking beyond the core

- Useful longer n-grams that did not make it to the core (D < 24)

4-grams: *v úvodu této kapitoly, jak je znázorněno na, cílem této kapitoly je, považovat za jeden z, být v rozporu s, vzhledem k tomu že…*

5-grams: *ale vzhledem k tomu že, dalo by se říct že, bez ohledu na to zda, a do jisté míry i, v tomto případě jde o...*

6-grams: *je třeba vzít v úvahu že, v tomto případě se jedná o, na první pohled se může zdát (že)...*

- Single words or 1-grams used in academic texts

A list of the 20 most frequent verbs (D = 24):

*pojednat, viz, charakterizovat, definovat, odvodit, rozlišovat, vymezit, odvozovat, vyskytovat, transformovat, formulovat, interpretovat, uplatňovat, uskutečňovat, redukovat, přiřadit, označovat, odlišovat, rozlišit, nabývat*

# AKALEX: beta version

https://jupyter.korpus.cz/shiny/kovarikova/akalex/

the CORE  all extracted n-grams with all the values

AKALEX

*Lexikon akademické češtiny*

**SYN2015:**

veškeré údaje jsou založeny na korpusu SYN2015

**Akademické texty:**

texty označené v korpusu SYN2015 jako textový typ *SCI: odborná literatura* (velikost 13 milionů textových pozic)

**Referenční korpus:**

subkorpus SYN2015 obsahující beletristické a žurnalistické texty (velikost 81 milionů textových pozic)

**Akce pro tabulku:**

*stahovaná data odpovídají aktuálnímu zobrazení dat včetně filtrování*

Stáhnout Aca2

**Kontakt:**

dominka.kovarikova@korpus.cz
lucie.lukesova@ff.cuni.cz
oleg.kovarik@gmail.com

Download of the current list, incl. filters

Akademické fráze | Kandidáti - lemmata | Kandidáti - slovní tvary

2-gramy | 3-gramy

Zobraz záznamů 10

Hledat:

| | Kategorie | LEMMA1 | LEMMA2 | Tvar | POS1 | POS2 | Poměr frekvencí | Distribuce | Disperze | MI |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | All | | | All | All | All | |
| 1 | fráze | následující | kapitola | následující kapitole | A | N | 369.6 | 24 | 0.64 | 7.27 |
| 2 | fráze | tento | kapitola | této kapitole | P | N | 180.54 | 24 | 1.07 | 5.01 |
| 3 | fráze | v | kapitola | v kapitole | R | N | 73.74 | 24 | 0.73 | 2.25 |
| 4 | fráze | definovat | jako | definována jako | V | J | 30.29 | 24 | 0.7 | 5.73 |
| 5 | reflexivum | se | uplatňovat | se uplatňuje | P | V | 28.45 | 24 | 0.69 | 4.28 |
| 6 | fráze | tento | vztah | těchto vztahů | P | N | 24.45 | 24 | 1.1 | 2.17 |
| 7 | fráze | být | charakteristický | je charakteristická | V | A | 20.78 | 24 | 0.63 | 3.23 |
| 8 | fráze | označovat | jako | označuje jako | V | J | 19.87 | 24 | 0.9 | 3.07 |
| 9 | syntax | se | označovat | se označuje | P | V | 18.82 | 24 | 1.72 | 2.77 |
| 10 | reflexivum | se | vyskytovat | se vyskytují | P | V | 17.59 | 24 | 1.42 | 4.75 |

Zobrazuji 1 až 10 z celkem 346 záznamů

Předchozí 1 2 3 4 5 … 35 Další

Most frequent word forms

# Noun – preposition (NR)

AKALEX

*Lexikon akademické češtiny*

**SYN2015:**

veškeré údaje jsou založeny na korpusu SYN2015

**Akademické texty:**

texty označené v korpusu SYN2015 jako textový typ *SCI: odborná literatura* (velikost 13 milionů textových pozic)

**Referenční korpus:**

subkorpus SYN2015 obsahující beletristické a žurnalistické texty (velikost 81 milionů textových pozic)

**Akce pro tabulku:**

*stahovaná data odpovídají aktuálnímu zobrazení dat včetně filtrování*

Stáhnout Aca2

**Kontakt:**

dominka.kovarikova@korpus.cz
lucie.lukesova@ff.cuni.cz
oleg.kovarik@gmail.com

Akademické fráze | Kandidáti - lemmata | Kandidáti - slovní tvary

2-gramy | 3-gramy

Zobraz záznamů 10

Hledat:

| | Kategorie | LEMMA1 | LEMMA2 | Tvar | POS1 | POS2 | Poměr frekvencí | Distribuce | Disperze | MI |
|---|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | All | ["N"] | ["R"] | All | All | All | A |
| 15 | fráze | vazba | mezi | vazby mezi | N | R | 15.22 | 24 | 1.37 | 5.51 |
| 24 | fráze | význam | pro | význam pro | N | R | 13.34 | 24 | 0.48 | 3.67 |
| 26 | fráze | vztah | mezi | vztah mezi | N | R | 13.08 | 24 | 0.48 | 6.44 |
| 27 | fráze | souvislost | mezi | souvislost mezi | N | R | 12.95 | 24 | 0.65 | 2.18 |
| 31 | fráze | orientace | v | orientaci v | N | R | 11.89 | 24 | 0.86 | 2.52 |
| 46 | fráze | příspěvek | k | příspěvek k | N | R | 9.22 | 24 | 0.78 | 4.78 |
| 62 | fráze | zdroj | pro | zdroje pro | N | R | 7.61 | 24 | 1.34 | 2.37 |
| 87 | fráze | kritérium | pro | kritéria pro | N | R | 6.39 | 24 | 0.57 | 4.42 |
| 99 | fráze | základ | pro | základ pro | N | R | 6.07 | 24 | 0.55 | 3.07 |
| 119 | fráze | poměr | mezi | poměr mezi | N | R | 5.73 | 24 | 0.84 | 5.09 |

Zobrazuji 1 až 10 z celkem 21 záznamů (filtrováno z celkem 346 záznamů)

Předchozí 1 2 3 Další

# Browsing through the list of candidates

AKALEX

*Lexikon akademické češtiny*

**SYN2015:**

veškeré údaje jsou založeny na korpusu SYN2015

**Akademické texty:**

texty označené v korpusu SYN2015 jako textový typ *SCI: odborná literatura* (velikost 13 milionů textových pozic)

**Referenční korpus:**

subkorpus SYN2015 obsahující beletristické a žurnalistické texty (velikost 81 milionů textových pozic)

**Akce pro tabulku:**

*stahovaná data odpovídají aktuálnímu zobrazení dat včetně filtrování*

Stáhnout Lemma2

**Kontakt:**

dominka.kovarikova@korpus.cz
lucie.lukesova@ff.cuni.cz
oleg.kovarik@gmail.com

Akademické fráze | Kandidáti - lemmata | Kandidáti - slovní tvary

1-gramy | 2-gramy | 3-gramy | 4-gramy | 5-gramy | 6-gramy

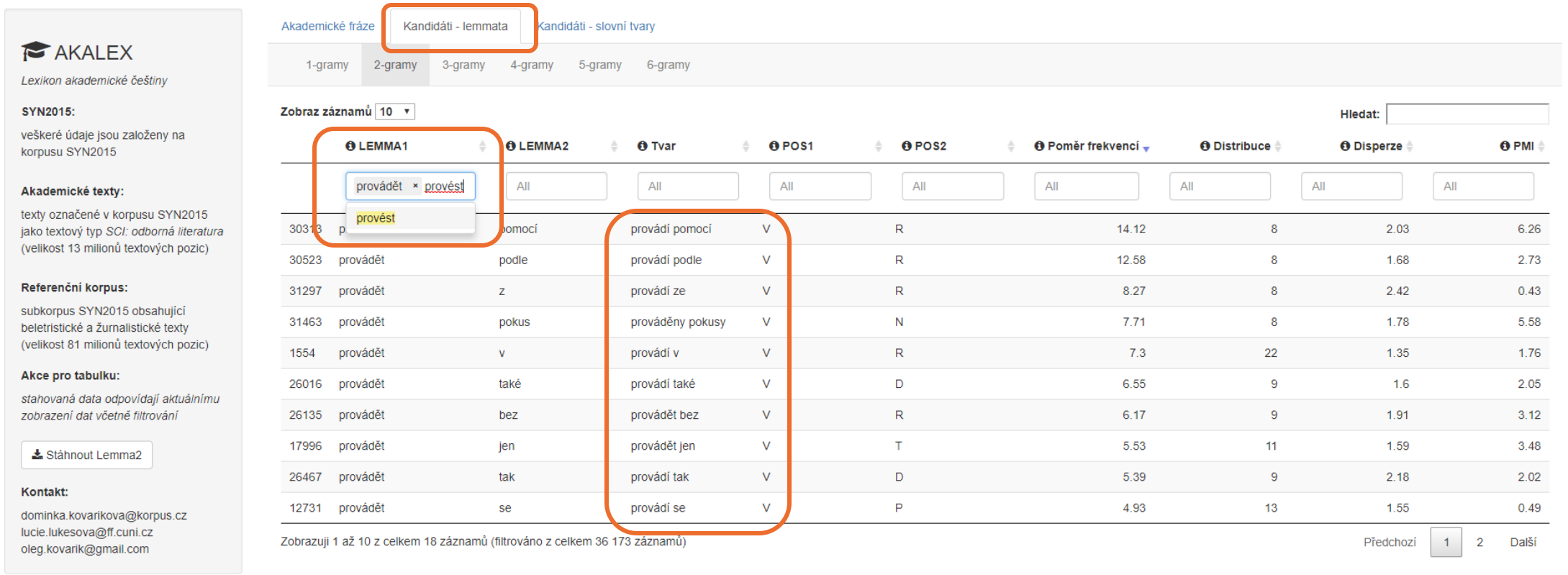
Zobraz záznamů 10 ▾ Hledat:

| | LEMMA1 | LEMMA2 | Tvar | POS1 | POS2 | Poměr frekvencí | Distribuce | Disperze | PMI |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | All | ["A"] | ["R"] | All | All | All | All |
| 29156 | znázorněný | na | znázorněný na | A | R | 276 | 8 | 2.42 | 5.71 |
| 16513 | kolmý | k | kolmé k | A | R | 214.63 | 11 | 3.53 | 5.81 |
| 23986 | analogický | k | analogické k | A | R | 116 | 9 | 2.19 | 3.38 |
| 23992 | kolmý | na | kolmé na | A | R | 107.15 | 9 | 3.16 | 4.5 |
| 29194 | lokalizovaný | v | lokalizované v | A | R | 101 | 8 | 1.75 | 2.88 |
| 19841 | zjištěný | v | zjištěná v | A | R | 87.67 | 10 | 1.54 | 1.37 |
| 24016 | vztažený | na | vztažené na | A | R | 84.8 | 9 | 2.47 | 4.81 |
| 19898 | uplatňovaný | v | uplatňovaných v | A | R | 57.33 | 10 | 1.62 | 3.27 |
| 24115 | pozorovaný | v | pozorovaná v | A | R | 54 | 9 | 1.93 | 0.74 |
| 29336 | odlišitelný | od | odlišitelné od | A | R | 54 | 8 | 2.03 | 7.5 |

Zobrazuji 1 až 10 z celkem 337 záznamů (filtrováno z celkem 36 173 záznamů)

Předchozí 1 2 3 4 5 … 34 Další

# Choosing a specific lemma



# Future plans

- Sort out more candidates and process further material (e.g. 1-grams, D>=16, longer n-grams)
- Improve the classification
- Include additional criteria (e.g. entropy)
- Revise incomplete n-grams and add them to the list
- Provide a guide on how to use the lists in academic writing
- Promote the tool among teachers, students and translators...

# Thank you for your attention!

dominika.kovarikova@ff.cuni.cz
oleg.kovarik@gmail.com
lucie.chlumska@ff.cuni.cz

CZECH NATIONAL CORPUS