CZECH NATIONAL
CORPUS

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS

FEAST & FAMINE

Sharing data through specialized corpus-based tools: the case of GramatiKat

Dominika Kováříková

Slovko, October 2021

Main points

- how to share data (and get something in return)
- quantitative analysis of GK is worthy of our attention
- experience with creating pre-processed and pre-analyzed data

representative corpus of contemp. written Czech, 100 million words
SYN2005, SYN2015, SYN2020

Reasons to share data (NŘ, March 2021)

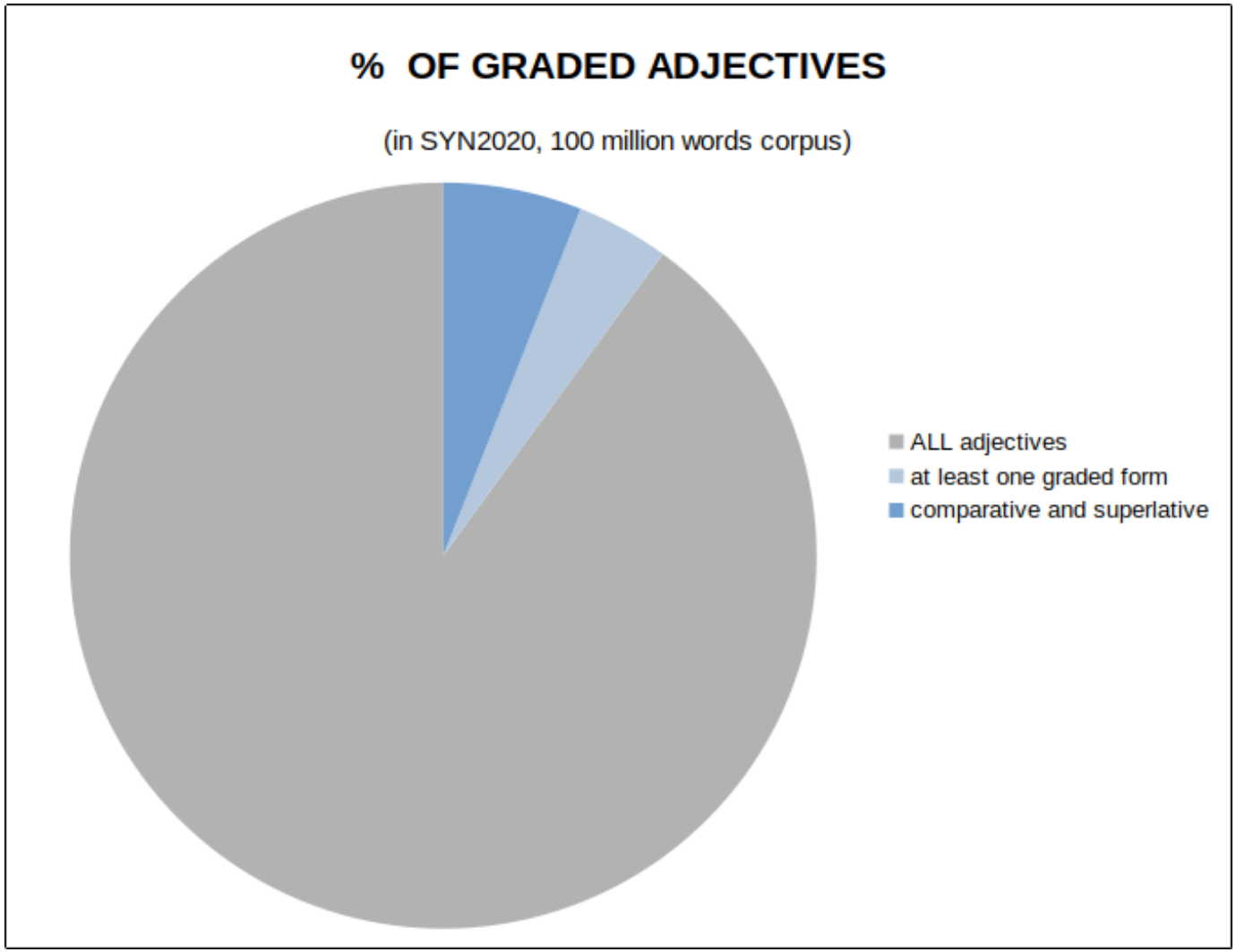
1. falsifiability/verifiability
2. incentive to be thorough in your research
3. replicability and possibility to further develop the original idea
4. synergy and opening new research possibilities
5. student training
6. social responsibility (publicly funded resarch)
7. FAIR data and possibility to re-interpret results
8. publishing the data as an regular research output
9. meta-analysis

(Chromý & Cvrček, 2021)

Reasons to share data (NŘ, March 2021)

1. falsifiability/verifiability
2. incentive to be thorough in your research
3. replicability and possibility to further develop the original idea
4. synergy and opening new research possibilities
5. student training
6. social responsibility (publicly funded resarch)
7. FAIR data and possibility to re-interpret results
8. publishing the data as an regular research output
9. meta-analysis

(Chromý & Cvrček, 2021)

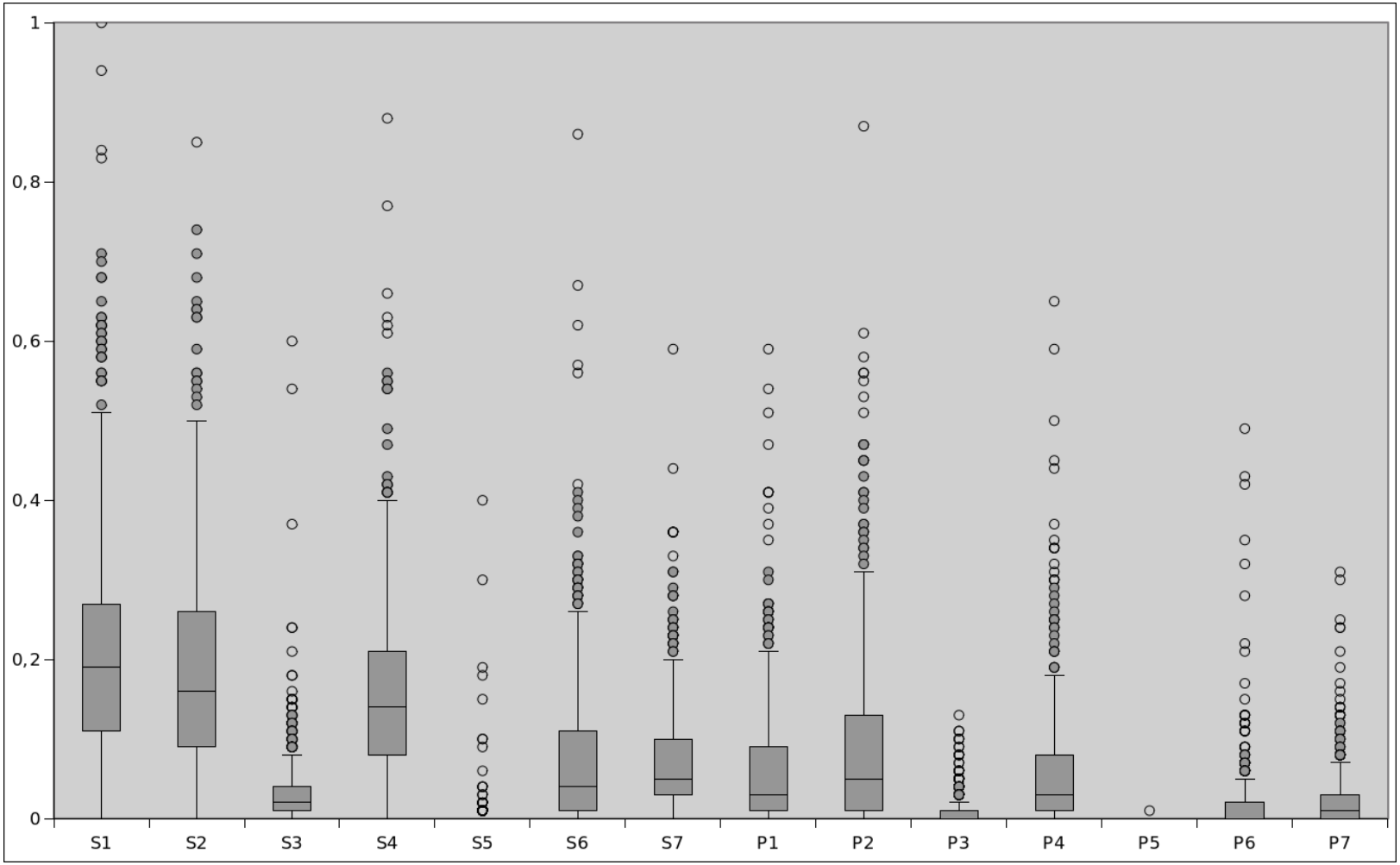

% OF GRADED ADJECTIVES
(in SYN2020, 100 million words corpus)
ALL adjectives
at least one graded form
comparative and superlative

GramatiKat: Pre-processed data

LEMMA	ROD	S1	S2	S3	S4	S5	S6	S7	P1	P2
All	All	All	All	All	All	0.20 – 0.98	All	All	All	All
bača	M	0.48	0.05	0.02	0.07			0.05	0.04	0.03
beruška	F	0.23	0.07	0.01	0.09			0.05	0.09	0.06
brach	M	0.15	0	0.22	0.05	0.33	0.01	0	0.14	0.05
čubka	F	0.46	0.02	0.03	0.09	0.28	0.01	0.05	0.02	0.02
děvenka	F	0.08	0.02	0.01	0.01	0.73	0	0.01	0.03	0.02
drahoušek	M	0.08	0.01	0.42	0.01	0.42	0.02	0	0.03	0
hajzl	M	0.26	0.04	0.03	0.12	0.34	0.01	0.01	0.13	0.04
inspektorka	F	0.47	0.07	0.05	0.07	0.24	0	0.05	0.02	0.02
kámoš	M	0.22	0.04	0.04	0.05	0.3	0	0.04	0.13	0.05
lhářka	F	0.56	0.02	0	0.1	0.21	0	0.05	0.02	0.02
miláček	M	0.14	0.04	0.3	0.06	0.24	0.01	0.07	0.04	0.02
mylord	M	0.02	0.01	0	0	0.98	0	0	0	0
pán	M	0.17	0.05	0.03	0.04	0.49	0.01	0.03	0.09	0.03
píča	F	0.32	0.15	0.04	0.13	0.22	0.06	0.06	0.01	0
signor	M	0.35	0.09	0.02	0.07	0.46	0	0	0	0
sir	M	0.44	0.15	0.07	0.04	0.26	0.01	0.03	0	0

Standard behavior of a word class

Standard case distribution in Czech nouns (SYN2015!)



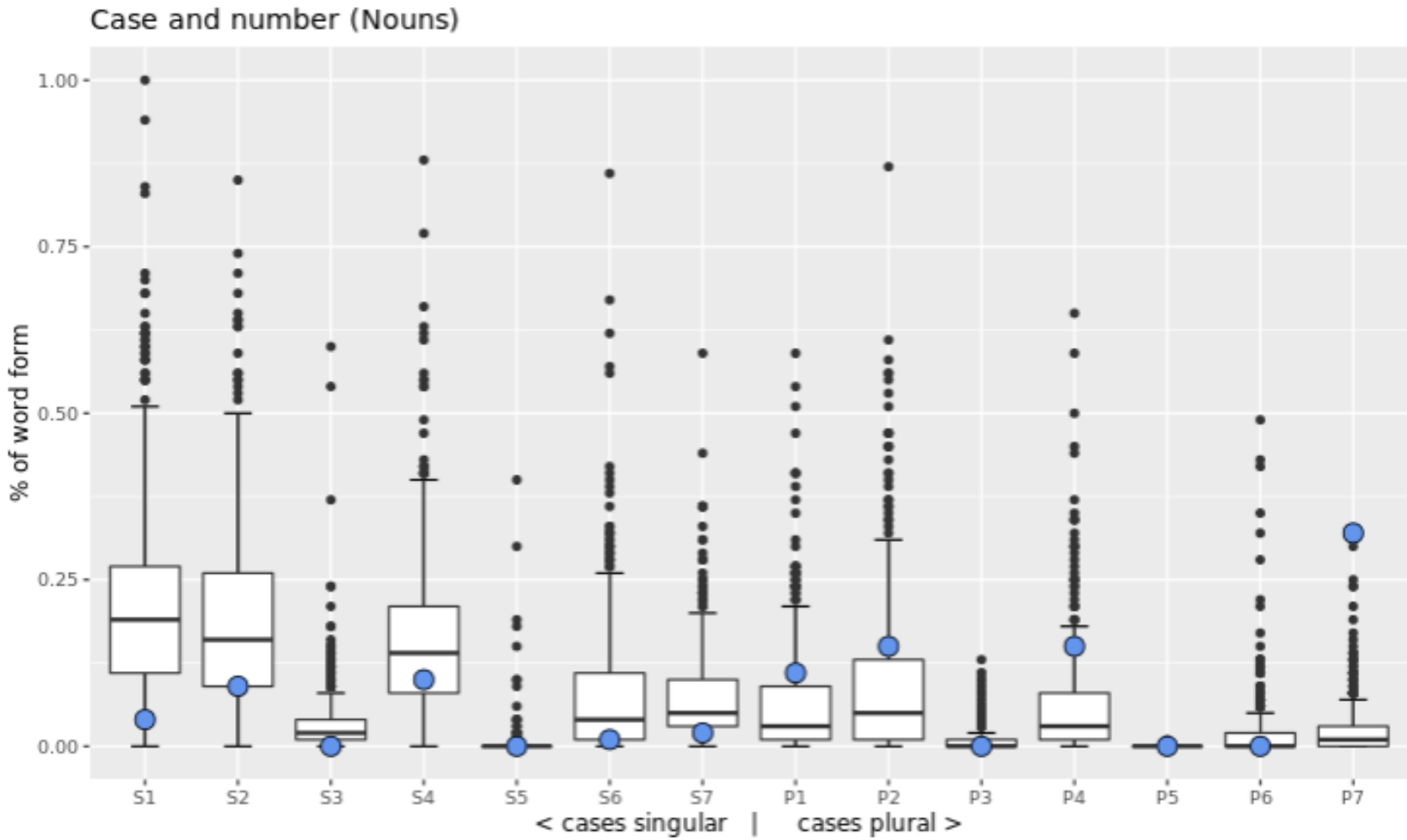
Grammatical profiles of lemmas

distribution of grammatical category values within individual lemmas

$\rightarrow$ e.g. distribution of case in nouns

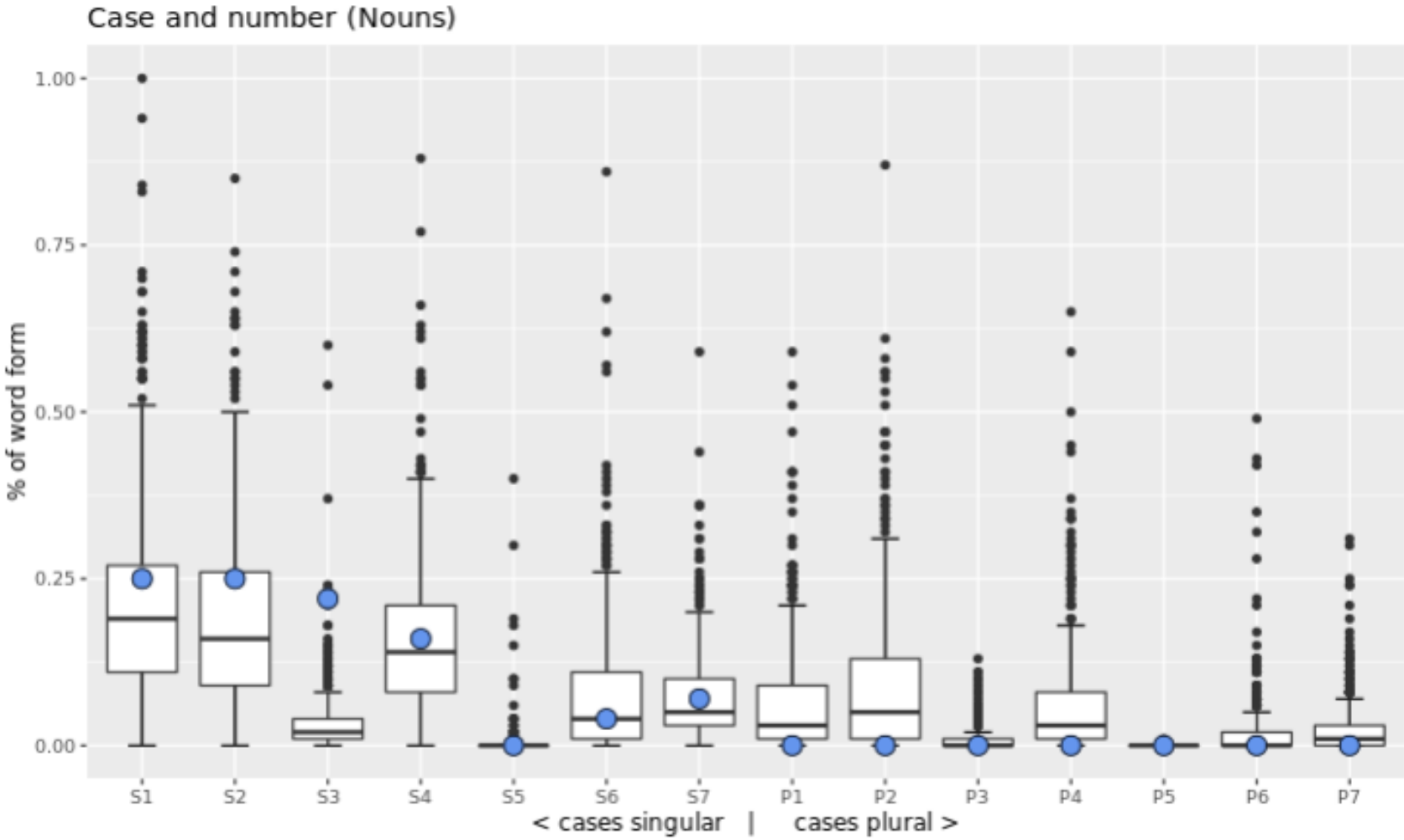
Grammatical profiles of lemmas

třpytka ‘glitter’



Grammatical profiles of lemmas

šílenství ‘madness’



Expect the unexpected

- anomaly is quite common (14 ways to go wrong)
- at least one anomalous paradigm cell in 40% of nouns (missing or excess)
- frequency dependent phenomenon!

Expect the unexpected

- anomaly is quite common (14 ways to go wrong)
- at least one anomalous paradigm cell in 40% of nouns (missing or excess)
- frequency dependent phenomenon!

$\Longrightarrow$ **blurry boundary between lexicon and grammar**

Lists of anomalous lemmas

emerging patterns
(semantic fields, grammatical phenomena...)

$\Longrightarrow$ Extremely high percentage of Locative plural:

v džínsách ('in jeans')
v montérkách ('in ovealls')
v kraťasech ('in shorts')
v trenýrkách ('in boxer shorts')
v šortkách ('in shorts')
ve spodkách ('in long johns')
ve slipech ('in underpants')
v teplákách ('in sweatpants')
v boxerkách ('in boxers')

APPLICATIONS

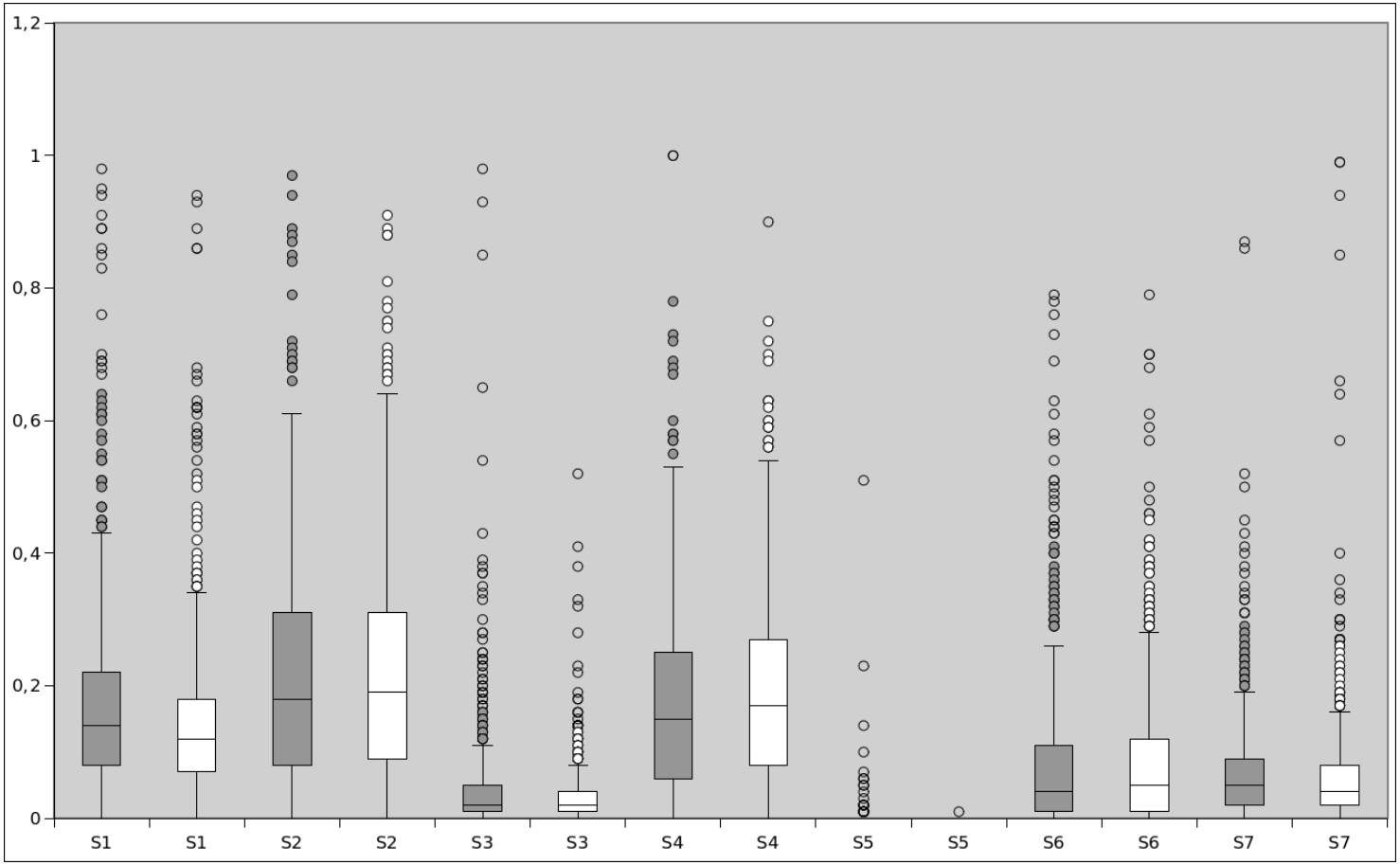
- morphology
- lexicography
- teaching on all levels, speech therapy, CzSL
- theoretical research of paradigm defectivity and language potentiality

Morphology

- in-depth analysis of grammatical categories
- synergy of grammatical and lexicological research
- comparison of languages

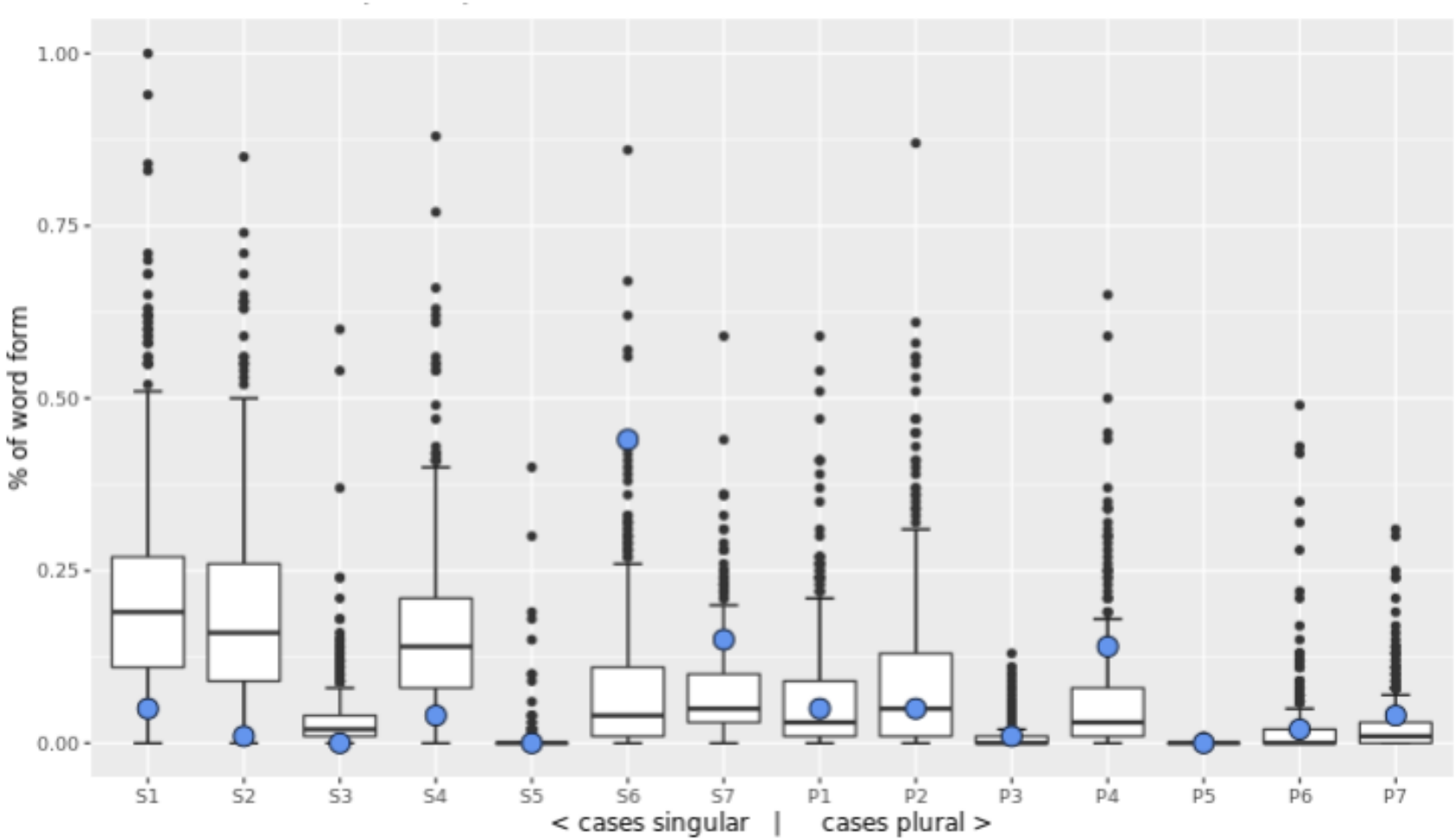
Czech vs Slovak sg

InterCorp parallel corpus, 50 mil word for Czech and for Slovak



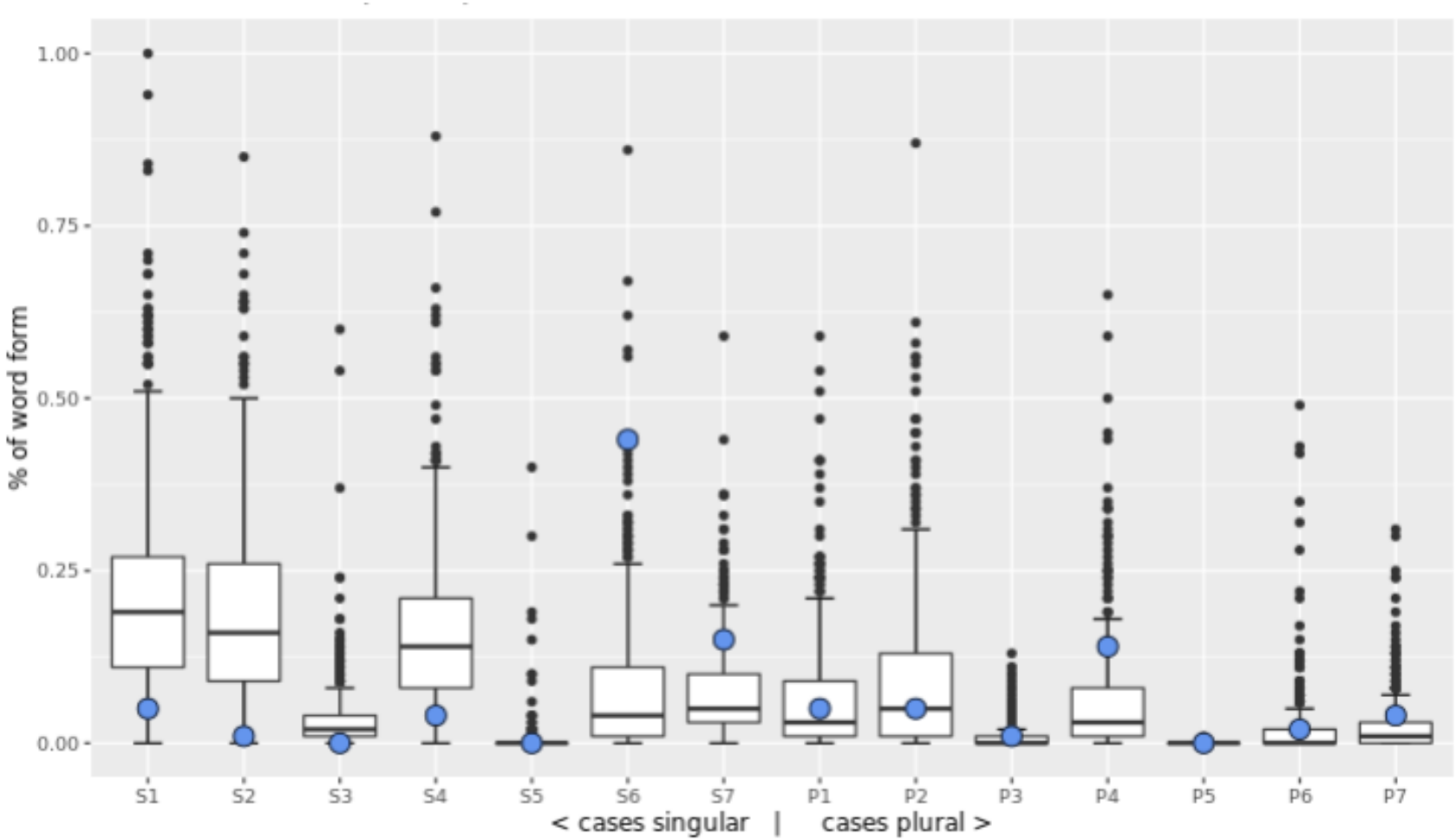
Lexicography (academic Czech)

důsledek (N) ‘consequence’



Lexicography (academic Czech)

v důsledku (preposition) ‘as a result of’



Defectivity and language potentiality

Project Feast and Famine:
Confrontation of overabundance and defectivity in language

Internetová jazyková příručka

- Hlavní stránka
- O příručce
- Nápověda
- Mobilní verze
- Návštěvnost
- English version

Související odkazy:

- Jazyková poradna
- ČSN 01 6910
- Zajímavé dotazy
- Databáze dotazů

Hledání konkrétního slova nebo tvaru slova.

Hledej

jih

dělení: jih[1]

rod: m. neživ.

	jednotné číslo	množné číslo
1. pád	jih	jihy[5]
2. pád	jihu[2]	jihů
3. pád	jihu	jihům
4. pád	jih	jihy
5. pád	jihu[3]	jihy
6. pád	jihu[4]	jizích[6]
7. pád	jihem	jihy

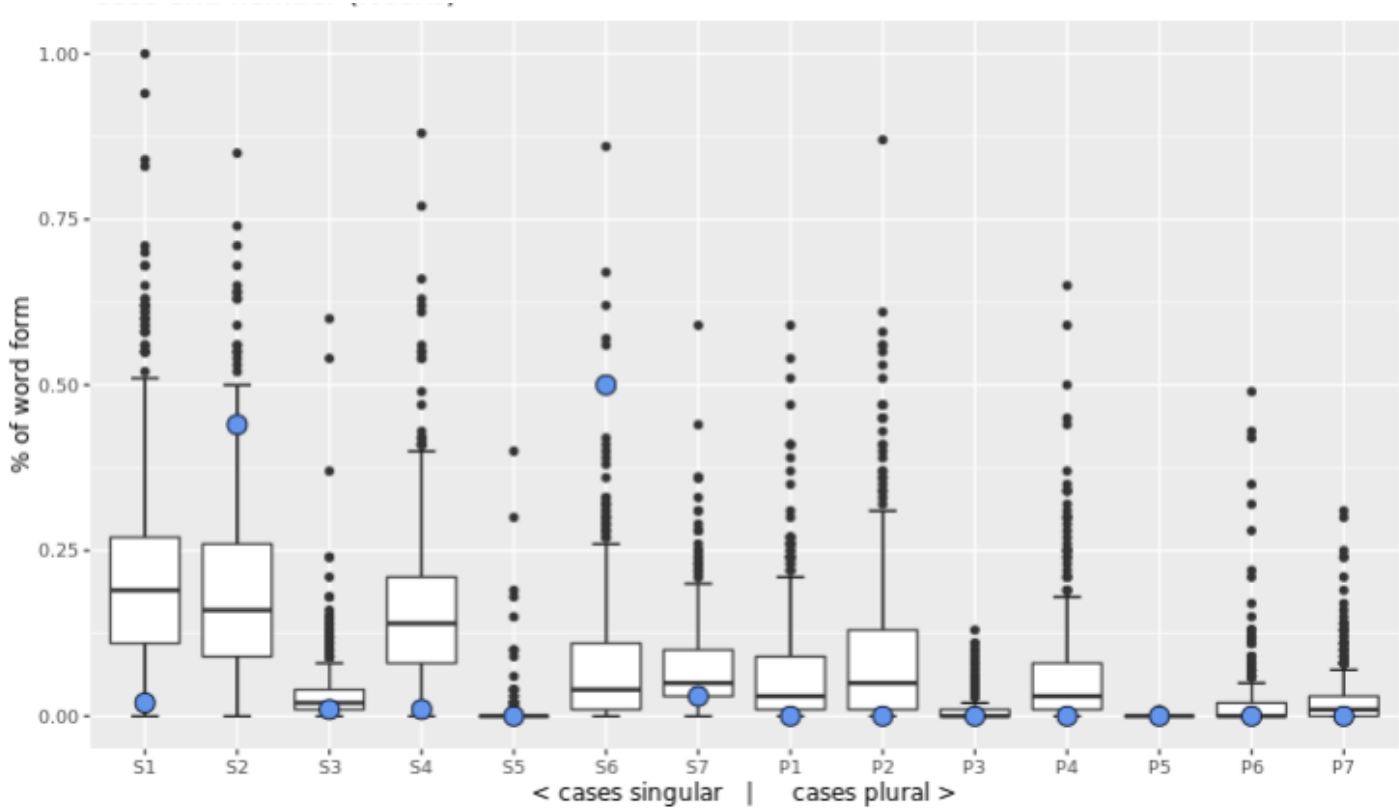
příklady: *na jihu Čech (Moravy, Evropy, USA);*
válka Severu proti Jihu[7]

Heslové slovo bylo nalezeno také v následujících slovnících: SSČ, SSJČ

Teaching

Janda and Tyers (2018) suggest that *"learning may be enhanced by focusing only on the word forms most likely to be encountered"* (p. 28).

cizina 'foreign countries, abroad'



Conclusion

- sharing data has its benefits: more ideas, more collaborations
- sharing ideas is easier, if we have too many (↑)
- quantitative research of GK is fascinating

References

Cvrček, V. et al. (2009). *Mluvnice současné češtiny I.: Jak se píše a jak se mluví.* Praha: Karolinum.

Čermák, F. et al. (2009). *Statistiky češtiny*. Prague: NLN.

Janda, L.A. and Tyers, F.M. (2018). Less is more: why all paradigms are defective, and why that is a good thing. *Corpus linguistics and linguistic theory*, 14(2). Accessible at https://doi.org/10.1515/cllt-2018-0031.

Křen, M. et al. (2015). *SYN2015: reprezentativní korpus psané češtiny.* Prague: ÚČNK FF UK. Accessible at http://www.korpus.cz.

Kováříková, D. and Kovářík, O. (2021). GramatiKat. Prague: ÚČNK FF UK. Praha 2021. Accessible at http://www.korpus.cz/gramatikat.

Kováříková, D. et al. (2019). Lexicographer's Lacunas or How to Deal with Missing Representative Dictionary Forms on the Example of Czech. *International Journal of Lexicography*, 33(1), pages 90–103. Accessible at https://doi.org/10.1093/ijl/ecz027.