CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

Lexicographer's Lacunas
or
How to Deal with Missing Dictionary Forms (On the Example of Czech)

Dominika Kováříková, Michal Škrabal, Václav Cvrček, Jiří Milička, Lucie Chlumská

OUTLINE:

- Introduction
- Lacunas and their types
- Identifying lacunas
- Recommendations for lexicographers
- Summary

Introduction

Lacunas in general

- words with incomplete paradigm

Lacunas in lexicography

- words without a dictionary form

This issue is rather neglected in literature: Wolski 1989, Schnorr 1991, Svensén 2009: 105–106; cf. in Czech lexicography: Čermák 1995, Filipec 1995, Kochová – Opavská 2016

Should the lexicographer reconstruct the *unattested* word form?

Dictionary forms

typically the *singular nominative* for nouns or the *infinitive* for verbs

×

- Latin, Greek or Bulgarian: 1st person singular present tense
- Hungarian or Macedonian: 3rd person singular present tense
- classical Arab dictionaries (listing roots instead of whole lexemes)

There will always be lexemes that lack the representative form = **lacunas**

Identifying lacunas and their types

Data

SYN2015 (100M) & SYN v6 (4G)
- 400 nouns
- 400 verbs
- 1000 adjectives

Classification of lacuna types

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

Types of lacunas

1. **systemic**: dictionary form cannot exist due to some reason (such as *vdaná* "married [woman]" in masculine) → calculate the expected frequency

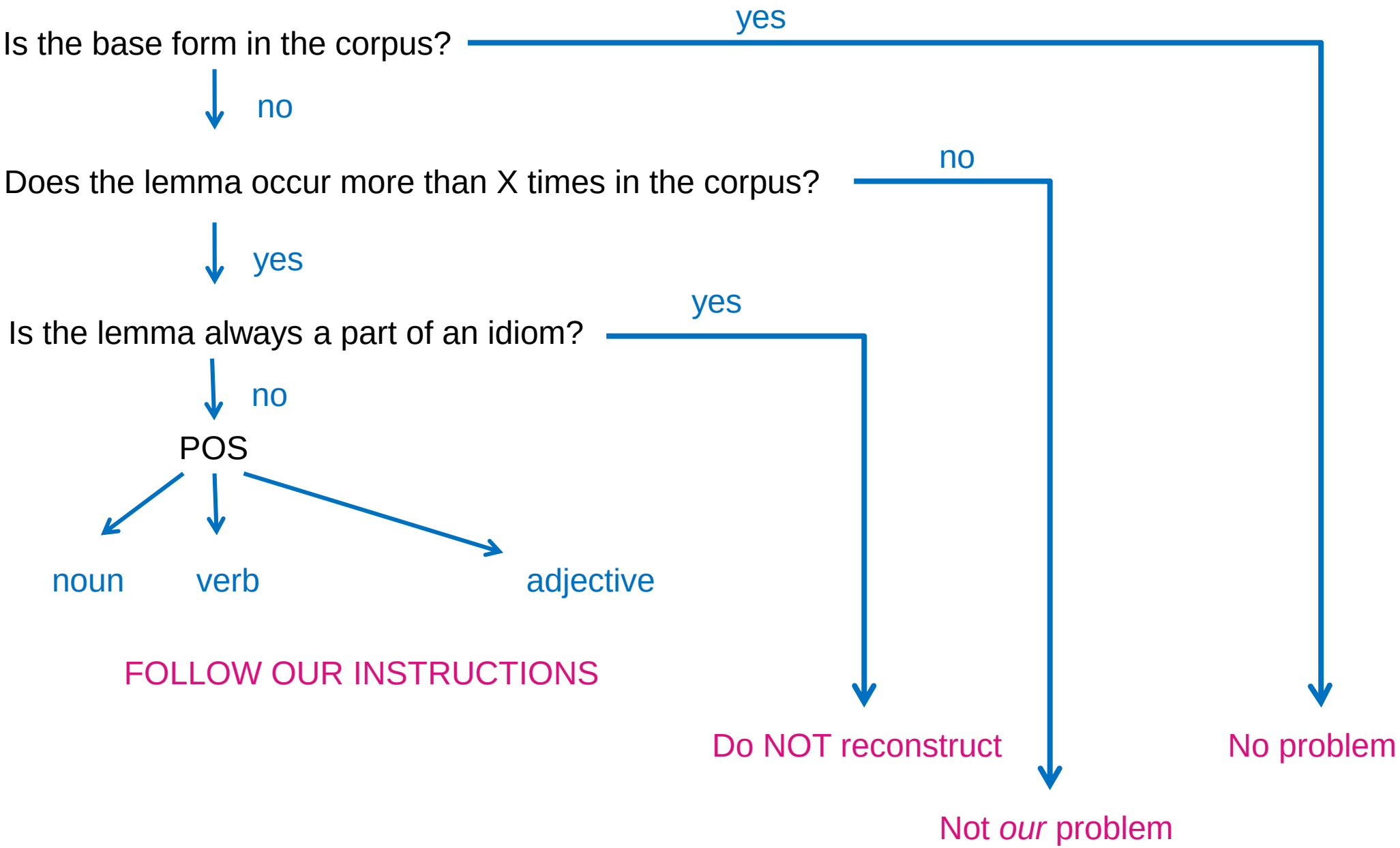
2. as the result of a **lack of data** → verify the lacuna in a larger corpus (SYN v6: 4G)

Expected frequency of dictionary form

$$\mathbb{E}[Fq_{DF}] = p(\mathrm{RGC}) \times Fq(\mathrm{lemma}) = \frac{Fq(\mathrm{RGC})}{Fq(\mathrm{POS})} \times Fq(\mathrm{lemma})$$

Only cases where $E[Fq_{DF}] \geq 3.8$ were considered – lower limit of binomial confidence interval for 3.8 (in 100M corpus) is 1

Recommendations for lexicographers


Is the base form in the corpus?
yes
no
Does the lemma occur more than X times in the corpus?
no
yes
Is the lemma always a part of an idiom?
yes
no
POS
noun
verb
adjective
FOLLOW OUR INSTRUCTIONS
Do NOT reconstruct
Not *our* problem
No problem

Verbs (400)

in most cases, infinitive form attested in a larger corpus

Rare exceptions:
A. semantic preference for past tense:
infrequent iteratives (*ptávat se* “to ask [repeatedly]”, *mluvívat* “to speak [repeatedly]”, *scházívat* “to miss [repeatedly]”)

B. restricted collocability (idioms):
mít vystaráno (“problem solved”), *bejvávalo* (“Those were the days.”)

→ recommendation: reconstruct
with exception of idioms

Adjectives (1000)

supporting POS (noun modifications)
formally very regular; Nsg.m.: *-ý/-í*

A1. limited collocability – multi-word terms:
kyselina hyaluronová (“hyaluronic acid”), *euklidovská geometrie* (“Euclidean geometry”), *lymská borelióza* (“Lyme borreliosis”)

A2. limited collocability – non-term MWE/idioms:
slonová kost (“ivory”), *jáma lvová* (“lion’s den”), *zcuchané vlasy* (“tousled hair”), *ustlaná postel* (“made-up bed”)

Adjectives (cont'd)

B1. semantic incompatibility with masculine
těhotná ("pregnant"), *vdaná* („married [woman]"),
vnadná ("luscious")

B2. semantic incompatibility with singular
nesčetní ("countless")

→ recommendations:

- A1 (terms): reconstruct
- A2 (MWE): headword = MWE
- B: as close as possible to dictionary form

Nouns (400)

restricted collocability + reduced paradigm

A. restricted collocability

- idioms: *pozdě bycha honiti* (“to go chasing ifs”), *být v čudu* (“to be gone”), *jít na kutě* (“hit the hay”), *k nezaplacení* (“priceless”)
- compounds written separately: *na blízko* (“close to”), *do červena* (“reddish”), *z loňska* (“from the last year”); *bez meškání* („without delay“)

Nouns (cont'd)

B. incompatibility with singular – pluralia tantum: *dveře* ("door"), *brýle* ("glasses"), *noviny* ("newspaper")

C. preference for plural forms
(dictionary form can be found in a larger corpus)
ančovička ("anchovy"), *paterče* ("quintuplet"), *cisterciák* ("Cistercian")

→ recommendations:

- A (MWE): headword = MWE

cf. practice in older Czech dictionaries (hypotethical forms as headwords marked by *)

- B: as close as possible to the dictionary form (= Npl)
- C: reconstruct

Summary

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

→recommendation:
reconstruct

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

→recommendation:

do not reconstruct
and use MWE as headword

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

→recommendation:

do not reconstruct
and keep as close as possible to the dictionary form

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

→recommendation:

reconstruct
and add a marker/gloss/usage note

lacuna type		restricted collocability		incompatibility with RGC	preference for non-RGC	lack of data
		terminology (MWT)	idioms/ non-term. MWE			
POS	nouns	0	++	+	+	+
	adjectives	++	+	+	0	+
	verbs	0	+	0	+	++

→recommendation:
reconstruct

References

Čermák, F. (1995): Překladová lexikografie. In F. Čermák & R. Blatná (eds.), *Manuál lexikografie.* Jinočany: H&H, p. 230-248.

Filipec, J. (1995): Teorie a praxe jednojazyčného slovníku výkladového. In F. Čermák & R. Blatná (eds.), *Manuál lexikografie.* Jinočany: H&H, p. 14-49.

Kochová, P. & Opavská, Z. (eds.) (2016): *Kapitoly z koncepce Akademického slovníku současné češtiny.* Praha: Ústav pro jazyk český AV ČR.

Kováříková, D., Chlumská, L. & Cvrček, V. (2012): What belongs to a dictionary? The example of negation in Czech. In R. V. Fjeld & J. M. Torjusen (eds.), *Proceedings of the 15th EURALEX International Congress.* Oslo: University of Oslo.

Schnorr, V. (1991): Problems of Lemmatization in the Bilingual Dictionary. In F. J. Hausmann et al. (eds.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie, 3. Teilband.* Berlin – New York: Walter de Gruyter, p. 2813-2817.

Seidensticker, T. (2008): Lexicography: Classical Arabic. In K. Versteegh et al. (eds.). *The Encyclopaedia of Arabic Language and Linguistics Vol. 3.* Leiden – Boston: Brill Academic, p. 30-37.

Svensén, B. (2009): *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Wolski, W. (1989): Das Lemma in texttheoretischer Sicht. In F. J. Hausmann et al. (eds.), *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie, 1. Teilband.* Berlin – New York: Walter de Gruyter, p. 366-371.

Thank you for your attention!

This study was written within the programme Progres Q08 *Czech National Corpus* implemented at the Faculty of Arts, Charles University.