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


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

<table>
<thead>
<tr>
<th>POS</th>
<th>nouns</th>
<th>adjectives</th>
<th>verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>POS</strong></td>
<td><strong>restricted collocability</strong></td>
<td><strong>incompatibility with RGC</strong></td>
<td><strong>preference for non-RGC</strong></td>
</tr>
<tr>
<td></td>
<td>terminology (MWT)</td>
<td>idioms/ non-term. MWE</td>
<td></td>
</tr>
<tr>
<td>nouns</td>
<td>0</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>adjectives</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>verbs</td>
<td>0</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

\[ E[F_{q_{DF}}] = p(\text{RGC}) \times F_q(\text{lemma}) = \frac{F_q(\text{RGC})}{F_q(\text{POS})} \times F_q(\text{lemma}) \]

Only cases where \( E[F_{q_{DF}}] \geq 3.8 \) 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

No problem

Not our 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]”, mluvít “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* ("luscieous")

B2. semantic incompatibility with singular

*nesčetná* ("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 (hypothetical forms as headwords marked by *)
• B: as close as possible to the dictionary form (= Npl)
• C: reconstruct
<table>
<thead>
<tr>
<th>POS</th>
<th>lacuna type</th>
<th>restricted collocability</th>
<th>incompatibility with RGC</th>
<th>preference for non-RGC</th>
<th>lack of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>nouns</td>
<td>terminology (MWT)</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>adjectives</td>
<td>idioms/ non-term. MWE</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>verbs</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>+</td>
<td>++</td>
</tr>
</tbody>
</table>

→ recommendation: reconstruct
<table>
<thead>
<tr>
<th>lacuna type</th>
<th>restricted collocability</th>
<th>incompatibility with RGC</th>
<th>preference for non-RGC</th>
<th>lack of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>terminology (MWT)</td>
<td>idioms/ non-term. MWE</td>
<td>incompatibility</td>
<td>preference</td>
</tr>
<tr>
<td>nouns</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>adjectives</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>verbs</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>

→ recommendation:

do not reconstruct and use MWE as headword
<table>
<thead>
<tr>
<th>lacuna type</th>
<th>restricted collocability</th>
<th>incompatibility with RGC</th>
<th>preference for non-RGC</th>
<th>lack of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>terminology (MWT)</td>
<td>idioms/ non-term. MWE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nouns</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>adjectives</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>verbs</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>

→ recommendation:

do not reconstruct
and keep as close as possible to the dictionary form
<table>
<thead>
<tr>
<th>lacuna type</th>
<th>restricted collocability</th>
<th>incompatibility with RGC</th>
<th>preference for non-RGC</th>
<th>lack of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>terminology (MWT)</td>
<td>idioms/ non-term. MWE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nouns</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>adjectives</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>verbs</td>
<td>0</td>
<td>+</td>
<td>0</td>
<td>+</td>
</tr>
</tbody>
</table>

→recommendation: reconstruct and add a marker/gloss/usage note
<table>
<thead>
<tr>
<th>POS</th>
<th>lacuna type</th>
<th>restricted collocability</th>
<th>incompatibility with RGC</th>
<th>preference for non-RGC</th>
<th>lack of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>terminology (MWT)</td>
<td>idioms/ non-term. MWE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nouns</td>
<td>restricted collocability</td>
<td>0</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>incompatibility with RGC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>preference for non-RGC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>verbs</td>
<td>lack of data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Recommendation:**
reconstruct
References


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.