Martina Waclawičová

DIALECT DICTIONARY AND LEXICALIZATION OF DIALECT PHENOMENA

Abstract This paper investigates Czech territorial dialect lexicography, particularly the incorporation of lexicalized phonological and morphological phenomena into differential dialect dictionaries. It examines the methodologies used in current Czech dialect dictionary creation. The study analyzes corpora of the Czech National Corpus (CNC), including ORAL v1, ORTOFON v2, and DIALEKT, to trace the lexicalization of dialect features. Observing lexicalization in Czech dialects through corpus analysis elucidates its various phases, from initial frequency increase to subsequent restriction in specific lexical units. Through illustrative examples (semi-semiconsonantal *u̯*, hard *ł*, the nominative plural ending of the masculine animatum noun *-í*), the study sheds light on lexicalization's evolution and distribution in contemporary spoken Czech. Additionally, it addresses challenges in documenting regular and irregular dialectal variations and proposes that lexicalized forms, even when they do not alter the lemma, should be included in differential dictionaries. Such an approach would enhance the representation of dialectal differences, contributing to a more comprehensive understanding of Czech dialectology.

Keywords Czech; dialects; lexicalization; dialect dictionary

1. Dialect Lexicography

From the outset of dialectological research, studying and describing dialect vocabularies has been central. Dictionaries of territorial dialects vary in their approach, including territorial scope, time span, and entry organization, aiming for either comprehensive or differential vocabularies (Sochová, 1995).

Regarding Czech dialects, the following types exist (Vojtová, 2002): differential, alphabetically ordered dictionaries (Kellner; 1949, Lamprecht, 1963), onomasiological glossaries combined with differential dictionaries (Bachmannová, 1997; Holub, 2001; Sochová, 2001), complete subject explanatory dictionaries (Roudný, 1952). The ongoing Dictionary of Czech Dialects aims to map all Czech dialects as a relatively complete, non-differential dictionary (Ireinová, 2019). It is an undeniably valuable, but extremely time-consuming project (the author's team has been working on it since 2011), which for the time being publishes electronically the processed part of the passwords (Slovník nářečí českého jazyka, 2016–). Due to the complexity of mapping entire vocabularies, most local dialect dictionaries use a differential approach, either alone or with subject dictionaries. The differential dialect dictionary (Sochová, 1995, p. 260) captures lexical units that are specifically dialectal. These include namely dialect-specific lexical units (they do not have a standard counterpart), proper lexical dialectisms (they have a standard counterpart), semantic dialectisms (they have a different meaning than the standard form), frequency dialectisms (their frequency

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

is lower in standard language), stylistic dialectisms (the word is similar to the standard form, but its form is varied according to the phonological or morphological dialectal features) and contextual dialectisms (phrases). Challenges arise in including lexicalized phonetic or morphological phenomena in differential dictionaries. This study tracks lexicalization in spoken Czech corpora, focusing on integrating these phenomena into differential dialect dictionaries.

2. Corpora of Spoken Czech in the Czech National Corpus

The Czech National Corpus (CNC) showcases spoken Czech's regional diversity through several corpora. For observing lexicalization of lexical and morphological dialectal phenomena, the most relevant corpora are ORAL v1 (Kopřivová et al.), ORTOFON v2 (Kopřivová et al.), and the specialized corpus DIALEKT (Goláňová et al.). ORAL v1 (5.4 million tokens) captures contemporary informal spoken Czech from 2002–2011. Its single transcription plane highlights regular deviations from written pronunciation. ORTOFON v2 (2.1 million tokens) represents informal spoken Czech from 2012–2019, captured on two transcription levels (orthographic and phonetic) linked to audio recordings. Both corpora include adult speakers from all generations across the Czech Republic. DIALEKT (223,000 tokens) documents traditional territorial dialects, mostly through monologic speeches by the oldest generation of speakers (born between 1875 and 1957) from all Czech dialect areas, including Czech language islands in Poland. By comparing contemporary spoken Czech (ORAL v1 and ORTOFON v2) with traditional dialects (DIALEKT), we will investigate the presence of lexicalization in selected phenomena.

3. Development of Czech Dialects and Lexicalization of Phonetic and Morphological Phenomena

Czech dialects, like those of other languages, resulted from historical development. The main driver of divergent linguistic evolution was feudal fragmentation. However, from the 19th century onwards, population migration and the growth of large cities initiated dialect leveling, where narrowly regional features receded and broader regional features persisted, leading to dialect convergence. Differentiating phonetic and morphological phenomena do not vanish immediately across the entire vocabulary but undergo a gradual process. Regular and productive phenomena often undergo lexicalization, remaining limited to specific lexical units or word forms for a time. Eventually, they may either disappear entirely or persist as linguistic relics.

3.1 Examples of Lexicalization Observed in Spoken CNC Corpora

3.1.1 Semi-Semiconsonantal u̯

The pronunciation of 'v' as a semi-semiconsonantal u̯ was originally common in the dialect regions of northeastern and central Bohemia (Český jazykový atlas 5, 2005, p. 429). The DIALEKT corpus documents 143 lemmas where u̯ occurs in

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Northeast Bohemian and Central Bohemian dialects. Among these lemmas, the following are the most frequent (with an occurrence rate of over 50 instances per million words):

1.	povídat	brzy	všechen	zrovna	opravdu[1]
	tell	soon	every	just	really

Most frequent word forms (i.p.m. over 50):

2.	poṷdá	poṷdám	dřiṷ	zroṷna ṷšechno	vopraṷdu[2]
	povídá	povídám	dřív	zrovna všechno	pravdu (standard forms)
	he tells	I tell	earlier just	everything	really

The ORAL and ORTOFON corpora identically record *ṷ* (in the subcorpus of speakers coming from the northeastern and central Bohemia region) already lexicalized in the following forms (i.p.m. over 1), ORTOFON:

3.	(v)opraudu (i.p.m. 24,64)	prauda (8,96)	zrouna (6,72)
	opravdu	pravda	zrovna (standard forms)
	really	true	just

ORAL:

4.	(v)opraudu (i.p.m. 4,69)	prauda (4,35)	houno (3,35)
	opravdu	pravda	hovno (standard forms)
	really	true	shit

The occurrence of *ṷ* pronunciation is not limited to the oldest generation as the bearers of the historical form of the dialect, but is distributed evenly in both general corpora among all generations of speakers (from 18 to 72 years old). In this case, the DIALEKT corpus captures the initial process of lexicalization and the ORAL and ORTOFON corpora represent the advanced stage.

3.1.2 Hard ł

The historical pronunciation of the hard *ł*, primarily characteristic of eastern Moravia and Silesia, was recorded in most of the peripheral dialect areas during the 20th century (Český jazykový atlas, 2005, p. 172). Examples of its lexicalization can be found in the Northeast Bohemian dialect area in the form of pronunciation relics, documented in the DIALEKT corpus (with an occurrence rate of 1,022.1 instances per million words). Among the 48 lemmas, *být* – "to be" is significantly more prevalent (with an occurrence rate of 58.03 instances per million), while others have rates below

[1] povídat (i.p.m. 306,25), brzy (122,5), všechen (96,71), zrovna (70,92), opravdu (61,25)

[2] poṷdá (i.p.m. 119,28), poṷdám (74,15), dřiṷ (70,92), zroṷna (70,92), ṷšechno (54,8), vopraṷdu (51,58)

10, indicating random occurrences. Specifically, verbs in the past participle[3], where *ł* serves as the formant, exhibit this phenomenon:

5.	był	šeł	dał
	byl	šel	dal (standard forms)
	he was	he went	he gave

Occurrences of other forms are mostly random and have rates below 10 i.p.m. The lexicalization of this phenomenon is evident. While the ORTOFON corpus documents lexicalized pronunciation in the forms of the verb *být* ("to be") in other regions, it is no longer observed in Northeastern Bohemia. Additionally, the ORAL corpus does not record the pronunciation of the hard *ł*.

3.1.3 The Nominative Plural Ending of the Masculine Animatum Noun -í

Among the morphological phenomena, lexicalization is evident in the nominative plural ending of the masculine animatum noun *-í*, which originally characterized the dialect in the southwestern half of Bohemia as a regular and typical feature (Český jazykový atlas 4, 2002, p. 152). The DIALEKT corpus demonstrates its consistent occurrence across the vocabulary (with 61 lemmas). Higher frequencies (over 10 instances per million words) are observed in the following forms:

6.	klucí	chlapcí	Ňemcí[4]	Američaňí	vojácí[5]
	kluci	chlapci	Němci	Američani	vojáci (standard forms)
	boys	boys	Germans	Americans	soldiers

In the ORAL and ORTOFON corpora, the phenomenon is already clearly lexicalized in the forms of *klucí* ("boys", with occurrences of 14.78 per million words in ORAL and 8.59 per million words in ORTOFON). However, in other instances, the phenomenon appears randomly (with occurrences of less than 1 per million words), often exclusively in the speech of a single speaker, demonstrating the influence of idiolect.

3.2 Detecting Lexicalization Phases Through Corpus Analysis

These examples show how observing the distribution of phenomena in the corpora of older dialect forms and contemporary spoken Czech can help us reveal lexicalization and its phases. The first phases are typically characterized by an increase in the frequency of the observed phenomenon for certain lexical units or their groups. In the next phase, the phenomenon is restricted exclusively to certain lexical units, the frequency of which may gradually decrease.

[3] tag VpYS---XR-AA– (i.p.m. 119,28), VpNS---XR-AA– (22,57), VpQW---XR-AA– (12,9), VpYS---XR-NA– (12,9)

[4] The high frequency of the words *Ňemcí*, *Američaňí* and *vojácí* is influenced by the choice of topics the dialect speakers monologue about, in this case the topic of World War II.

[5] klucí (i.p.m. 106,38), chlapcí (54,8), Ňemcí (45,13), Američaňí (29,01), vojácí (12,9)

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

4. Capturing Lexicalized Phenomena in a Differential Dictionary

How are lexicalized phonological and morphological dialect phenomena incorporated into differential dialect dictionaries? The current practice for creating Czech dialect dictionaries involves comparing entries with Slovník spisovného jazyka českého (Dictionary of the written Czech language, SSJČ) (Bachmannová, 2016, p. 13). Entries are included if they are absent from the SSJČ, have spelling differences, or exhibit vowel quantity variations. Additionally, words labeled as dialectal, regional, or folk in the SSJČ are also included. In other dictionaries, only significant lexical differences are noted, excluding minor variations like vowel quantity (Bachmannová, 1998, p. 11).

Typically, words that differ from the standard form only by a regular variation in the basic form are usually not included in dictionaries. Such cases are usually discussed in introductory essays that describe the dialectal features being mapped. Words with regular dialectal changes in non-lemma forms are often covered in the morphological descriptions of the dialect rather than as dictionary entries.

More challenging are lexicalized (irregular or residual) phonological or morphological changes. These are generally not included as separate entries but are mentioned in introductory sections, such as (Bachmannová, 2016, p. 9): Archaic Feature: Pronunciation of the hard "l" (ł) in past participles (*był, přišeł, měła*[6]) and occasionally elsewhere (*ke stołu, na dłuh, hołt, vokoło*[7]).

If a lexicalized phenomenon occurs in a word's lemma, it should be a separate dictionary entry. For instance (Bachmannová, 2016, p. 286):

7. povdat (poudat)
   povídat (standard form)
   to tell

However, if the phenomenon appears only in non-lemma forms, such lexical units are not included. Previously regular but now lexicalized forms, such as *klucí* (3.1.3), do not appear in dictionaries because their lemmata match the standard form and the change occurs only in other forms.

These cases should be included in differential dialect dictionaries, even if they do not show lemma differences. Lexicalized differential forms in other word forms should be noted and commented on within the entry. The question remains whether to list them under the non-differential lemma:

8. kluk – lexikalizovaný tvar nom. pl. *klucí*
   kluk – lexicalized form nom. pl. *klucí*

[6] He was, he came, she had.

[7] To the table, on credit, simply, around.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

or introduce the entry with the differential form:

9. klucí – nom. pl. lexikalizovaný tvar (zákl. tvar *kluk*)
   klucí – nom. pl. lexicalized form (lemma *kluk*)

The first approach aligns with existing entries, while the second allows easier retrieval.

5. Conclusion

Tracking the occurrence of phenomena in various spoken language corpora reveals the progressive lexicalization of originally regular phonological or morphological phenomena. Including these lexicalized forms in differential dialect dictionaries, especially when they appear in non-lemma forms, would provide a more comprehensive dialect description. Even if the lemma matches the standard form, describing the dialectal difference within the entry enhances the overall depiction of the dialect.

References

Bachmannová, J. (1997). *Podkrkonošský slovník.* Academia.

Bachmannová, J. (2016). *Slovník podkrkonošského nářečí.* Academia.

Český jazykový atlas 4 (2002). Academia.

Český jazykový atlas 5 (2005). Academia.

Goláňová, H., Waclawičová, M., & Lukeš, D. (2021). DIALEKT: nářeční korpus, verze 2 z 23. 12. 2021. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from http://www.korpus.cz

Holub, Z. (2001). *Lexikon nejjižnějšího úseku českých nářečí.* Aleš Čeněk.

Ireinová, M. (2019). O Slovníku nářečí českého jazyka. In É. Császári, & M. Imrichová (Eds), *Király Péter 100. Tanulmánykötet Király Péter tiszteletére I.* (pp. 223–229). Szláv Filológiai Tanszék.

Kellner, A. (1949). *Východolašská nářečí II.* Dialektologická komise při Matici moravské.

Kopřivová, M., Laubeová, Z., Lukeš, D., Poukarová, P., & Škarpová, M. (2020). ORTOFON v2: Korpus neformální mluvené češtiny s víceúrovňovým přepisem. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from http://www.korpus.cz

Kopřivová, M., Lukeš, D., Komrsková, Z., Poukarová, P., Waclawičová, M., Benešová, L., & Křen, M. (2017). ORAL: korpus neformální mluvené češtiny, verze 1 z 2. 6. 2017. Ústav Českého národního korpusu FF UK. Retrieved January 10, 2024, from http://www.korpus.cz

Lamprecht, A. (1963). *Slovník středoopavského nářečí.* Krajské nakladatelství Ostrava.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Roudný, J. (1952). *Slovník nářečí podle zápisů z Úlibic na Jičínsku.* [Candidate dissertation, Univerzita Karlova v Praze].

Slovník nářečí českého jazyka (2016–). Dialektologické oddělení Ústavu pro jazyk český AV ČR, v. v. i. Retrieved January 10, 2024, from http://sncj.ujc.cas.cz

Slovník spisovného jazyka českého (1960–1971). Academia.

Sochová, Z. (1995). Nářeční lexikografie. In F. Čermák, & R. Blatná (Eds.), *Manuál lexikografie* (pp. 249–264). H&H.

Sochová, Z. (2001). *Lašská slovní zásoba.* Academia.

Vojtová, J. (2002). Perspektivy nářeční lexikografie. *Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná, 51*(A50), 105–110. https://hdl.handle.net/11222.digilib/100913

Acknowledgements

This work has been supported by Charles University Research Centre program No. 24/SSH/009.

Contact information

Martina Waclawičová
Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague
martina.waclawicova@ff.cuni.cz