sciendo
DOI 10.2478/jazcas-2025-0011

THE SONORITY SEQUENCING PRINCIPLE IN HISTORICAL CZECH: A CORPUS-BASED STUDY

MARKÉTA ZIKOVÁ[1] – RADEK ČECH[2] – MARTIN BŘEZINA[3] – PAVEL KOSEK[4]
[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-0635-8893)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-4412-4588)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6986-9754)
[4]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0001-6678-9989)

ZIKOVÁ, Markéta – ČECH, Radek – BŘEZINA, Martin – KOSEK, Pavel: The Sonority Sequencing Principle in Historical Czech: A Corpus-based Study. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 123 – 131.

Abstract: This paper investigates the application of the Sonority Sequencing Principle (SSP) in historical Czech through a corpus-based approach. Drawing on texts from the 14th and 17th centuries, we examine the structure of word-initial and word-final consonant clusters with respect to both the strict and mild versions of the SSP. The results reveal two frequent types of violations: those involving liquids—specific to the diachronic development of Czech—and those involving sibilants, which are common cross-linguistically. Our findings provide new empirical evidence for the study of historical phonotactics in Slavic languages.

Keywords: Sonority Sequencing Principle, syllable structure, phonotactics, consonant clusters, corpus linguistics, historical Czech

1 INTRODUCTION

The Sonority Sequencing Principle (SSP) accounts for cross-linguistic phonotactic patterns in syllable structure. It states that syllables follow a universal sonority contour, with sonority peaking at the syllable nucleus and decreasing toward the margins, i.e. the onset and coda (Clements 1990; Zec 1995). This contour is determined by sonority—a scalar property reflecting relative loudness and acoustic energy—which follows a universal hierarchy: vowels > glides > liquids > nasals > fricatives > plosives (Parker 2011).

Although the SSP is considered a *phonological* principle, its application is strongly influenced by *morphological* structure. Specifically, the principle applies more strictly within words than at word edges, where SSP violations are more common. This asymmetry between word-internal and word-peripheral positions is

illustrated by two Czech words: *rvala* 'she tore' and *larva* 'larva'. Both are bisyllabic and contain the consonant cluster *rv*, in which a more sonorous liquid is followed by a less sonorous fricative. Whether this cluster is tautosyllabic or heterosyllabic depends on its position within the word.

In *rvala*, the cluster occurs word-initially and thus forms the onset of the first syllable (*rva.la*). In contrast, in *larva*, where the cluster is word-internal, the analogous syllabification is ungrammatical: **la.rva*. Instead, the cluster is heterosyllabic (*lar.va*), meaning that only the fricative forms the onset (of the second syllable), while the liquid serves as the coda (of the first syllable). In sum, the word-internal cluster *rv* conforms to the SSP, whereas the word-initial cluster in *rva.la* violates it. In the syllable #*rva*, sonority does not decrease away from the nucleus; instead, it rises within the onset, contrary to the expected sonority contour.

Such SSP-violating peripheral clusters are typical of Slavic languages, including Czech, and originate in their historical development (Bethin 1998). They evolved from Proto-Slavic forms containing *jer* vowels (ь/ъ), which were lost in weak positions. The SSP-violating form *rva.la* thus evolved from the SSP-conforming *rъ.va.la*.

In this paper, we examine peripheral consonant clusters in Czech. While the phonotactic patterns of contemporary Czech are relatively well documented—albeit typically without reference to the SSP (cf. Bičan 2013; Lukeš and Šturm 2017)—our analysis focuses on historical Czech, which remains understudied from this perspective. This research provides new empirical evidence that may serve as a foundation for future comparative work on the development of syllable structure in Czech.

The aim of this paper is to determine the extent to which peripheral clusters in historical Czech conform to the SSP, and the extent to which they violate it. In the case of violating clusters, we focus on distinguishing between accidental violations—those arising from the diachronic development of Proto-Slavic—and systematic ones, which are attested cross-linguistically and not limited to Slavic languages, as discussed, for example, in Yin et al. (2023).

The paper is organized as follows. Section 2 introduces the theoretical background of the SSP. Section 3 presents the corpus of historical Czech and outlines the methodology used for data extraction. Sections 4 provides the results of our corpus-based analysis and their interpretation, respectively. Finally, Section 5 concludes the paper.

2 THE SONORITY SEQUENCING PRINCIPLE

Following Parker (2011), we adopt a seven-level sonority scale, with vowels at the highest levels and obstruents at the lowest levels, as illustrated in Tab. 1. The table shows the correspondences between sonority classes, IPA segments, and graphemes.

sonority level	sonority class	segments	graphemes
7	non-high vowels	/a aː e eː o oː/	*a á e ě é o ó*
6	high vowels	/i iː u uː/	*i y í ý u ú ů*
5	glides	/j/	*j*
4	liquids	/r l/	*r, ŕ, l, ľ, ĺ, ł*
3	nasals	/m n ɲ/	*m n ň*
2	fricatives	/f v s z ʃ ʒ ʂ x ɦ/	*v, f, s, ś, z, ź, š, ž, ř, ch, h*
1	stops and affricates	/p b t d t͡s t͡ʃ c ɟ k g/	*p, b, t, d, c, ć, č, ť, ď, k, g*

Tab. 1. The sonority scale and the segmental inventory of historical Czech

Based on this scale, sonority profiles of words can be constructed, as illustrated below. Fig. 1 displays the sonority profiles of the words *trám* ‘beam’ and *nárt* ‘instep’, which contain consonant clusters in onset and coda positions, respectively. In both cases, the sonority profile features a single peak—formed by a vowel—which corresponds to the syllable nucleus (N). From this nucleus, sonority decreases toward both syllable margins, in accordance with the predictions of the SSP.

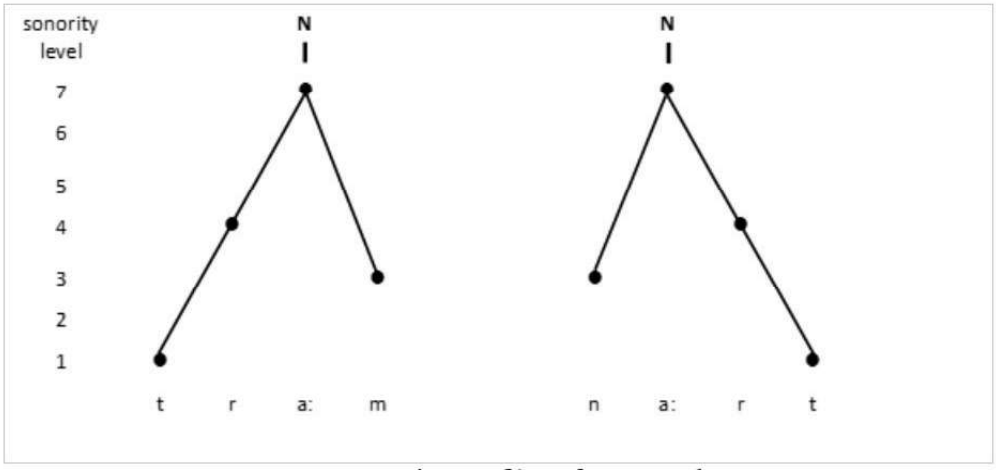

Fig. 1. Sonority profiles of *trám* and *nárt*

Fig. 2 displays two further examples attested in Czech, where additional sonority peaks are formed by consonants at word margins. In the word *lotr* ‘rascal’, the peak-forming consonant (a liquid *r*) appears word-finally. It functions as a syllable nucleus, and the resulting bisyllabic structure *lo.tr* is thus consistent with the SSP.

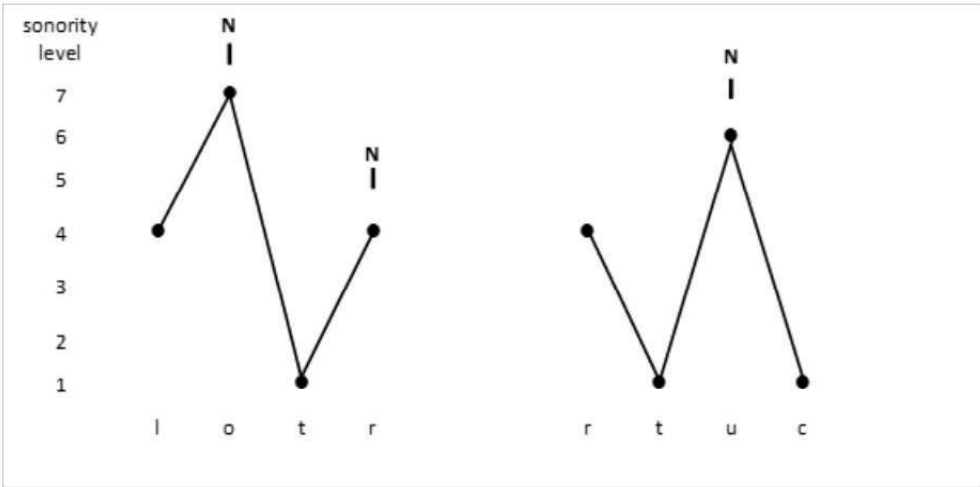

Fig. 2. Sonority profiles of *lotr* and *rtuť*

In the monosyllabic word *rtuť* 'mercury', the peak-forming liquid appears word-initially. Unlike in the previous example, however, it does not function as a syllable nucleus, and the onset cluster #*rt* therefore violates the SSP.

To distinguish between two types of peak-forming consonants attested in Czech, we adopt the terms *syllabic* consonants for those that function as syllable nuclei, and *trapped* consonants for those that violate the SSP (cf. Scheer 2009). In Czech, these two types differ in two main respects. First, they exhibit an asymmetrical distribution at word margins: syllabic consonants occur only at the right margin (as in *lo.tr* above), while trapped consonants appear at both the left margin (e.g. *r* in *rtuť*) and the right margin (e.g. *s* in *koks* 'coke'). Second, they differ in their segmental properties: syllabic consonants are limited to liquids (and sometimes nasals). Syllabic and trapped liquids are discussed further in Section 4.1.

Summing up, liquids that form sonority peaks in word-final position are syllabic consonants. Other consonants in word-final peak position are trapped and violate the SSP. Likewise, all peak-forming consonants in word-initial position are trapped.

Trapped consonants that form sonority peaks represent one SSP-violating type. A second type involves word-peripheral consonants such as the plosive *p* in *pták* 'bird' and *t* in *fakt* 'fact'. Like trapped consonants, these plosives violate the SSP because sonority does not decrease toward the onset margin in #*pt* or toward the coda margin in *kt*#. However, unlike trapped consonants, these plosives do not form sonority peaks, as they are adjacent to another consonant of the same sonority level.

The contrast between trapped consonants such as *r* in #*rt* and non-trapped SSP violations such as *p* in #*pt* is sometimes captured by distinguishing two versions of the SSP: a strict version and a mild version. According to the strict version, sonority must decrease continuously from the syllable nucleus toward both syllable margins. In contrast, the mild version requires only that sonority must not rise toward the

onset or the coda. Under the strict SSP, both #*rt* and #*pt* violate the principle, as sonority does not decrease in either cluster. Under the mild version, however, #*pt* is a well-formed onset: it constitutes a sonority plateau, where sonority remains constant but does not rise, unlike in the trapped cluster #*rt*.

In what follows, we test both versions of the SSP on historical Czech data. Our aim is to provide a typology of SSP violations attested in a corpus of Czech texts written between the 14th and 17th centuries, and to interpret them from a broader cross-linguistic perspective.

3 CORPUS DATA EXTRACTION

The corpus of historical Czech used in our analysis is based on 26 texts of varying token counts, written between the 14th and 17th centuries.[1] It comprises 113,159 tokens, understood as graphical words, i.e. sequences of graphemes delimited by spaces on both sides. To obtain relevant data for testing the validity of both the strict and mild versions of the SSP, tokens were processed according to the following algorithm.

In the first step, non-syllabic prepositions such as *k* 'to', *s* 'with', *v* 'in', and *z* 'from', were joined with the tokens that followed them. This means that originally separate tokens in prepositional phrases—such as *z toho* 'from it'—were treated as a single unit for analysis, yielding onsets like #*zt*. The rationale behind this step is that non-syllabic prepositions do not form independent phonological units; they always procliticize to the following word. Similarly, the non-syllabic second-person auxiliary *s* forms a single unit with the preceding host, as in the verbal token *byls* 'you were'. In this case, however, the *s*-encliticization is already reflected in the orthography.

In the next step, each token was annotated according to the sonority scale presented above in Tab. 1. The annotation was carried out automatically using a sonority parser developed by the authors, as described in Ziková et al. (2023).

From these sonority annotated data, we extracted word-initial and word-final demisyllables. The demisyllable is defined as a unit consisting of an onset or coda together with its adjacent vowel nucleus. For example, a word *křtalt* 'character' contains the word-onset demisyllable #*křta* and the word-coda demisyllable *alt*#.

[1] Rather than working with original manuscripts, we relied on published editions to facilitate automatic processing. We used the following critical editions. Cejnar, J. (1964) Nejstarší české veršované legendy; Daňhelka, J. (1952) *Husitské skladby Budyšínského rukopisu*; Flajšhans, V. (1882) *Staročeská píseň o božím těle ze XIII. století*; Hrabák, J. – Vážný, V. (1959) *Dvě legendy z doby Karlovy*; Janošík-Bielski, M. (2008) *Modlitba Kunhutina*; Kolár, J. (1959) *Frantové a grobiáni: z mravokárných satir 16. věku v Čechách*; Lomnický z Budče, Š. (1572) *Instrukcí aneb Krátké naučení*; Patera, A./Černá, A. M. (1881/2008) *Hradecký rukopis*; Vážný, V. (1963) *Alexandreida*; Žampachová, K. (2021) *Umučení rajhradské – edice a jazykový rozbor*.

In the next step, we extracted the consonantal parts of these demisyllables, yielding 702 word-onset types (e.g. *#křt*) and 132 word-coda types (e.g. *lt#*). The observed asymmetrical distribution of onsets and codas confirms the tendency toward open syllables, that is syllables lacking codas. This finding aligns with previous observations for contemporary Czech, as reported by Lukeš and Šturm (2017).

The collected word onsets and codas were then analyzed according to both versions of the SSP, as presented in the following section.

4 TWO TYPES OF SSP VIOLATIONS

A closer inspection of the data reveals two frequent types of SSP violations, involving clusters with liquids and sibilants, respectively. While the first type is specific to the historical development of Czech (and Slavic languages more generally), the second type is well attested beyond Slavic as well.

4.1 Liquids

Liquids are default components of well-formed complex onsets (as in *trick*) or codas (as in *culp*). Moreover, liquids are often syllabic. Although consonant strings containing syllabic liquids such as *pl#* in *people* may appear to violate the SSP, this is not the case, as they are not true clusters: the syllabic liquid forms the syllable's nuclear peak, just like a vowel.

In Section 2, we discussed word-final syllabic liquids (such as *r* in *lo.tr*). In addition to these, word-internal syllabic liquids are also attested in Czech. For example, the words *držet* 'to hold' and *pokrm* 'food' may at first appear to contain an onset cluster *#drž* or a coda cluster *krm#*, respectively. However, these are not true clusters, as the liquid is syllabic. As a result, bisyllabic forms *dr.žet* and *po.krm* are fully consistent with both versions of the SSP.[2]

In addition to clear-cut cases of syllabic liquids, our corpus also includes ambiguous forms such as *slza* 'tear' or *řekl* 'he said'. These words may be either monosyllabic or bisyllabic, and we currently lack sufficient evidence to definitively support either interpretation; see Ziková et al. (2025) for details on the method used to distinguish between the liquid types.

This ambiguity stems from diachronic development. Originally, these words were monosyllabic, containing trapped liquids that violated the SSP. Over time, however, these trapped liquids gradually shifted to SSP-conforming syllabic consonants between the 14th and 16th centuries, resulting in bisyllabic forms *sl.za* and *ře.kl*, as attested in contemporary Czech.

[2] Recall that our analysis is restricted to complex onsets and codas. This criterion also applies to tokens with syllabic liquids. Accordingly, words like *čtvr.tek* 'Thursday' (with the complex onset *#čtv*) and *prst* 'finger' (with the complex coda *st#*) were included in the analyzed sample, whereas words like *dr.žet* and *po.krm*, which contain both a simple onset (*#d*, *#p*) and a simple coda (*t#*, *m#*), were excluded.

As shown in Tab. 2, these ambiguous forms constitute a relatively large share of all recorded SSP violations in word codas—approximately one third under the mild version and one quarter under the strict version of the SSP.

	mild SSP		strict SSP	
	word onset	word coda	word onset	word coda
all violating types	418	43	574	57
violating types with liquids	66	15	67	15
proportion of liquid types	15.8%	34.8%	11.6%	26.3%

Tab. 2. The proportion of SSP-violating types with ambiguous liquids

4.2 Sibilants

The second type of SSP-violating forms attested in our corpus involves sibilants. Unlike ambiguous liquids, which are specific to historical Slavic, sibilants violate the SSP cross-linguistically. For example, Harris (1994) shows that nearly every two-segment word onset in English that conforms to the strict SSP has a corresponding variant expanded by a sibilant that violates both versions of the SSP; cf. pairs such as /pr/*ize* – /spr/*ead*, /tr/*ick* – /str/*ike*, or /pl/*ain* – /spl/*it*, where the first member conforms to the SSP in onset position, while the second violates it due to the initial sibilant.

The reason why sibilants behave in this specific way remains a matter of debate in the literature (Goad 2011). Nevertheless, it is clear that sibilants—including fricatives /s z ʃ ʒ/ and affricates /t͡s t͡ʃ/—contribute significantly to SSP violations at word edges in historical Czech as well, as illustrated in Tab. 3. The table shows that both word-onset clusters with sibilants (e.g. #*spr* in *spravuje* ‘(s)he manages’) and word-coda clusters (e.g. *dž#* in *poněvadž* ‘whereas’) account for more than 50% of all violating types across all examined parameters.

	mild SSP		strict SSP	
	word onset	word coda	word onset	word coda
all violating types	418	43	574	57
violating types with sibilants	299	23	400	33
proportion of sibilant types	71.5%	53.5%	69.7%	57.9%

Tab. 3. The proportion of SSP-violating types with sibilants

5 CONCLUSION

This paper has examined the extent to which the Sonority Sequencing Principle is reflected in the syllable structure of historical Czech. Using a corpus-based approach, we evaluated both strict and mild versions of the SSP and identified substantial violation rates in both word-initial and word-final positions.

The most frequent sources of these violations are sibilants and, to a lesser extent, liquids that alternate between trapped and syllabic status. The latter type of violation is specific to historical Czech and is connected to the evolution of Proto-Slavic *jer* vowels. Since all trapped liquids were gradually eliminated in word-internal and word-final positions, we expect SSP conformity to differ significantly between historical and contemporary Czech in this regard.

In contrast, sibilants are well-known SSP violators cross-linguistically. In this respect, historical Czech follows a general pattern, and we expect the same for contemporary Czech. Moreover, sibilants are often involved in morphologically complex clusters. For example, English features three productive affixes consisting of the sibilant /s/ that mark possessive, nominal plural, and verbal agreement. The concatenation of these sibilant markers produces complex codas, as seen in *hy*/mn's/, *atte*/mpt-s/, and *he tru*/st-s/, which are not found in monomorphemic words.

A similar situation is observed in Czech. As mentioned above, sibilants appear both in proclitic prepositions and in the enclitic verbal auxiliary. As a result, they give rise to phonotactically specific clusters, such as the onset *#zž* in the prepositional phrase *z života* 'from (the) life', which, once again, have no counterparts in morphologically simplex words. Since the set of proclitic prepositions also includes non-sibilant forms like *v* 'in' and *k* 'to', these too contribute to the formation of word onsets that violate the strict version of the SSP, as *#vb* and *#kpr* in prepositional phrases *v bázni* 'in fear' and *k práci* 'to work', respectively.

In general, morphological complexity undoubtedly influences violations of the SSP. The extent to which this applies to both historical and contemporary Czech—where proclitic prepositions also serve as productive verbal prefixes—remains an open question for future research.

ACKNOWLEDGEMENTS

We express our gratitute to Tomáš Vlček for his assistance with the processing of historical data. This research was supported by the Czech Science Foundation under the project *Development of Syllabic Sonorants in Czech* (GA23-04719S), awarded to Markéta Ziková.

References

Bethin, Ch. (1998). Slavic prosody. Language change and phonological theory. Cambridge: Cambridge University Press.

Bičan, A. (2013). Phonotactics of Czech. Frankfurt am Main: Peter Lang.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In: J. Kingston – M. E. Beckman (eds.): Papers in laboratory phonology I: Between the grammar and physics of speech. Cambridge: Cambridge University Press, pp. 283–333.

Goad, H. (2011). The representation of *s*C clusters. In: M. van Oostendorp et al. (eds.): The Blackwell companion to phonology. Oxford: Wiley-Blackwell, pp. 898–923.

Harris, J. (1994). English sound structure. Oxford: Blackwell.

Parker, S. (2011). Sonority. In: M. van Oostendorp et al. (eds.): The Blackwell companion to phonology. Oxford: Wiley-Blackwell, pp. 1160–1184.

Scheer, T. (2009). Syllabic and trapped consonants in the light of branching onsets and licensing scales. In: G. Zybatow et al. (eds.): Studies in formal Slavic phonology, morphology, syntax, semantics, and information structure. Frankfurt am Main: Peter Lang, pp. 411–426.

Šturm, P., and Lukeš, D. (2017). Fonotaktická analýza obsahu slabik na okrajích českých slov v mluvené a psané řeči. Slovo a slovesnost, 78(2), pp. 99–118.

Yin, H., van de Weijer, J., and Round, E. (2023). Frequent violation of the sonority sequencing principle in hundreds of languages: How often and by which sequences? Linguistic Typology, 27(1), pp. 131–175.

Zec, D. (1995). Sonority constraints on syllable structure. Phonology, 12, pp. 85–129.

Ziková, M., Březina, M., Čech, R., and Kosek, P. (2023). Syllabic consonants in historical Czech and how to identify them. Jazykovedný časopis, 74(1), pp. 391–400.

Ziková, M., Březina, M., Čech, R., and Kosek, P. (2025). The shift away from the marked: Syllabic consonants in historical Czech. Glossa: a journal of general linguistics, 10(1), pp. 1–24.