# A Corpus Analysis of Collocations of Toponyms in Standard Latvian and Latgalian

prof. Sanita Martena

Izstrādāts ar ES ANM atbalstu projekta Nr. 5.2.1.1.i.0/2/24/I/CFLA/003
P&A grantā “Spatiotemporālā modeļa izstrāde un pielietošana Latgaliešu runas korpusa datu daudzveidīgas analīzes nolūkos”, Nr. RTU-PA-2024/1-0063

DE GRUYTER

Sociolinguistica
European Journal of Sociolinguistics

# Do Speakers of a National and a Collateral Regional Language Live in Different Mental Spaces? A Corpus Analysis of Collocations of Toponyms in Standard Latvian and Latgalian



| | |
|---|---|
| Journal: | *Sociolinguistica* |
| Manuscript ID | soci.2026.0001 |
| Manuscript Type: | research article |
| Keywords: | Corpus-Assisted Discourse Studies (CADS), Regional Collateral Language, Socio-onomastics, Toponyms, Latvia, Latgalian |




Abstract:

The paper discusses differences between the use of toponyms among speakers of Latvian as a small national language, and of Latgalian as a related regional collateral language. For understanding differences of perceptions related to Rīga / Reiga, the capital of Latvia, and Daugavpils / Daugovpiļs and Rēzekne, the two main cities in the region of Latgale, the paper analyses collocations in the main academic corpora of written Latvian and Latgalian, LVK2022 and MuLa2022. The collocations are grouped into general semantic categories, followed by a detailed analysis of the categories Places / Directions / Mobility, Education, People, and Languages and Speaking. A quantitative overview is followed by a qualitative analysis, and an interpretation of how these collocations

# Salīdzinājumam izmantotie tekstu korpusi

**Mūsdienu latgaliešu tekstu korpuss (MuLa 2022)**
**https://korpuss.lv/id/MuLa2022**

- Latgaliešu rakstu valodā publicēti (**1988–2021**) teksti (**2 milj**. vārdlietojumu).
- Metadati: teksta autors, nosaukums, publicēšanas vieta un gads, kā arī informācija par teksta veidu un žanru (publicistika (55%), daiļliteratūra (37%), populārzinātniski, zinātniski teksti(8%)).
- Korpuss **nav lemmatizēts un nav morfoloģiski marķēts**.

**Līdzsvarotais mūsdienu latviešu valodas korpuss (LVK 2022)**
**https://korpuss.lv/id/LVK2022**

- Latviešu literārajā valodā publicēti (**2000 – 2021**) teksti (**101 milj**. vārdlietojumu).
- Žanri:periodika (60%), daiļliteratūra (10%), zinātniski teksti (10%), teksti no „Vikipēdijas" (7%), normatīvie akti (7%), Saeimas stenogrammas (3%) un subtitri (3%)
- Automātiski **gramatiski marķēts**.

# RQ

1. What do collocations of the names of the three cities reveal about discourse prosodies related to them?
2. What differences are there regarding the conceptualisation of the three cities in Latgale in contrast to Latvia in general?
3. What do these differences imply for the perceptions of centre and periphery in a Latvian context and, more generally, for distinctions between mental spaces by speakers of a collateral regional language vis-à-vis the related national language?
4. What conclusions can there, finally, be drawn for the way how CADS (corpus-assisted discourse studies) can be connected to socio-onomastics?

# Pētījuma metodoloģija

1. **Frekvence** – atslēgvārdu biežums MuLa2022 un LVK2022.
2. **Kolokācijas** – vietvārdu piecas labās un kreisās puses kolokācijas (100 biežākās).
3. **Diskursīvo modeļu noteikšana** - diskursa prosodijas (*discourse prosodies*) jeb kā noteikti vārdi kontekstā iegūst kolokācijas ar pozitīvu vai negatīvu konotāciju un kādi diskursīvie modeļi izveidojas (Baker, 2006). Dažāda korpusu apjoma dēļ - %.
4. **Konkordanču analīze** (lietojums kontekstā).

# Frekvence

**MuLa2022**:

***Rēzekne*** (2,502), *Riezekne* (26), Režica (12), Rositten (5)

**Reiga** (1,991), Rīga (52)

***Daugovpiļs*** (684), *Daugavpiļs* (615) Daugpiļs (80), *Dinaburga* (60), *Dvinska* (63), *Borisoglebska* (2).

**LVK2022**:

***Rīga*** (203,946), *Reiga* (2)

**Daugavpils** (14,817), *Dinaburga* (214), *Dvinska* (67), *Borisogļebska* (5), *D-pils* (2), *Dpils* (1);

***Rēzekne*** (8,459) *Režica* (10), *Rositten* (4).

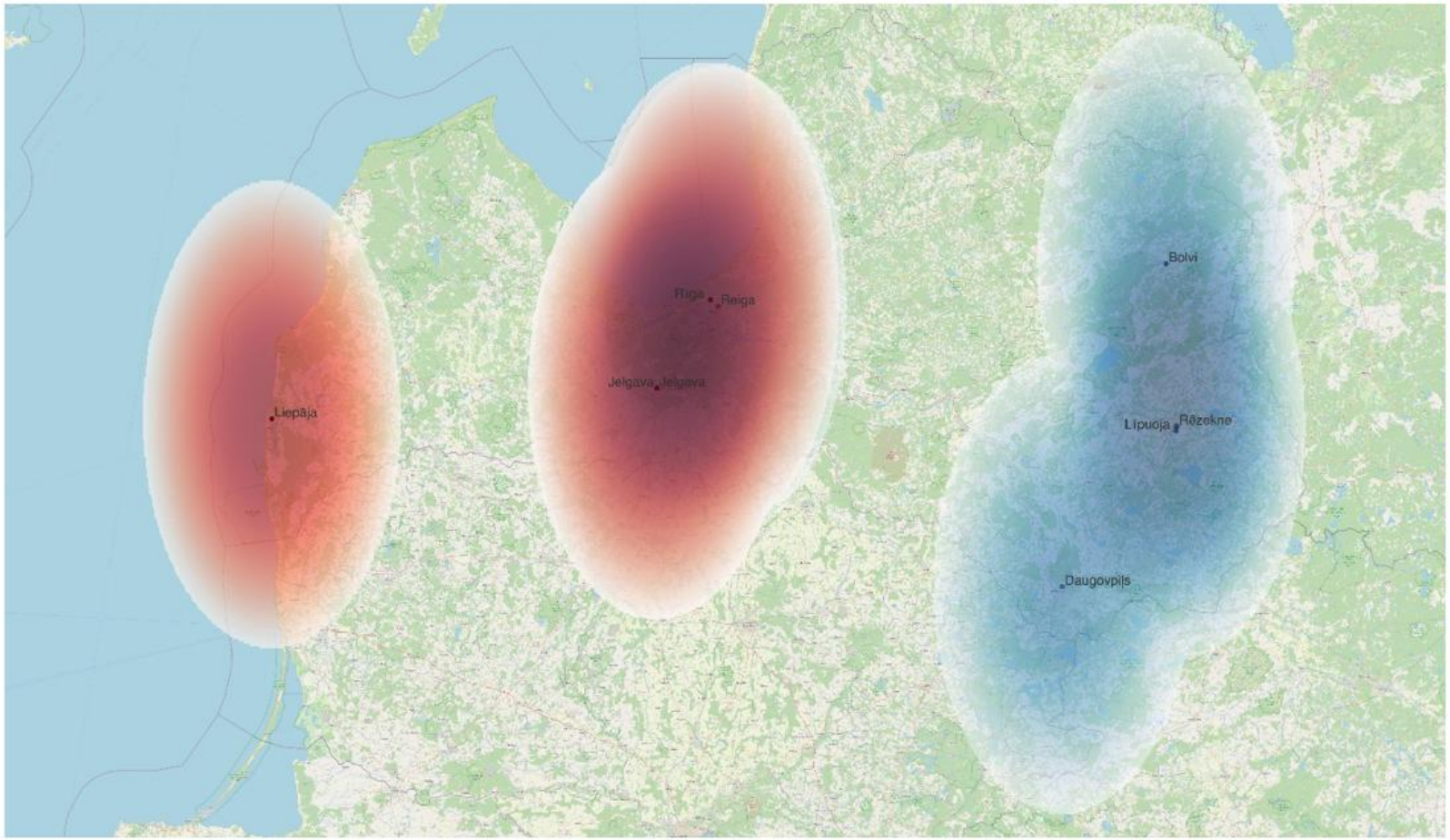


Figure 2: Collocations of *Riga* (LVK, red) and *Reiga* (MuLa, blue) with other toponyms

or highways. Figure 2 displays the collocations of *Rīga* / *Reiga* with other toponyms in LVK (red) and MuLa (blue). As is clearly visible, the Latgalian corpus (blue) relates *Reiga* mostly to places in Latgale, whereas collocations of *Rīga* in the Latvian corpus (red) are oriented towards central and Western Latvia. In the area around Riga, blue and red areas overlap.

# Discourse Prosodies (MuLa2022, LVK2022)

## (A.Kļavinska, S.Martena, H.Martens, A. Āboliņa, 2025)

| | MuLa2022 | | | LVK2022 | | |
|---|---|---|---|---|---|---|
| | Daugovpiļs | Rēzekne | Reiga | Daugavpils | Rēzekne | Rīga |
| Places, Directions, Mobility | 40,01% | 46,99% | 49,01% | 48,48% | 58,13% | 35,56% |
| Other | 14,06% | 10,48% | 15,55% | 14,54% | 9,02% | 8,62% |
| Education | 12,11% | 13,32% | 4,39% | 4,08% | 4,80% | 3,93% |
| Time | 9,08% | 10,28% | 11,65% | 4,38% | 5,49% | 17,86% |
| Culture and Arts | 8,94% | 5,37% | 4,54% | 3,43% | 1,70% | 2,92% |
| People | 5,26% | 4,48% | 4,54% | 1,54% | 1,58% | 3,81% |
| Religion | 2,74% | 2,70% | 1,88% | 0,00% | 0,00% | 0,31% |
| Economics and Business | 2,67% | 0,69% | 2,81% | 2,41% | 3,41% | 3,35% |
| Administration | 2,45% | 1,66% | 0,00% | 15,46% | 12,05% | 6,60% |
| Institutions and Organizations | 1,73% | 2,67% | 4,10% | 4,17% | 2,91% | 11,68% |
| Languages and Speaking | 0,65% | 0,00% | 0,74% | 0,34% | 0,00% | 0,00% |
| Sport | 0,29% | 0,00% | 0,00% | 0,77% | 0,31% | 4,13% |
| Politics | 0,00% | 1,38% | 0,79% | 0,41% | 0,59% | 1,23% |

# Use in the context: MuLa2022 Corpus

Excerpt 6: *Daudzi **latgalīši** **Reigā** pruovej puorvarēt gravitacejis spāku i lidoj tik augši, augši, ka aizmierst par savom saknem. (MuLa 2022, doc 1424)*

*Many Latgalians in Riga may overcome the force of gravity and fly so high, high that they can forget their roots.*

Excerpt 7: *i gondreiž vysod bejuši lapni, par tū sovā vydā ari Reigā **runoj latgaliski** i gondreiž vysod ir bejuši lapni par tū, ka īt nu Latgolys piecbarikažu laikā izauguse jauna audze, kura na tik da Krystapila, a I Reigys vydā **runoj latgaliski*** [MuLa2022]

*And almost always* [they] *were proud to speak Latgalian among themselves even in Riga and almost always were proud that also the generation that had grown up in the post-barricade times in Latgale speaks Latgalian not only until Krustpils but also in the middle of Riga.*

# Conclusions (1)

Differences between the Latvian and Latgalian corpora mostly indicate the regional orientation of users of Latgalian. Riga is an important centre for both users of Latvian and of Latgalian, but there is a considerably stronger Latgalian component in *MuLa2022* than in *LVK2022* – and other regions of Latvia are much less relevant for Latgalians. Latvian texts conceptualise cities through the national lens: They use standardised administrative terminology and frame urban centres in terms of institutional, infrastructural and statistical relations. In contrast, Latgalian texts embed cities within regional identity, preserving historical toponyms and foregrounding regional networks. Thus, Latvian discourse prioritises centralisation and governance, whereas Latgalian discourse highlights belonging, cultural continuity and everyday regional perspectives. At the same time, the Latgalian corpus reflects, for instance, that moving away from Latgale to Riga can promote individual expressions of regional identity and stimulate Latgalians to participate in social groups.

# Conclusions (2)

Our data show that users of Latgalian indeed display different mental representations of the cities than the users of Standard Latvian. These are connected to different places, important persons, cultural and social institutions, and life realities. The use of toponyms thereby indicates different realities and identities of a regional language and social structures of a peripheral region. This becomes particularly apparent when conducting a deeper qualitative analysis of the collocations.

# Conclusions (3)

In this sense, the analysis of urban toponyms in Latvian and Latgalian text corpora shows that linguistic communities‘ associations are closely tied to identity and ideology. Using CADS methods (an analysis of collocations, concordance, and discourse prosody), we identified distinct evaluative patterns and discursive framings in both corpora. Combined with a socio-onomastic interpretation and a geo-spatial visualisation through heat maps, our data show how linguistic choices are clustered geographically and reflect regional identities. This integrated approach demonstrates that city names function as discursive markers of belonging and cultural positioning, and thereby offer a powerful framework for studying language, space, and ideology.