The Changing Slovenian Smellscape, 1870–1919: Measuring Olfactory Discourse

Inna Novalija
inna.koval@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Janez Brank
janez.brank@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Marko Grobelnik
marko.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Counts of smell references in a historical corpus rise when the corpus itself grows, so raw frequencies cannot tell whether a culture talked about smell more often. We re-examine 366,333 passages of Slovene print from 1870–1919, previously processed with a curated smell lexicon and an LLM reader, and correct three biases without any manual annotation: exposure, word sense, and hallucination. Sense confirmation (the rate at which the LLM confirms a lexicon match in the same passage) separates olfactory uses of polysemous entries from non-olfactory ones and recovers explicit Slovene olfactory terms unaided (AUC 0.92, split-half $r = 0.98$). It shows that 52% of all keyword matches belong to term types the LLM corroborates in under 2% of their occurrences, dominated by polysemy such as *pot* ('sweat' / 'path'). After correction the apparent growth of olfactory discourse disappears: raw counts rise 5.5-fold while the corpus grows 6.5-fold, and the exposure-corrected rate is flat ($\tau = -0.07$, $p = 0.46$). What changes is composition. Food and drink rise from 19% to 32% of mentions and industry from 12% to 18%, while nature falls from 41% to 24%. Contrary to the deodorisation thesis, smell talk becomes *more* pleasant, and the drift is carried by a doubling in the share of purchasable scented goods — a commercialisation of the printed smellscape rather than a retreat from odour.

Keywords

Digital Humanities, Cultural Heritage, LLM, Text Mining, Sensory History

1 Introduction

Olfactory history has an evidence problem. Smell leaves no artefact, so its past is reconstructed almost entirely from what people wrote about it [4], and large digitised newspaper and book collections are now the obvious place to look. But a digitised collection is not a sample of a culture; it is a sample of what was printed, survived, was scanned, and passed OCR. All four grow steeply across the late nineteenth century. A count of smell references that rises with time may therefore be reporting nothing about smell.

Our earlier study [1] detected olfactory references in a subset of the Slovenian Digital Library (dLib.si) covering 1870–1919, using two independent methods: string matching against an expert-curated index of Slovene smell expressions [11], and passage-level extraction with an instruction-tuned LLM. It reported that mentions rise toward the turn of the century and that the two methods agree on only a minority of detections. Both findings were left as observations: the rise was measured in raw counts, and the disagreement was never traced to a cause.

This paper turns those observations into measurements, under a constraint we consider realistic for most heritage projects: **no manual annotation of any kind**. We replace human validation with three automatic corrections and validate the corrections themselves automatically. Our contributions are:

(1) **Sense confirmation**, a label-free method for deciding which entries of a smell-source lexicon are actually being used olfactorily, validated by known-item recovery and split-half reliability;
(2) **an exposure-corrected measurement** of Slovene olfactory discourse showing that its apparent nineteenth-century growth is an artefact of corpus growth;
(3) **a diachronic account of what did change** — the composition and hedonic balance of the printed smellscape — including a result that runs against the standard deodorisation narrative.

2 Related Work

Computational work on historical olfaction is anchored by Odeuropa [9], which produced the European Olfactory Knowledge Graph [6] and an annotation scheme for olfactory events, sources and qualities [7]. That line of work builds gold-standard annotations; we ask what can be established when annotation is unavailable. For Slovene, monolingual resources such as SloBERTa [12] and reference corpora such as Gigafida [5] and JANES [3] support modern text, but not the OCR-degraded, pre-standardised orthography of the period studied here.

Our framing follows the historiography of smell rather than of NLP. Corbin's account of nineteenth-century deodorisation [2] predicts a growing intolerance of odour, and Jenner [4] cautions against reading such narratives off unexamined sources. The corpus-composition problem we address is well documented in culturomics: inferences drawn from raw frequencies in Google Books are strongly confounded by what the corpus contains in each period [10, 8]. To our knowledge this correction has not been applied to sensory mining, and the present study is the first exposure-corrected diachronic measurement of olfactory discourse in Slovene.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.94

3 Data

We use the stored detection output of [1]: for every passage, the list of smell expressions returned by the LLM and the list of matches from the expert keyword index. The archive available to us covers 16,514 documents, 366,333 passages and 109.1 million words published between 1870 and 1919.[1] The keyword method fires in 53.7% of passages and the LLM in 36.4%; 0.68% of passages returned malformed LLM output and are counted as empty.

The exposure problem is visible in the data itself: the corpus grows from 6.3 million words in the 1870s to 41.0 million in the 1910s, a factor of 6.5.

4 Method

All three corrections operate on the existing detections; nothing is re-run against the LLM and nothing is labelled by hand.

4.1 Exposure normalisation

Every quantity we report is a rate per million words with exact Poisson intervals, never a count. Trends are tested with the Mann–Kendall statistic, which assumes neither linearity nor normality. As a negative control we shuffle the year label across passages 200 times and recompute the trend, giving an empirical null for $|\tau|$.

4.2 Sense confirmation

The expert index lists smell *sources* — sweat, earth, wine, incense — and string matching cannot tell whether a given occurrence concerns that thing's smell. In Slovene the problem is acute: *pot* is both 'sweat' and 'path', *duh* both 'odour' and 'spirit', *mira* both 'myrrh' and 'peace'. We therefore ask a second, independent reader. For each term type t we compute

$$c(t) = \frac{|\{\text{passages where the LLM also returns } t\}|}{|\{\text{passages where the keyword index matches } t\}|},$$

after folding diacritics and conservative suffix stripping so that inflected forms of one lemma count as one type, and counting each (passage, type) pair once. Types with at least 20 matches and $c(t) \geq 0.10$ are treated as olfactory; the rest are excluded. The threshold is deliberately low: it is meant to remove types the second reader essentially never corroborates, not to rank the survivors.

Because $c(t)$ is defined without labels, it can be validated without labels. We check (i) *known-item recovery*: whether it ranks a list of explicit Slovene olfactory terms (*vonj*, *smrad*, *dišava*, *zaudarjati*, …), written independently of the score and never shown to it, above the rest of the index; and (ii) *split-half reliability*: whether the score computed on two disjoint random halves of the documents agrees.

4.3 Groundedness

An LLM term that does not occur in its own passage is either an inference or a fabrication. We test each returned term for literal presence after folding and drop those that fail, which bounds the contribution of hallucination without inspecting a single passage by hand.

4.4 Typology

Term types are assigned a domain (food, nature, body, medical, industrial, domestic, ritual), a hedonic value and an explicit/source distinction by a deterministic Slovene lexicon built over the most frequent extracted types. The lexicon labels 50.0% of mention tokens; unlabelled tokens are excluded from composition analyses rather than imputed.

[1]The passages are the same units the 2025 study queried; the archive is a subset of that run, preserving its per-year structure.

Table 1: Automatic validation. No item requires human labels.

Check	Result
Known-item recovery (AUC, explicit terms)	0.92
mean $c(t)$, explicit terms	0.75
mean $c(t)$, other index entries	0.15
Split-half reliability of $c(t)$ (Pearson / Spearman)	0.98 / 0.92
LLM terms present in their own passage	94.8%
Malformed LLM responses	0.68%
Negative control, permuted years ($\lvert\tau\rvert$ at 95th pct.)	0.21
Typology vs. expert domain labels (token-weighted)	77.1%

5 Results

5.1 The corrections

Sense confirmation separates the index sharply. Of 263 term types with at least 20 matches, 119 clear the threshold but carry only 19.2% of keyword matches; the 144 that fail carry 80.5%. Types corroborated in under 2% of their occurrences account for 51.7% of *all* keyword matches in the corpus. The largest are transparently polysemous: *pot* (50,862 matches, $c = 0.001$), *zemlja* 'earth' (25,300, $c = 0.007$), *spomin* 'memory' (14,825, $c = 0.004$) and *svetnik* 'saint' (10,542, $c = 0.0006$), against *smrad* 'stench' ($c = 0.99$) and *vonj* 'scent' ($c = 0.98$). The score is not merely following volume: it recovers the independently written list of explicit olfactory terms with AUC 0.92, and reproduces itself across disjoint halves of the corpus at $r = 0.98$ (Table 1).

Groundedness is high: 94.8% of LLM terms occur literally in their own passage, falling only from 95.4% to 92.8% between the cleanest and noisiest OCR decile. Retaining grounded LLM terms and sense-confirmed keyword matches leaves 355,124 of 723,403 mentions (49.1%).

Two negative findings are worth recording. First, normalising inflection explains almost none of the method disagreement reported in [1]: term-level agreement rises only from 6.9% to 9.0% of detected passages, so 2.2% of the apparent disagreement is a string artefact and the rest is real. Second, the LLM shows no OCR advantage — its passage hit rate falls 42% from the cleanest to the noisiest OCR decile, against 46% for string matching.

5.2 Volume: the rise is the archive

Raw counts reproduce the 2025 result and rise strongly ($\tau = 0.66$, $p < 10^{-10}$), growing 5.5-fold while the corpus grows 6.5-fold. Once expressed per million words the trend vanishes and if anything reverses: $\tau = -0.21$ ($p = 0.03$) over all detections and $\tau = -0.07$ ($p = 0.46$) over the sense-confirmed subset (Fig. 1). The permutation control puts the 95th percentile of $|\tau|$ under shuffled years at 0.21, so even the weak negative trend is at the edge of what year-shuffling produces by chance. The honest statement is that the density of olfactory discourse in Slovene print is *flat* across the fifty years. OCR does not rescue the rise: the junk-token rate has no trend ($p = 0.21$), and conditioning the rate on it leaves the year–rate partial correlation at −0.30.

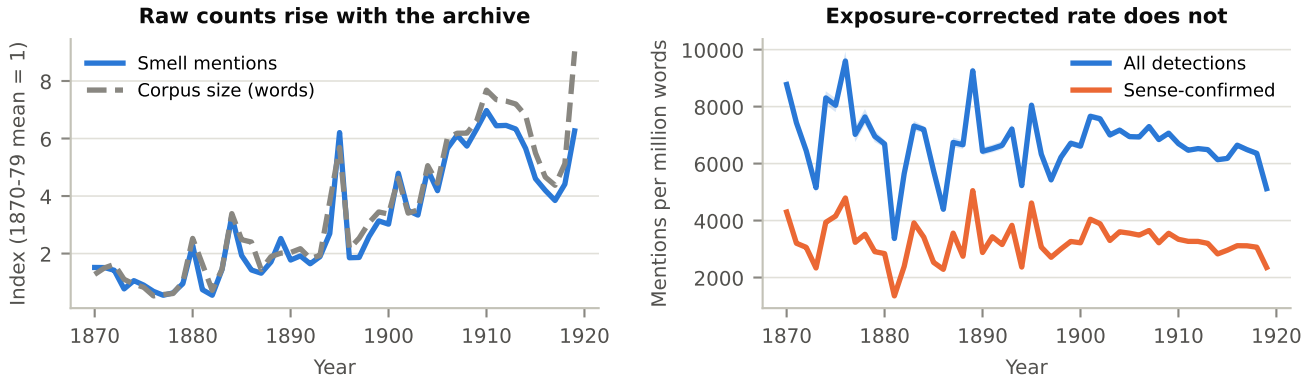

Figure 1: The apparent growth of olfactory discourse is corpus growth. Left: smell mentions and corpus size, each indexed to its 1870–79 mean, move together. Right: expressed as a rate per million words, neither all detections nor the sense-confirmed subset increases.

Table 2: Domain rates per million words, 1870s vs. 1910s, on the sense-confirmed set. Mann–Kendall over all 50 years.

Domain	1870s	1910s	×	p
Food & drink	177.8	326.3	1.84	$< 10^{-5}$
Medicine & hygiene	30.7	50.9	1.66	0.025
Industry & fuel	120.1	183.2	1.53	$< 10^{-5}$
Household	46.7	53.6	1.15	0.027
Ritual	31.0	33.2	1.07	0.99
Body & decay	146.2	115.7	0.79	0.12
Nature & plants	413.2	249.8	0.60	0.040

5.3 Composition: what did change

The mix of smells changes decisively even as the total does not (Fig. 2; $\chi^2(24) = 2023$, $p < 10^{-300}$, Cramér's $V = 0.07$, $n = 110{,}593$). Between the 1870s and the 1910s, food and drink rise from 18.8% to 31.9% of confirmed mentions and industry and fuel from 12.2% to 18.3%, while nature and plants fall from 41.0% to 24.4%. Expressed as rates per million words, food rises 1.8-fold ($\tau = 0.46$, $p < 10^{-5}$) and industry 1.5-fold ($\tau = 0.44$, $p < 10^{-5}$), while nature falls to 0.60 of its initial rate ($\tau = -0.20$, $p = 0.04$) (Table 2). The smellscape of Slovene print moves indoors and into the market, and away from field and forest.

An independent expert labelling of the curated index into these same seven domains (662 terms) reproduces the composition shift and the food, industry and nature results, and agrees with our lexicon on 77.1% of mentions (Table 1). It does not reproduce the medical rise or the flat household row, so we draw no conclusion about either.

5.4 Hedonics: no deodorisation

The deodorisation thesis predicts that smell talk grows more negative. We find the opposite. The share of hedonically marked mentions that are unpleasant falls from 39.3% in the 1870s to 28.5% in the 1910s (52,062 marked mentions; logistic odds ratio 0.93 per decade, $z = -8.2$, $p < 10^{-15}$; Fig. 3).

The drift has a traceable source. A closed list of purchasable scented goods — cigars, spirits, beer, wine, coffee, soap, perfume, ointments, lamp oil — written down before the hedonic result was computed, rises 1.7-fold ($\tau = 0.32$, $p = 0.001$) while everything else falls slightly and non-significantly (0.83-fold, $p = 0.20$). Its share of all confirmed mentions nearly doubles, from 3.7% to 7.1%

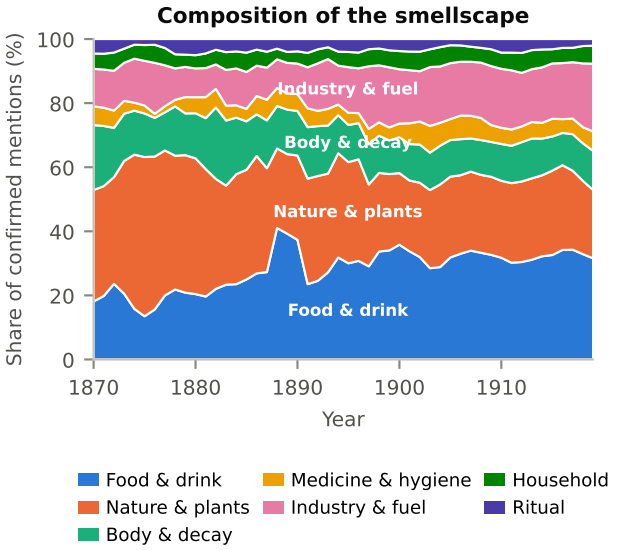

Figure 2: Composition of the sense-confirmed smellscape, three-year rolling mean. The total is flat; the mix is not.

($p < 10^{-4}$). These goods are 20.8% pleasant and 0% unpleasant in our lexicon, against 9.9% and 4.3% for everything else. What our data record is not a culture retreating from odour but a printed public sphere in which an increasing share of smell language is advertising for things that can be bought.

Consistent with this, only 4.1% of confirmed mentions use an explicit olfactory head from our list, and that share declines (odds ratio 0.97 per decade, $p < 10^{-5}$): the corpus talks about smell mostly by naming things that smell. The two methods differ in which things. The keyword index is weighted toward nature (33.4%) and the body (16.9%); the LLM toward food and drink (37.4%) and the household (8.5% vs. 0.5%), and 85.0% of its mentions use vocabulary absent from the expert index entirely.

5.5 Events

We probed three periods with windows fixed from the historical record before the rates were computed. Cholera (1886–87) and the 1895 Ljubljana earthquake give no elevation; the First World War does, with industrial smells at 1.33 and food at 1.30 times

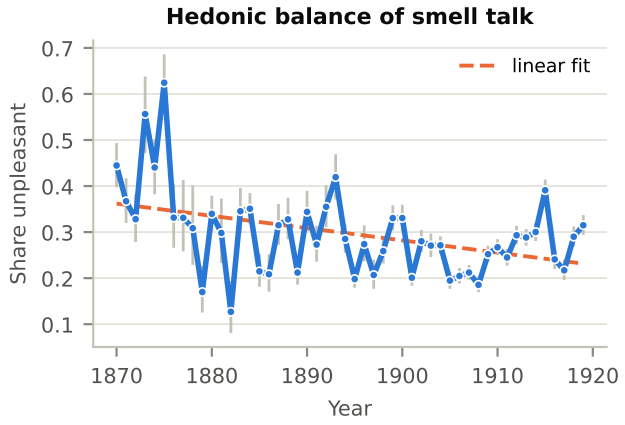

Figure 3: Share of hedonically marked mentions that are unpleasant, with Wilson intervals. The trend is downward, against the deodorisation prediction.

their out-of-window rate ($p = 0.035$ each, Mann–Whitney). Two of three probes are null, which we report as a check on ourselves.

6 Discussion and Limitations

The methodological lesson generalises beyond smell. A curated lexicon of *sources* for any sensory or affective phenomenon will be dominated, in running text, by non-target senses of its most frequent entries; in our corpus half of all matches are such. Where no annotation budget exists, a second detector of a different kind is a usable substitute, not because agreement implies truth, but because systematic non-agreement on a term type is strong evidence of polysemy.

That is also the main limitation. Sense confirmation measures agreement between two detectors, not correctness, and cannot detect an error they share. It is biased toward what the LLM finds legible, and it is a filter, not a precision estimate. We therefore make no absolute claims about detection quality; every claim above is relative or diachronic, where a bias that is constant over time cannot manufacture a change. Valence and domain are assigned at the type level, so an ironic or hostile passage about a perfume still counts as pleasant; the hedonic result is a statement about which vocabulary appears, not about sentiment in context. The available metadata give year but not genre or place of publication, so we cannot separate a change in olfactory culture from a change in the genre mix of what was printed, it is likely that some of both is present. Finally, the corrections use no labelled passages, but domain and valence rest on a word list: ours labels half of all mention tokens, and the expert labelling covers 56.6% of confirmed keyword mentions against only 11.2% of grounded LLM ones, most of which lie outside the curated index.

7 Conclusion

Re-examining 109 million words of Slovene print, we find that the growth of olfactory discourse reported for this period is an artefact of a growing archive: the exposure-corrected rate is flat. What changed is the composition of the smellscape: toward food and industry, away from nature; and its hedonic balance, which becomes more positive rather than more negative, carried by a doubling of the share of purchasable scented goods. Along the way, sense confirmation shows that half of the matches produced by an expert smell lexicon are not about smell at all. We release the per-term table (volume by method, confirmation rate, domain, valence) as an automatically sense-filtered vocabulary of historical Slovene olfactory language, with the expert domain labelling and the pipeline. Future work should recover genre and place of publication from dLib metadata.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency via project J7-50233. The smell index is due to Mojca Ramšak [11]. The authors acknowledge the use of large language models for language optimisation. While these tools contributed to re- fining the presentation, all conceptual frameworks, analyses and interpretations remain the sole responsibility of the authors.

References

[1] Janez Brank, Inna Novalija, Dunja Mladenić, and Marko Grobelnik. 2025. LLM based approach to extracting smells in Slovenian corpora. In *Proceedings of the 28th International Multiconference Information Society (SiKDD)*. Ljubljana, Slovenia. doi:10.70314/is.2025.sikdd.5.

[2] Alain Corbin. 1986. *The Foul and the Fragrant: Odor and the French Social Imagination*. Originally published as *Le miasme et la jonquille*, 1982. Harvard University Press, Cambridge, MA.

[3] Darja Fišer, Nikola Ljubešić, and Tomaž Erjavec. 2020. The JANES project: language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54, 1, 223–246.

[4] Mark S. R. Jenner. 2011. Follow your nose? smell, smelling, and their histories. *The American Historical Review*, 116, 2, 335–351.

[5] Simon Krek, Špela Arhar Holdt, Tomaž Erjavec, Jaka Čibej, Andraz Repar, Polona Gantar, Nikola Ljubešić, Iztok Kosem, and Kaja Dobrovoljc. 2020. Gigafida 2.0: the reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*. Marseille, France, 3340–3345.

[6] Pasquale Lisena, Thibault Ehrhart, and Raphaël Troncy. 2024. European olfactory knowledge graph. Zenodo. (2024). doi:10.5281/zenodo.10709703.

[7] Stefano Menini, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2023. Scent mining: extracting olfactory events, smell sources and qualities. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, Dubrovnik, Croatia, 135–140. doi:10.18653/v1/2023.latechclfl-1.15.

[8] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331, 6014, 176–182. doi:10.1126/science.1199644.

[9] ODEUROPA Project Consortium. 2021–2023. ODEUROPA: negotiating olfactory and sensory experiences in cultural heritage practice and research. https://odeuropa.eu/. EU Horizon 2020, GA No. 101004469. (2021–2023).

[10] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10, 10, e0137041. doi:10.1371/journal.pone.0137041.

[11] Mojca Ramšak. 2025. *Antropologija vonja*. AMEU-ISH, Ljubljana.

[12] Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of the 24th International Multiconference Information Society (SiKDD)*.