Smelling the Past: Investigating Historical Models for Olfactory Event Extraction

Teresa Paccosi Marijn Koolen
KNAW Humanities Cluster, DHLab, Amsterdam
teresa.paccosi@dh.huc.knaw.nl, marijn.koolen@di.huc.knaw.nl

Abstract

In this paper, we present a series of experiments using historical language models to investigate the impact of pretraining on data that more closely resembles the task domain, focusing on the case study of automatic olfactory event extraction. We tested five monolingual historical and contemporary pretrained models on the task of extracting olfactory events using a benchmark spanning several centuries. The languages investigated are English, French, Italian, German, and Dutch. The aim of our research is not only to assess whether historical models can improve performance on this diachronically oriented task, but also to gain deeper insight into the factors influencing model performance through a detailed analysis of performance patterns. We examine potential sources of variation and previously proposed hypotheses to account for lower performance observed in this task, thereby offering a more comprehensive understanding of model behavior in this context.

Keywords: historical smell, historical models, model evaluation

1. Introduction

Within the traditional hierarchy of the five senses, olfaction has long been relegated to a marginal role, particularly in Western societies where vision has held a dominant position (Winter et al., 2018). This neglect, coupled with the widespread assumption that smell lacks a well-defined semantic field (Sperber, 1975), has contributed to the limited number of studies devoted to this sensory domain. Only in recent years olfaction has begun to attract renewed scholarly attention (Howes, 2006), and its cultural significance has started to be acknowledged, including its value as a historical example of intangible cultural heritage (Tullett et al., 2022). In the fields of Natural Language Processing (NLP) and Digital Humanities (DH), engagement with the domain of olfaction is likewise relatively recent. Precisely because smell has received comparatively little systematic attention, it constitutes a promising area for computational investigation. Nevertheless, the structured analysis of olfactory terminology and the automatic extraction of smell-related information remain underexplored within NLP research, especially from a diachronic point of view. This paper presents an experimental investigation of language models pre-trained on historical texts, applied to the specific case study of olfactory language. Previous work on this case study had relied solely on models trained on contemporary corpora. Using a similar setting to the one presented in Menini (2024), we compare models pre-trained on contemporary texts against those pre-trained on historical texts. We combine multiple evaluation perspectives aiming at assessing the impact of using historical models for this task, through (a) standard metrics such as precision, recall, and F_1 score, (b) a comparison of strict versus lenient F_1 scores to account for span boundary choices (previously identified as major source of problems for some entity types), and (c) detailed error analysis including overlap assessment and manual inspection. By situating our work within the broader line of research on historical model performance (Plank, 2016; Ehrmann et al., 2023), especially the influence of pre-training data similarity to target tasks (Gururangan et al., 2020; Gladstone et al., 2025), we focus on a specific case study that has not previously been explored from this perspective.

Our contributions are twofold: (a) the release of fine-tuned historical models for extracting olfactory events, and (b) a systematic analysis highlighting concrete factors that can lead to performance improvements.[1]

2. Related Work

Most research on sensory language has primarily focused on developing structured resources to model perceptual domains. In this vein, Tekiroğlu et al. (2014) introduce *Sensicon*, a large-scale sensory lexicon that automatically associates English words with the five senses, while other studies target domain-specific descriptive vocabularies, such as those used by whiskey and wine reviewers, as in Hamilton et al. (2023) and Lefever et al. (2018). Focusing more specifically on smell-related events, Brate et al. (2020) introduced an annotation framework and semi-supervised methods for capturing olfactory experiences, while McGregor and McGillivray (2018) employed distributional

[1] Code and data are available at this GitHub repository.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 389–399.

semantics to isolate smell descriptions. More recently, Tonelli and Menini (2021) built a FrameNet-inspired annotation scheme and a benchmark for olfactory events in historical texts (Menini et al., 2022), which Menini (2024) employed to conduct experiments on the same task we are investigating in this paper. The author used BERT models pretrained on contemporary language, which have also previously shown strong performance on the extraction of sensory-related information (Stojanov et al., 2021). However, as computational methods are increasingly applied in historical research, it becomes clear that the use of appropriately designed systems and models is essential. Simply increasing the size of a language model or modifying its architecture does not guarantee improved performance (Alajrami and Aletras, 2022; Sun et al., 2022). Instead, the effectiveness depends heavily on the alignment of data distributions between pretraining and fine-tuning phases (Li et al., 2022). Supporting this point, Verkijk et al. (2025) show that a model pretrained exclusively on in-domain historical data outperforms models trained on much larger amounts of data of a more diverse nature.

3. Extracting Olfactory Events

In this section, we outline the theoretical background of our study, focusing on the principles of Frame Semantics on which the olfactory event extraction task is based. We briefly discuss how Frame Semantics is applied to model olfactory events in text, and then introduce the models used in this paper along with the details of our experimental setup.

3.1. Theoretical Background

Our work builds on the olfactory benchmark presented in Menini et al. (2022)[2] and the work of Menini (2024), who finetuned various BERT models on this benchmark for the task of extracting smell-related concepts, specifically targeting the olfactory frame. The original benchmark covers six languages, English, French, Italian, German, Dutch, and Slovenian. We excluded Slovenian, as we lack the necessary linguistic expertise to perform a reliable error analysis for this language, and, to our knowledge, no monolingual historical model is currently available for it. The benchmark is based on Frame Semantics theory (Fillmore et al., 2006), modeling olfactory events as a distinct frame with its Lexical Units (LUs) and Frame Elements (FEs). LUs are words that evoke the olfactory frame, defined as "pairing[s] of a word with a meaning" (Ruppenhofer et al., 2016), and in this case include terms describing odours, such as *smell*, *perfume*, or *stench* (referred to here as *Smell_Word*). Each frame also defines a set of FEs, representing the semantic roles or participants associated with the LU. For example, in "The cook carefully smelled the soup", *smell* is the LU, while *the cook* and *the soup* serve as the *Perceiver* and *Smell_Source* FEs, respectively. In Table **??**, we report a list of all the FEs in the benchmark, along with brief descriptions, which constitute the set of tags targeted in the olfactory frame extraction task.

3.2. Models

Since our experiments focus mainly on historical language models, and to our knowledge, the only existing multilingual historical model, hmBERT (Schweter et al., 2022), covers German, English, French, Swedish, and Finnish, but not Italian or Dutch, we opted for a strictly monolingual setting for each language. In this setup, both the training data and the models are language-specific, and we include contemporary models for comparison alongside the historical ones.

The language-specific historical models we use for the olfactory extraction task are the following:

- **English**: MacBERTh[3] , (Manjavacas and Fonteyn, 2021)
- **Italian**: BERToldo[4] (Aprosio et al., 2022),
- **French**: D'AlemBERT[5] (Gabay et al., 2022),
- **Dutch**: GysBERT[6] (Manjavacas Arevalo and Fonteyn, 2022),
- **German**: Historical German BERT[7]

The language-specific contemporary models we used are instead:

- **English**: bert-base-cased[8] (Devlin et al., 2019),
- **Italian**: bert-base-italian-cased[9] (Schweter, 2020),

[2]Detailed information about the benchmark is provided in Menini et al. (2022) and the corresponding github repository: https://github.com/Odeuropa/benchmarks_and_corpora.

[3]https://huggingface.co/emanjavacas/MacBERTh

[4]https://github.com/dhfbk/historical-bert?tab=readme-ov-file

[5]https://huggingface.co/pjox/dalembert

[6]https://huggingface.co/emanjavacas/GysBERT

[7]https://huggingface.co/redewiedergabe/bert-base-historical-german-rw-cased

[8]https://huggingface.co/google-bert/bert-base-cased

[9]https://huggingface.co/dbmdz/bert-base-italian-cased

Frame Element	Example Sentence
Smell Source	The person, object or place that has a specific smell. *The odour [of tar] and [pitch] was so strong.*
Odour Carrier	The carrier of an odour, either an object (e.g. handkerchief) or atmospheric elements (wind, air) *The unpleasant smell [of the vapour] of linseed oil extended for a considerable distance.*
Quality	A quality associated with a smell and used to describe it. *Earth has a [strong], [aromatic] odour.*
Perceiver	The being that perceives an odour, who has a perceptual experience, not necessarily on purpose. *The scent is described by [Dr. Muller] as delicious.*
Evoked Odorant	The object, place or similar that is evoked by the odour, even if it is not in the scene. *In offensive perspiration of the feet [a peculiar cabbage-like] stench is given off.*
Location	The location where the smell event takes place. *And, particularly, [at the foot of the garden], where he felt so very offensive a smell that has sickened him.*
Time	An expression describing when the smelling event occurred. *Galeopsis smells fetid [at first handling], [afterwards] aromatic.*
Circumstances	The state of the world under which the smell event takes place. *[When stale] the lobster has a rank stench.*
Effect	An effect or reaction caused by the smell. *An ill smell [gives a nauseousness].*

Table 1: Overview of the Frame Elements (FEs) as reported in (Menini et al., 2022). Lexical Units (i.e., *Smell_Words* are underlined and the FE of interest is in square brackets. The same definitions hold for all languages included in the benchmark.

- **Dutch**: BERTje[10] (De Vries et al., 2019),
- **French**: FlauBERT[11] (Le et al., 2020),
- **German**: bert-base-german-cased[12] (Chan et al., 2020).

For the contemporary setting, we re-ran all experiments using the same models investigated in Menini (2024) to create a fully controlled experimental setup with consistent test data and predictions across languages. Beyond simply reproducing previous results, this setup enabled us to explore more deeply the factors underlying performance differences. Specifically, we examined how predictions overlap with gold annotations, analysed the nature of errors made by different models in span selection, and compared the behaviour of their tokenisers, with the goal of gaining a deeper understanding of how historical data influences model performance.

3.3. Multitask Approach for Olfactory Frame Extraction

Given that previous results on this task showed that the multitask setting significantly outperforms the single-task setting, we decided to focus our analysis on the former. In this setting, we employ multitask learning (Caruana, 1997), an approach in which the model learns multiple tasks simultaneously while sharing a common representation, enabling knowledge transfer across tasks. We model the classification of the LU and each FE class as separate tasks, following previous work on this task that suggested that detecting LUs, typically expressed as single tokens, can serve as an auxiliary task, supporting the identification of more complex FEs through shared representations. For this setup, we use MaChAmp (Van Der Goot et al., 2021), a toolkit for fine-tuning contextualised embeddings in multitask NLP scenarios. MaChAmp initialises from a pre-trained encoder and fine-tunes it using an inverse square root learning rate schedule with linear warm-up (Howard and Ruder, 2018). Each task is assigned a dedicated decoder, and prediction is performed using Conditional Random Fields (Lafferty et al., 2001), which are well suited for sequence labeling. We configure MaChAmp to run 10 BIO-based sequence labeling tasks, each corresponding to a LU or FE. Evaluation is conducted using span-level strict and lenient precision (P), recall (R), and F_1. We experiment with the same setting in all languages using uniform weighting at 1 for all tasks, a learning rate of $1e-4$, a batch size of 32, and 30 epochs.

4. Results

In Table 3, we show the results of the experiments for both contemporary and historical models, in terms of strict and lenient precision (P), recall (R) and F_1. From a general perspective, it is noticeable that using historical models has a positive impact on F_1 scores for almost all labels in En-

[10] https://huggingface.co/GroNLP/bert-base-dutch-cased

[11] https://huggingface.co/flaubert/flaubert_base_cased

[12] https://huggingface.co/google-bert/bert-base-german-cased

glish and Dutch, and to a lesser extent in German, whereas for Italian and French contemporary models consistently perform better. A more detailed discussion of these findings is provided in Section 5.1. Examining the results more closely, we observe some cross-linguistic and cross-label patterns. Although the majority of labels do not achieve particularly high scores with either contemporary or historical models, certain labels consistently perform well: notably, *Smell_Word*, and *Quality*, and, compared to the other labels, also *Smell_Source*, and *Evoked_Odorant*. The strong performance of *Smell_Word* is not surprising, as it often corresponds to lexically consistent single words, making its identification relatively straightforward. The other labels, however, probably benefit from regular syntactic realisations: *Quality* frequently appears as an adjective preceding the *Smell_Word*, *Smell_Source* often functions as a direct object, and *Evoked_Odorant* typically occurs in predictable prepositional constructions such as "like" or "as". Still, both *Smell_Source* and *Evoked_Odorant* present a variability of expression that likely do not help model generalisation. More complex or longer entities, such as *Circumstances*, and *Effect*, consistently show low performances and a lower recall, reflecting the difficulty of correctly identifying multi-token spans with variable realisation. Comparing strict and lenient evaluations, we observe notable differences for labels with longer spans, especially in English, German and Dutch. For example, in *Circumstances*, where recall shows notable improvement in languages such as English, German, and Italian, supporting the idea that incorrect span boundaries are one of the main sources of errors for this entity type. We provide a more detailed discussion of the differences between strict and lenient evaluation results in Section 5.3.

Model	tokens	data	coverage
BERToldo$_all$ (it)	∼10M	970 MB	1200-1900
D'AlemBERT (fr)	∼185M	–	1500-1790
MacBERTh (en)	∼3.9B	–	1450-1950
GysBERT (nl)	∼7.1B	–	1618-1999
Hist BERT (de)	–	–	1840-1920

Table 2: Size and temporal coverage of pretraining data for each historical model.

5. Discussion and Analysis

In this section, we provide a more detailed discussion of the results presented in Section 4. This includes an analysis of the difference in performance under lenient versus strict evaluation, a comparison of Subword Fertility Rates across models, an examination of the gain or loss in F1 when comparing historical and contemporary models, previously suggested as a possible factor influencing model performance (Rust et al., 2021), and an analysis of the most common errors made by the models in terms of span selection.

5.1. Contemporary Versus Historical Models

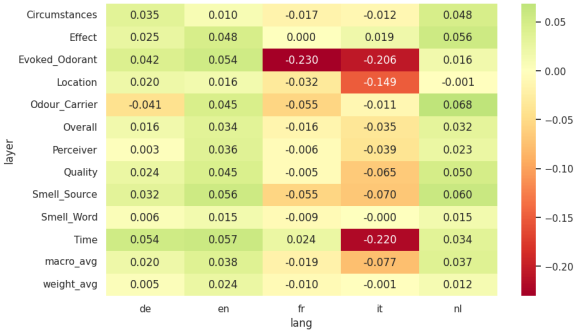


Figure 1: F1 differences (historical − contemporary) across layers and languages.

Comparing the potential gain from using historical models for this task, a clear pattern emerges: Italian and French achieve better performance with contemporary models, whereas English and Dutch, and, to a lesser extent, German, benefit more from the historical setting. These differences may be partly related to the size and temporal coverage of the pretraining data for each historical model, as summarised in Table 4. French and Italian historical models appear to perform worse, likely due to the comparatively limited historical data available (approximately 185M and 10M tokens, respectively), which may constrain their ability to capture complex or less frequent historical patterns. Moreover, the French historical model was primarily pretrained on 17th-century texts, whereas the majority of texts in the French benchmark come from the late 18th century. This suggests that performance may not only be influenced by the size of the pretraining data but also by the similarity of the pretraining domain, a factor previously observed in Dutch by Verkijk et al. (2025) as potentially impacting model effectiveness.

In contrast, the English and Dutch historical models achieve comparatively stronger results, plausibly benefiting from larger pretraining corpora spanning a broader temporal range, which may enable them to model both frequent and rare phenomena more effectively. German shows only a marginal improvement: its historical pretraining corpus spans 1840–1920, which is only slightly earlier than the texts used for the contemporary model. As a result, the model gains some benefit from exposure to older linguistic patterns, but the over-

		Strict						Lenient					
		Contemporary			Historical			Contemporary			Historical		
Entity type	Lang.	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Circumstances	de	0.240	0.112	0.142	**0.262**	**0.143**	**0.177**	0.532	0.243	0.312	**0.540**	**0.301**	**0.372**
	en	**0.363**	0.283	0.311	0.328	**0.318**	**0.321**	0.554	0.449	0.486	**0.572**	**0.562**	**0.564**
	fr	**0.040**	**0.057**	0.041	0.028	0.017	0.024	**0.494**	**0.428**	**0.368**	0.156	0.057	0.093
	it	0.203	**0.178**	**0.177**	**0.278**	0.123	0.164	0.505	**0.495**	**0.470**	**0.534**	0.222	0.297
	nl	0.196	0.076	0.107	**0.223**	**0.119**	**0.155**	0.314	0.122	0.173	**0.343**	**0.185**	**0.240**
Effect	de	0.207	0.066	0.100	**0.241**	**0.088**	**0.125**	**0.472**	0.151	0.228	0.452	**0.165**	**0.235**
	en	0.160	0.153	0.155	**0.212**	**0.207**	**0.204**	0.358	0.342	0.348	**0.382**	**0.345**	**0.352**
	fr	0.000	0.000	0.000	0.000	0.000	0.000	**0.250**	**0.200**	**0.250**	0.000	0.000	0.000
	it	0.075	0.040	0.060	**0.133**	**0.045**	**0.079**	**0.375**	**0.135**	**0.222**	0.287	0.089	0.159
	nl	0.221	0.097	0.122	**0.241**	**0.145**	**0.178**	**0.473**	0.209	0.272	0.468	**0.294**	**0.356**
Evoked Odorant	de	0.350	0.297	0.317	**0.401**	**0.330**	**0.359**	**0.706**	**0.607**	**0.644**	0.670	0.546	0.596
	en	0.326	0.343	0.312	**0.374**	**0.385**	**0.367**	0.594	**0.608**	0.574	**0.601**	0.596	**0.577**
	fr	**0.423**	**0.381**	**0.397**	0.375	0.086	0.167	**0.451**	**0.410**	**0.426**	0.375	0.086	0.167
	it	**0.356**	**0.353**	**0.329**	0.124	0.125	0.123	**0.501**	**0.552**	**0.483**	0.328	0.302	0.300
	nl	0.362	0.281	0.312	**0.337**	**0.323**	**0.328**	0.561	0.440	0.487	**0.562**	**0.540**	**0.548**
Location	de	0.228	0.089	0.126	**0.253**	**0.106**	**0.146**	**0.330**	0.132	0.185	0.323	**0.140**	**0.192**
	en	0.424	0.424	0.423	**0.432**	**0.462**	**0.439**	**0.569**	0.568	**0.567**	0.542	**0.588**	0.555
	fr	0.206	**0.166**	**0.174**	**0.225**	0.106	0.143	0.436	**0.317**	**0.338**	**0.504**	0.238	0.321
	it	**0.312**	**0.178**	**0.219**	0.197	0.050	0.070	0.498	**0.300**	**0.365**	**0.500**	0.176	0.222
	nl	**0.350**	0.156	**0.214**	0.273	**0.177**	0.213	**0.517**	0.226	0.312	**0.436**	**0.285**	**0.343**
Odour Carrier	de	**0.313**	**0.130**	**0.160**	0.202	0.093	0.118	**0.313**	**0.130**	**0.160**	0.309	0.113	0.152
	en	0.355	0.320	0.329	**0.435**	**0.352**	**0.375**	0.468	0.423	0.435	**0.550**	**0.448**	**0.474**
	fr	0.412	**0.159**	**0.253**	**0.452**	0.130	0.197	**0.457**	**0.174**	**0.278**	0.452	0.130	0.197
	it	0.334	**0.151**	**0.194**	**0.374**	0.125	0.183	0.470	**0.215**	**0.273**	**0.512**	0.162	0.241
	nl	0.247	0.106	0.146	**0.319**	**0.163**	**0.215**	0.303	0.127	0.176	**0.381**	**0.193**	**0.254**
Perceiver	de	**0.393**	**0.236**	0.286	0.378	**0.236**	**0.289**	**0.436**	0.264	0.319	**0.436**	**0.273**	**0.334**
	en	0.395	0.438	0.412	**0.431**	**0.468**	**0.448**	**0.531**	**0.581**	**0.551**	0.529	0.576	0.550
	fr	0.157	**0.058**	**0.084**	**0.188**	0.050	0.078	**0.417**	**0.146**	**0.215**	0.216	0.060	0.092
	it	0.297	**0.204**	**0.231**	**0.427**	0.150	0.192	0.433	**0.301**	**0.338**	**0.474**	0.183	0.224
	nl	**0.377**	0.210	0.268	0.358	**0.248**	**0.291**	**0.426**	0.237	0.303	0.421	**0.291**	**0.342**
Quality	de	0.553	0.565	0.556	**0.596**	**0.569**	**0.579**	0.758	**0.773**	0.761	**0.794**	0.757	**0.771**
	en	0.674	0.696	0.684	**0.716**	**0.742**	**0.729**	0.824	0.852	0.837	**0.840**	**0.871**	**0.855**
	fr	0.390	**0.451**	**0.415**	**0.419**	0.403	0.410	**0.640**	**0.740**	**0.682**	0.601	0.578	0.589
	it	**0.732**	**0.739**	**0.734**	0.682	0.661	0.670	**0.807**	**0.815**	**0.810**	0.784	0.759	0.769
	nl	0.633	0.599	0.614	**0.671**	**0.661**	**0.664**	0.741	0.700	0.718	**0.757**	**0.749**	**0.751**
Smell Source	de	0.395	0.312	0.346	**0.415**	**0.352**	**0.378**	0.587	0.462	0.514	**0.597**	**0.508**	**0.545**
	en	0.475	0.456	0.465	**0.505**	**0.541**	**0.520**	**0.636**	0.610	0.621	0.631	**0.677**	**0.651**
	fr	**0.359**	**0.336**	**0.329**	0.300	0.262	0.274	**0.581**	**0.557**	**0.534**	0.446	0.392	0.408
	it	**0.400**	**0.418**	**0.405**	**0.400**	0.289	0.335	0.562	**0.588**	**0.569**	**0.588**	0.426	0.493
	nl	0.351	0.283	0.312	**0.401**	**0.348**	**0.372**	0.549	0.442	0.487	**0.571**	**0.497**	**0.530**
Smell Word	de	0.801	**0.808**	0.800	**0.816**	0.807	**0.806**	0.815	**0.822**	0.814	**0.831**	**0.822**	**0.821**
	en	0.889	0.855	0.871	**0.900**	**0.873**	**0.886**	0.898	0.864	0.880	**0.911**	**0.884**	**0.897**
	fr	**0.796**	0.817	**0.801**	0.780	**0.819**	0.792	**0.845**	**0.870**	**0.852**	0.802	0.841	0.814
	it	0.856	**0.905**	**0.877**	**0.897**	0.859	0.876	0.870	**0.920**	**0.891**	**0.910**	0.872	0.889
	nl	0.740	0.784	0.761	**0.746**	**0.808**	**0.776**	0.754	0.799	0.775	**0.763**	**0.826**	**0.793**
Time	de	0.234	0.087	0.122	**0.376**	**0.137**	**0.176**	0.420	0.172	0.232	**0.524**	**0.229**	**0.286**
	en	0.471	0.387	0.416	**0.521**	**0.441**	**0.473**	0.624	0.507	0.549	**0.633**	**0.531**	**0.573**
	fr	0.244	**0.086**	0.120	**0.452**	0.077	**0.144**	**0.646**	**0.286**	**0.347**	0.524	0.121	0.207
	it	**0.490**	**0.425**	**0.453**	0.262	0.213	0.233	**0.768**	**0.654**	**0.703**	0.700	0.581	0.629
	nl	**0.296**	0.140	0.181	0.291	**0.182**	**0.216**	**0.445**	0.240	**0.297**	0.346	0.227	0.265
macro avg.	de	0.371	0.270	0.295	**0.394**	**0.286**	**0.316**	0.537	0.376	0.417	**0.548**	**0.385**	**0.430**
	en	0.453	0.435	0.438	**0.485**	**0.479**	**0.476**	0.606	0.580	0.585	**0.619**	**0.608**	**0.605**
	fr	0.302	**0.251**	**0.263**	**0.344**	0.195	0.244	**0.529**	**0.413**	**0.432**	0.437	0.250	0.316
	it	**0.412**	**0.359**	**0.375**	0.384	0.264	0.297	**0.583**	**0.497**	**0.518**	0.570	0.377	0.428
	nl	0.377	0.273	0.304	**0.386**	**0.317**	**0.341**	**0.508**	0.354	0.400	0.505	**0.409**	**0.442**
weighted avg.	de	0.633	0.611	0.615	**0.644**	**0.608**	**0.620**	0.719	**0.674**	0.686	**0.731**	0.673	**0.693**
	en	0.681	0.667	0.672	**0.697**	**0.699**	**0.696**	0.756	0.742	0.746	**0.762**	**0.770**	**0.763**
	fr	0.606	**0.629**	**0.608**	**0.608**	0.609	0.598	**0.703**	**0.743**	**0.705**	0.653	0.637	0.634
	it	0.708	**0.735**	**0.717**	**0.747**	0.697	0.717	0.754	**0.778**	**0.759**	**0.790**	0.720	0.746
	nl	0.601	0.590	0.591	**0.604**	**0.608**	**0.603**	0.661	0.623	0.635	**0.664**	**0.651**	**0.653**

Table 3: Evaluation scores per entity type and overall, for contemporary and historical models for all five languages, using both strict and lenient interpretations.

all similarity to contemporary language limits the potential advantage of historical specialisation.

Overall, these observations suggest that the quantity and relevance of training data may play a key role in affecting the model performance on this particular task.

5.2. Subword Fertility Rate

One reason for creating historical models from scratch is that the tokeniser is also trained from scratch, on the same historical texts as the pre-training data. The underlying idea is that the resulting vocabulary may better capture the spelling and

Model	tokens	data	coverage
BERToldo$_all$ (it)	∼10M	970 MB	1200-1900
D'AlemBERT (fr)	∼185M	–	1500-1790
MacBERTh (en)	∼3.9B	–	1450-1950
GysBERT (nl)	∼7.1B	–	1618-1999
Hist BERT (de)	–	–	1840-1920

Table 4: Size and temporal coverage of pretraining data for each historical model.

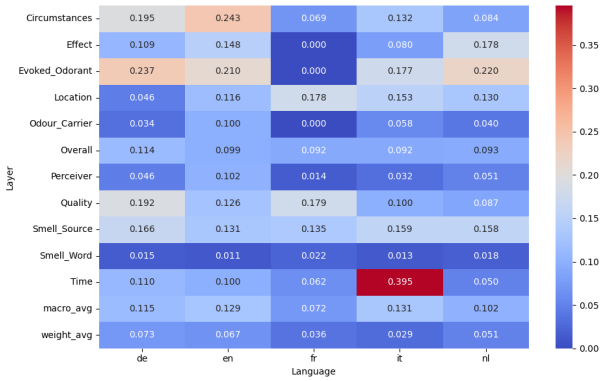
word usage of the historical period. If the spelling of many words in a language has changed strongly between the period of the task data (e.g. the 17th century) and the period of the pretraining data of contemporary models (e.g. 20th and 21st century), then the contemporary tokeniser may split many tokens in historical texts into multiple subtokens, while a tokeniser trained on the same historical texts would use fewer subtokens and for most common words would need only a single subtoken. To test whether the difference in performance could be related to the difference in tokenisers, we look at the Subword Fertility Rate (SFR) (Rust et al., 2021). This is the mean number of subtokens per token for a representative samples of texts. A higher SFR may correspond to a decrease in performance, although Ali et al. (2024) found that this not a reliable measure, at least not for English. We compute the SFR for the tokenisers of contemporary and historical models of each of the five languages using the ground truth datasets, with SFR scores for all tokens in the dataset, and for tokens in the spans of each entity type.

For overall SFR, there is little difference between historical and contemporary models for English (SFR of 1.13 vs. 1.23 respectively), French (1.27 vs. 1.21) and Italian (1.15 vs. 1.25). The differences are bigger for Dutch (1.23 vs. 1.48). Interestingly, for German, the historical tokeniser has a substantially higher SFR than the contemporary model (1.84 vs. 1.29). Based on these numbers, there seems to be no clear relation between differences in SFR and differences in performance. This is further demonstrated when we zoom in on specific entity types. Most models perform better on *Smell_Words* than on other entities, partly because *Smell_Words* are among the most frequent entity types (no Frame Element is annotated in their absence), and partly because they are typically single words, whereas most other entity types span longer sequences, as discussed in Section 4. But for most languages the tokens of *Smell_Words* have the highest SFR of all entity types, and for entity types with the lowest SFR, such as *Time* and *Circumstances* the performance is much lower. When we compare contemporary and historical models on *Smell_Words*, we find that differences in performance in terms of F_1 are minimal (see Table 3), while the SFR differences are big (and in different directions) for English (2.14 for contemporary vs. 1.54 for historical), German (2.18 vs. 2.63) and Italian (1.89 vs. 1.55). Only for Dutch (2.01 vs. 1.94) and French (1.58 vs 1.70) are the SFR differences small. In other words, there is no direct relationship between SFR and performance, although it may be confounded with other factors, such as the amount of orthographic shift in spelling between historical and contemporary texts in a language, the amount of pretraining data that a base model has seen and the amount of task-specific training data per language and entity type. Of course, SFR is not the only aspect that matters in comparing tokenisers of contemporary and historic models, but we include this analysis because SFR can be analysed without recreating the models from scratch.

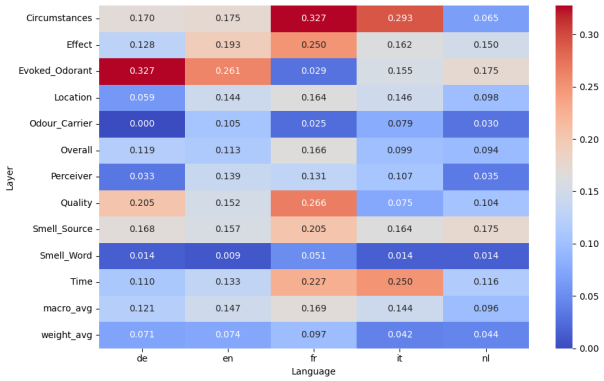
5.3. Exact Versus lenient

In the work of Menini (2024), it was hypothesised that the low performance on certain labels could be partly due to issues with span boundary selection. To test this intuition, we compute lenient F_1 scores, where a predicted tag is considered correct if it partially overlaps with the ground truth tag and has the same entity type. This is useful because, with exact F_1, we risk discarding cases where the assigned tag is actually correct, but one or more words in the expression were not perfectly aligned between the gold and predicted annotations. This issue is particularly relevant for labels that often include entire phrases or sentences, such as *Circumstances*, *Evoked_Odorant* and *Effect*, but it can also provide insights for other labels.

The results reported in Table 3 were compared in terms of loss and gain on exact F_1 in Fig. 2. This comparison clearly shows that while using lenient evaluation, gains are particularly pronounced in the contemporary case. In the case of the contemporary models, a possible correlation emerges between the span length of a label and the gain observed under the lenient evaluation. In particular, *Circumstances*, *Effect* and *Evoked_Odorant*, which tend to correspond to longer spans, show the largest improvements in almost all languages. By contrast, in the historical models, gains are minimal for almost all labels across all languages, with the notable exception of *Time* in Italian. The marked improvement for *Time* is also visible in the contemporary setting. This pattern may be explained by the fact that Italian is the language with the highest average number of tokens per span for this label, followed by French, but further investigation is needed to confirm this hypothesis. These findings support the previously proposed intuition regarding the low performance observed for specific labels. At the same time, the improvements



(a) Lenient F1 gain for historical models



(b) Lenient F1 gain for contemporary model

Figure 2: Comparison of exact and lenient F1 scores across languages and labels for historical (a) and contemporary (b) models.

remain relatively limited for the vast majority of labels, indicating that the primary sources of error are likely to lie elsewhere.

5.4. Span Analysis

Based on random samples of 50-70 partial matches per language, we find that longer ground truth spans and predicted spans often differ by a one or two words, where the added or missing words are arguably not necessary for a meaningful tag. In the following examples we report some recurring span boundary mismatches between the ground truth (GT) annotations and the predicted spans (PRED) in English.[13]

Separate vs. merged span instances:

- *Smell_Word* [nl]
  - GT: “reuk en geur” (smell and scent) as single entity
  - PRED: “reuk” (smell) and "geur" (scent) as two entities

Inclusion or omission of prepositions:

- *Smell_Source* [nl]
  - GT: “hare bloemen” (her flowers)
  - PRED: “bloemen” (flowers)
- *Smell_Source* [en]
  - GT: “fine incense”
  - PRED: “with fine incense”

Unclear span boundaries in longer expressions:

- *Effect* [it]
  - GT: “che serva per generare il vento” (which serves to generate the wind)
  - PRED: “per generare il vento” (to generate the wind)
- *Circumstances* [de]
  - GT: “besonder, wenn sie getrocknet” (particularly, when they are dried)
  - PRED: “wenn sie getrocknet” (when they are dried)

Partial specifications or reduced mention:

- *Location* [fr]
  - GT: “dans tout l’Orient” (in all of the Orient)
  - PRED: “l’Orient” (the Orient).
- *Evoked_Odorant* [it]
  - GT: “leggermente aromatico” (slightly aromatic)
  - PRED: “aromatico” (aromatic)

We argue that the words added in or missing from the predicted spans are often not necessary for the tag to be useful, so lenient F_1 may be a more useful measure. Also, the examples point at an inconsistency in the ground truth labels, as the preposition is sometimes included and sometimes not. Updating the ground truth could therefore also result in clearer and more consistent training targets. To further investigate where models tend to make the most errors, we also analysed how often they incorrectly identified spans by examining the positions of the predicted start and end tokens, aiming at uncovering systematic tendencies in span predictions and identify the most common types of errors. The results (reported in Fig. 3) show that late start and early end predictions were the most frequent, meaning that models often missed the beginning of spans and closed them too early. Another notable pattern is that predicted spans rarely extended beyond the true end, indicating a tendency to under-predict span length. These findings further support that one of the main sources

[13] For a comprehensive overview of the types of errors, refer to this notebook in the GitHub repository.

	side	Start			End		
	location	early	exact	late	early	exact	late
lang	model_type						
de	contemporary	0.30	0.29	0.41	0.38	0.49	0.13
	historical	0.33	0.26	0.41	0.35	0.54	0.10
en	contemporary	0.21	0.37	0.42	0.46	0.47	0.07
	historical	0.26	0.35	0.38	0.42	0.48	0.10
fr	contemporary	0.17	0.26	0.57	0.52	0.34	0.14
	historical	0.17	0.30	0.53	0.55	0.36	0.09
it	contemporary	0.24	0.31	0.45	0.42	0.48	0.10
	historical	0.23	0.41	0.36	0.50	0.41	0.09
nl	contemporary	0.22	0.30	0.48	0.41	0.50	0.09
	historical	0.22	0.29	0.49	0.34	0.55	0.11

Figure 3: Late and early start and end in model predictions across all five languages, comparing contemporary and historical models.

Lang.	#	Add	Skip	One vs. multi GT_1, P_m	One vs. multi P_1, GT_m
DE	49	0.31	0.51	0.12	0.08
EN	43	0.49	0.30	0.19	0.05
FR	44	0.00	0.82	0.18	0.00
IT	51	0.18	0.69	0.16	0.00
NL	51	0.25	0.63	0.16	0.00

Table 5: Types of errors per language, whether the predicted has added or skipped tokens w.r.t. the GT, or whether one GT span covers multiple predicted spans (GT_1, P_m) or vice versa (P_1, GT_m).

of difficulty lies in correctly identifying span boundaries.

Zooming in on specific types of errors, we distinguish between the model adding (Add) or skipping (Skip) tokens with respect to the GT, or whether a single GT span corresponds to multiple predicted spans (GT_1, P_m) or vice versa (P_1, GT_m). The percentage of errors per type is shown in Table 5. There are clear differences between languages. For English, the model is more likely to add additional tokens than to skip tokens, whereas for all the other languages it is more likely to skip tokens. For French, skipped tokens represent 81% of the boundary errors. A less common error is that a single GT span is tagged as multiple spans by the model. The reverse is very rare.

Finally, we look at linguistic error patterns in terms of POS tags. For French and Italian, adjectives are the most commonly skipped (30% and 31% of errors respectively). Other notable error groups are skipped determiners for German, Italian and Dutch (8%, 14% and 18% of errors respectively) and added or skipped prepositions for English (12% and 9%), and noun phrases for English, French, Italian and Dutch (9%, 16%, 10% and 8%). We note that far fewer mistakes are made with adverbs and verbs (5% and 4% across all five languages respectively). As mentioned above, these boundary issues could be due to inconsistent ground truth span selection, with adjectives and prepositions more often than not being excluded in the ground truth spans, leading models learning to mostly skip them too.

6. Conclusions

In this work, we conducted a series of experiments across five languages to investigate the impact of historical models on the extraction of olfactory frames, a task previously explored only in contemporary contexts. We fine-tuned multiple historical models and compared their performance with contemporary models in terms of strict and lenient precision, recall, and F_1, while also considering the potential influence of subtoken fertility rate and performing a detailed error analysis. Our results show that historical models improve performance in English, Dutch, and to a lesser extent German, whereas for Italian and French they lead to a decrease in performance. This effect is likely related to the amount and temporal specificity of the pretraining data. These findings raise interesting questions about the notion of a "historical domain". Our results suggest that assuming all historical texts constitute a single domain may obscure the real factors driving model performance. In particular, historical shifts language use (spelling, vocabulary and word usage) is only on aspect affecting the similarity between pretraining corpora and the benchmark data. But there may be many other factors, such as text type, genre and topic that, along with corpus size and coverage, play a key role. Future work should disentangle these factors to better understand the various dimensions of language difference and change and how such characteristics affect model learning. The analysis of lenient scores further highlights that historical models yield gains across nearly all labels and languages, suggesting that the low performance of many labels is at least partially due to span boundary issues, as previously hypothesised by Menini (2024), and confirmed by our span analysis. At the same time, the lenient evaluation also reveals some inconsistencies in the ground truth annotations, indicating areas for potential improvement in the dataset. Overall, our study demonstrates that the use of historical models can provide benefits for olfactory extraction tasks in older texts, while also pointing to important directions for understanding domain specificity, annotation quality, and label-level challenges.

7. Limitations

This study relies on a single experimental setting. While our analyses provide valuable insights into span prediction errors and model behavior, further experiments exploring alternative settings, parameters, and training strategies would be necessary to assess the generalisability of our findings for the task of olfactory event extraction.

8. Acknowledgments

This research was funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant nos. EINF-13810 and EINF-16838. Both authors contributed equally to this work.

9. Bibliographical References

Ahmed Alajrami and Nikolaos Aletras. 2022. How does the pre-training objective affect what large language models learn about linguistic properties? *arXiv preprint arXiv:2203.10415*.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, et al. 2024. Tokenizer choice for llm training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2022. Bertoldo, the historical bert for italian. In *Proceedings of the second workshop on language technologies for historical and ancient languages*, pages 68–72.

Ryan Brate, Paul Groth, and Marieke van Erp. 2020. Towards olfactory information extraction from text: A case study on detecting smell experiences in novels. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 147–155, Online. International Committee on Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Charles J Fillmore et al. 2006. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From FreEM to d'AlemBERT: a large corpus and a language model for early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3367–3374, Marseille, France. European Language Resources Association.

Clovis Gladstone, Zhao Fang, and Spencer Dean Stewart. 2025. Ground truth generation for multilingual historical nlp using llms. *arXiv preprint arXiv:2511.14688*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Leah Hamilton, Caitlin Neill, and Jacob Lahne. 2023. Flavor language in expert reviews versus consumer preferences: an application to expensive american whiskeys. *Food Quality and Preference*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

David Howes. 2006. Charting the sensorial revolution.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

E. Lefever, I. Hendrickx, I. Croijmans, A. Van den Bosch, and A. Majid. 2018. Discovering the language of wine reviews: A text mining account. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Junyi Li, Tianyi Tang, Zheng Gong, Lixin Yang, Zhuohao Yu, Zhipeng Chen, Jingyuan Wang, Xin Zhao, and Ji-Rong Wen. 2022. ElitePLM: An empirical study on general language ability evaluation of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3519–3539, Seattle, United States. Association for Computational Linguistics.

Enrique Manjavacas and Lauren Fonteyn. 2021. Macberth: Development and evaluation of a historically pre-trained language model for english (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Steven McGregor and Barbara McGillivray. 2018. A distributional semantic methodology for enhanced search in historical records: A case study on smell. In *14th Conference on Natural Language Processing*.

Stefano Menini. 2024. Semantic frame extraction in multilingual olfactory events. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)*, pages 14622–14627.

Stefano Menini, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, et al. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd workshop on computational approaches to historical language change*, pages 1–10.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *arXiv preprint arXiv:1608.07836*.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Stefan Schweter. 2020. Italian bert and electra models. *Zenodo*.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition. *arXiv preprint arXiv:2205.15575*.

Dan Sperber. 1975. *Rethinking symbolism*. 11. CUP Archive.

Riste Stojanov, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of medical Internet research*, 23(8):e28229.

Simeng Sun, Brian Dillon, and Mohit Iyyer. 2022. How much do modifications to transformer language models affect their ability to learn linguistic knowledge? In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 46–53, Dublin, Ireland. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically

constructed sensorial lexicon. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1511–1521, Doha, Qatar. Association for Computational Linguistics.

Sara Tonelli and Stefano Menini. 2021. Framenet-like annotation of olfactory information in texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20.

William Tullett, Inger Leemans, Hsuan Hsu, Stephanie Weismann, Cecilia Bembibre, Melanie A Kiechle, Duane Jethro, Anna Chen, Xuelei Huang, Jorge Otero-Pailos, et al. 2022. Smell, history, and heritage. *The American Historical Review*, 127(1):261–309.

Rob Van Der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637.

Bodo Winter, Marcus Perlman, and Asifa Majid. 2018. Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, 179:213–220.