A Multilingual Benchmark to Capture Olfactory Situations over Time

S. Menini[1] **T. Paccosi**[1] **S. Tonelli**[1] **M. Van Erp**[2] **I. Leemans**[2]
P. Lisena[3] **R. Troncy**[3] **W. Tullett**[4] **A. Hürriyetoğlu**[2] **G. Dijkstra**[2]
F. Gordijn[2] **E. Jürgens**[2] **J. Koopman**[2] **A. Ouwerkerk**[2] **S. Steen**[2]
I. Novalija[5] **J. Brank**[5] **D. Mladenić**[5] **A. Zidar**[5]
[1] FBK, [2] KNAW, [3] EURECOM, [4] ARU, [5] JSI,
{menini,tpaccosi,satonelli}@fbk.eu, inger.leemans@huc.knaw.nl
{marieke.van.erp, ali.hurriyetoglu}@dh.huc.knaw.nl
{lisena,troncy}@eurecom.fr, william.tullett@aru.ac.uk
{inna.koval,janez.brank,dunja.mladenic}@ijs.si

Abstract

We present a benchmark in six European languages containing manually annotated information about olfactory situations and events following a FrameNet-like approach. The documents selection covers ten domains of interest to cultural historians in the olfactory domain and includes texts published between 1620 to 1920, allowing a diachronic analysis of smell descriptions. With this work, we aim to foster the development of olfactory information extraction approaches as well as the analysis of changes in smell descriptions over time.

1 Introduction

Human experience is mediated through the senses, which we use to interact with the world. Since the perceptual world is so important to us, all languages have resources to describe the sensory perception. Nevertheless, previous research showed that, at least in Western European languages, the visual dimension is prevalent in language, with a richer terminology used to describe it, while the olfactory dimension is less represented (Winter, 2019). For example, in English, there are less unique words for the smell domain than for the other senses. They are also used less frequently and olfactory descriptions are often a target of cross-modal expressions.

Sensory terminology has been researched previously, with the goal to build resources and to analyse how the different senses are described in language (Tekiroğlu et al., 2014b,a). Some research is specifically devoted to smell (Lefever et al., 2018), but they all focus on contemporary language. One notable exception is the collection of essays published in Jędrzejowski and Staniewski (2021), where olfaction in different languages is analysed in a diachronic perspective. For example, Strik Lievers (2021) describes how the olfactory lexicon has changed from Latin to Italian.

In this work, we contribute to the diachronic analysis of olfactory language by annotating a multilingual benchmark with smell situations spanning three centuries. Compared to existing studies, our focus is not on the occurrences of single terms, but we rather capture smell events in texts, i.e. more complex structures involving different participants. The benchmark currently covers six languages (Dutch, English, French, German, Italian and Slovene). Annotation of Latin data is ongoing, but we do not include here the results for this language because they are still preliminary.

We describe the annotation guidelines and the document selection process. Our benchmark includes texts issued between 1620 and 1920 covering ten domains of olfactory interest to cultural history. We release the benchmark at https://github.com/Odeuropa/benchmarks_and_corpora and we present a first analysis of its content.

2 Related Work

Studies on olfactory language in cognitive science primarily focus on the verbal expressions of the odour perceived (Majid and Burenhult, 2014; Majid et al., 2018), while in historical studies, instead, they mainly deal with the textual accounts of experienced smells, as in Tullett (2019). Within the NLP community, little attention has been devoted to the automatic analysis of smell references in texts. Most works have focused on the creation of lexical databases, for example Tekiroğlu et al. (2014b,a) worked on the creation of Sensicon, representing the first systematic attempt to build a lex-

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 1 - 10

icon automatically associated to the five senses. Other studies have focused on synaesthetic aspects of language, dealing with the multisensoriality of sensory words. For instance, Lievers and Huang (2016) create a controlled lexicon of perception, while Girju and Lambert (2021) propose to use word embeddings for the extraction of sensory descriptors and their interconnections in texts.

As regards smell-specific works, Brate et al. (2020) propose both a simple annotation scheme to capture odour-related experiences and two semi-supervised approaches to automatically replicate this annotation. Lefever et al. (2018) present an automated analysis of wine reviews, where olfaction plays a fundamental role, while McGregor and McGillivray (2018) introduce an approach to automatically identify smell-related sentences in a corpus of historical medical records using distributional semantic modelling. More recently, Tonelli and Menini (2021) present FrameNet-inspired guidelines to annotate smell events in texts. We consider this work the starting point upon which we build our annotation task. In particular, we aim at assessing the underlying assumptions of such guidelines: whether frames can be applied diachronically and across languages using the same annotation scheme.

3 Annotation Guidelines

Annotation of olfactory events and situations in texts is a new task that was recently introduced in Tonelli and Menini (2021). We adopt the same framework in this work, whose guidelines are summarised below.

Olfactory annotation is inspired by the FrameNet project (Ruppenhofer et al., 2006)[1] which, focusing on the semantic dimension of situations and participants, should be easily applicable to multiple languages and constructions. In FrameNet, events and situations are so-called *frames* and are used as synonyms for schemata, semantic memory or scenarios. They represent the components of the internal model of the world that language users have created by interpreting their environment (Fillmore, 1976).

According to frame semantics, a frame includes two main components: *lexical units* (LUs) and *frame elements* (FEs). The former are words, multiwords or idiomatic expressions that evoke a specific frame, while the latter are frame-specific semantic roles that, in case of verbal LUs, are usually realized by the syntactic dependents of the verb. For example, the *Commerce pay* frame includes as lexical units 'pay', 'payment', 'disburse', 'disbursement', 'shell out', and has the following frame elements: Buyer, Goods, Money, Rate, Seller.

While FrameNet aims to be a general-purpose resource, the guidelines we follow only concern olfactory situations. Therefore, the scope of our annotation considers only smell-related lexical units and a single frame of interest, the *Olfactory event*. The same structure as the original FrameNet is adopted based on lexical units and related frame elements. When necessary, domain-specific semantic roles are introduced upon discussions with experts in olfactory heritage and history. For example, the roles *Smell source*, *Evoked odorant* and *Odour carrier* were not originally in FrameNet, while some generic roles such as *Perceiver*, *Time*, *Location* and *Circumstances* are borrowed from the original resource. An overview of the frame elements included in our annotation is shown in Table 2.

The list of lexical units (LUs) was defined with the help of domain experts, choosing smell-related lexical units that evoke olfactory situations and events. The LU lists were created in six languages, namely English, Dutch, Italian, French, German and Slovenian. They include basic smell-related terms, which are generally comparable across languages (for instance the translation of words such as 'to smell', 'odour' 'odorous', 'smelly', 'perfume'). The lists were extended with language- and culture-specific terms, such as German compound nouns created with the roots '-gestank' and '-geruch', e.g. *Regengeruch* ('rain smell') or *Viehgestank* ('cattle stink'). The initial version of the list is reported in Table 1.

We consider these guidelines appropriate for our task because they have been designed following a multilingual perspective, with no language-specific adaptations. Furthermore, as we annotate documents from different time periods, LU lists are not fixed, giving the possibility to add new items as the outcome of the annotation process.

4 Document selection

In close collaboration with cultural historians, we defined ten domains of interest, where we expected to find a high number of smell-related

[1] https://framenet.icsi.berkeley.edu

English
Nouns: stink, scent, scents, smell, smells, odour, odor, odours, odors, stench, reek, aroma, aromas, aromatic, whiff, foetor, fetor, fragrance, musk, rankness, redolence, pong, pungency, niff, deodorant, olfaction **Verbs**: smelling, smelled , reeked, sniff, sniffed, sniffing, whiffed, fragrance, deodorized, deodorizing, snuffing, snuffed **Adjectives**: stinking, stank, stunk, scented, odourless, odoriferous , odorous, malodorous , reeking, aromatic , whiffy, fetid, foetid, fragrant, fragranced, redolent, frowzy, frowsy, pungent, funky, musty, niffy, unscented, scentless, deodorized, noisome , smelly, mephitic, olfactory **Adverbs**: musky, pungently **Other**: atmosphere, essence, putrid.
Dutch
Nouns: Aroma, Damp, Geur, Geurigheid, Geurstof, Geurtje , Luchtje, Miasma, Mufheid, Odeur, Parfum, Parfumerie, Reuck, Reuk, Reukeloosheid, Reukerij, Reukje, Reukloosheid, Reukorgaan, Reukstof, Reukwater, Reukwerk, Reukzin, Riecking, Rieking , Ruiker, Snuf, Stank, Stinkbok , Stinker, Stinkerd, Stinkgat, Stinknest, Vunsheid, Waesem, Walm, Wasem, Deodorisatie, Desodorisatie **Verbs**: Aromatiseren, Deodoriseren, Desodoriseren, Geuren, Meuren, Neuzen, Ontgeuren, Opsnuiven, Parfumeren, Rieken, Riecken, Ruiken, Ruycken, Snuffelen, Stinken, Uitwasemen, Vervliegen, Wasemen, Zwemen **Adjectives**: Aromatisch, Balsemachtig, Balsemiek, Geparfumeerd, Geurig, Geurloos, Heumig, Hommig, Hummig, Muf, Muffig, Neuswijze, Onwelriekend, Penetrant, Pisachtig, Reukloos, Riekelijk, Ruikbaar, Schimmelig, Soetgeurig, Soetreukig, Stankloos, Stankverdrijvend, Stankwerend, Stinkend, Stinkerig, Vervliegend, Vuns, Vunze, Weeïg, Welriekend, Zwavelig **Adverbs**: neusgierig, neuswijs, neuswijsheid, neuswijze, reuklustig, welgeneusd **Kinds of smell**: aardgeur, aardlucht, aardreuk, aaslucht, ademlucht, ambergeur, amberlucht, amberreuk, anijsgeur, balsemgeur, balsemlucht , bosgeur, braadgeur, braadlucht, brandlucht, brandreuk, dennenlucht, gaslucht , gasreuk, graflucht, harslucht, houtlucht, Huim, lijklucht, Meur, modderlucht , Muf, muskusgeur, muskusreuk, pestlucht, roetlucht, rooklucht, rotlucht, rozengeur, wierookgeur, wierookwalm, wierookwolk, wijnreuk, zweetlucht, Pekgeur, Pikreuk (and anything ending with -geur or -reuk).
Italian
Nouns: lezzo, morbo, putidore, fiatore, puzzo, puzza, fetore, miasma, putrefazione, effluvio, esalazione, estratto, odore, aroma, olezzo, fragranza, profumo, aulimento, odoramento, afrore, tanfo, tanfata, zaffata **Verbs**: odorare, puzzare, profumare, deodorare, odorizzare, aromatizzare, fiutare, annusare, nasare, olezzare, ammorbare, appestare, impestare, impuzzare, impuzzire, impuzzolentire, impuzzolire, intanfare **Adjectives**: puzzolente, fetente, fetido, deodorizzato, putrefatto, odorato, odoroso, odorifero, aromatizzato, profumante, profumato, suave, soave, olfattivo, olfattorio, maleodorante, aromatico, pestilenziale, puzzoso, fragrante **Adverbs**: profumatamente, odorosamente **Other**: essenza, atmosfera, sentire
French
Nouns: puanteur, flair, odeur, odorat, parfum, arôme, déodorant, nez, narine, gaz, baume, senteur, fragrance, musc, senteur, aigreur, olfaction, odorat, effluve, exhalaison, fumet, relent, pestilence, fétidité, remugle **Verbs**: puer, flairer, exhalter, odoriser, renifler, schlinguer, chlinguer, empester, parfumer, désodoriser, humer, renifler, embaumer **Adjectives**: puant, odorant, fétide, aromatique, olfactif, odorifère, odoriférant, nasal, pestilentiel, infect, malodorant, parfumé, inodore, piquant, désodorisé, méphitique, olfactif, empesté, infect, nauséabond **Other**: émanation, bouquet (about wine), sentir, sniffer, dégoûtant, dégoutant, écoeurant, percevoir
German
Nouns: Geruch, Gestank, Aroma, Parfum, Parfüm, Parfümöl, Duft, Dampf, Dunst, Duftstoff, Riechwasser, Duftwasser, Riechorgan, Geruchsorgan, Nase, Riechstoff, Aromastoff, Riechwasser, Duftwasser, Riecher, Qualm, Zigarettenqualm Anything ending on -geruch / -gestank / -duft **Verbs**: aromatisieren, riechen, stinken, schnüffeln, schnuppern, beschnuppern, parfümieren, ausdünsten, duften, qualmen, einatmen, inhalieren, ausdünsten, exhalieren, verfliegen, verdampfen, evaporieren, sich verflüchtigen **Adjectives**: parfümiert, olfaktorisch, wohlriechend, stinkend, duftend, riechend, muffig, modrig, aromatisch, blumig, geruchlos, penetrant, durchdringend, schimmlig, schimmelig Anything ending on -duft / - duftig / -riechend **Kinds of smell**: Aasgestank, Abgasgeruch, alkoholisch, angebrannt, angenehm, anregend, Apfelduft, beißend, Babygeruch, blumig, brennend, durchdringend, dominant, ekelerregend, ekelhaft, erdig, erfrischend, erregend, fade, faul, frisch, fruchtig, harzduftend, harzig, herb, herbstlich, holzig, intensiv, kamillig, käsig, klinisch, ländlich, Lavendelduft, Lebkuchenduft, ledrig, Leichengeruch, Leichengestank, metallisch, mild, minzig, mosig, Moschusgeruch, muffig, muffelig, nussig, Pfefferminzgeruch, pilzig, Puderduft, ranzig, rauchig, Regengeruch, salbeiartig, salzig, Sandelholzduft, säuerlich, schal, schwefelig, schweißig, Schweißfußgeruch, sommerlich, schwer, seifig, staubig, stechend, steril, stickig, streng, süßlich, Tabakgeruch, unangenehm, Uringeruch, verbrannt, verfault, Viehgestank, Weihrauchduft, Wundgestank, würzig, zimtig, zitronig. Anything ending on - duft / -geruch
Slovenian
Nouns: vonj, smrad, duh, voh, vonjava, dišava, umetna dišava, parfum, aroma, dišavina, priduh, vzduh, aromatičnost, pookus, pikantnost, zatohlost, deodorant, dezodorant, zadah, zaudarjanje **Verbs**: smrdeti, zaudarjati, dišati, zadišati, zavonjati, zadehteti, zaduhteti, vohati, duhati, vonjati, ovohati **Adjectives**: gnil, smrdljiv, smrdeč, umazan, usmrajen, prijeten, dišeč, aromatičen, dišaven, zaudarjajoč, postan, zatohel, opojen, brez vonja, vohalen, žaltav, strupen, toksičen, ogaben, oster, pikanten, vohalen, odišavljen **Other**: plesniv, pokvarjen, zadušljiv, zadušen, čuten, zavdajati, buket

Table 1: Initial list of possible lexical units for each language of interest. We list under *Other* the terms that were initially not included because they are ambiguous, but that were annotated as lexical units during benchmark creation.

documents. These domains are: *Household & Recipes*, *Law and Regulations*, *Literature*, *Medicine & Botany*, *Perfumes & Fashion*, *Public health*, *Religion*, *Science & Philosophy*, *Theatre*, *Travel & Ethnography*. The additional category *Other* was included in the list for documents which are relevant to the olfactory dimension but do not fall within any of the previously mentioned categories. Ideally, the benchmark should contain 10 documents for each category, distributed evenly over the time period between 1620 and 1920, for a total of 100 documents. However, no strict length requirements were defined for each document, because their availability and characteristics change drastically across languages. In some cases, a document may be few pages with dense olfactory in-

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].*
Creator	The person that creates a (usually pleasant) smell. *The origin of perfume is commonly attributed [to the ancient Egyptians].*

Table 2: Overview of the Frame Elements (FEs) related to Olfactory situations and events with corresponding examples. Lexical units are underlined and the FE of interest is in square brackets. The same definitions hold for all languages included in the benchmark. For more details on FEs descriptions see (Tonelli and Menini, 2021).

formation, while in some other cases a book could contain smell references scattered throughout the volume. Therefore, each of the six annotation teams was free to apply the most appropriate criteria for the selection of documents to annotate. For example, Dutch annotators decided to focus on short text snippets of around 20 sentences. For Italian and English, longer passages up to a few hundred sentences are included. Other differences across languages concern the quality and variety of available documents in digital format. While for some languages, such as Dutch and English, large online repositories exist and it was possible to find documents belonging to each of the 10 domains and covering the time span of interest, the limited availability of digital repositories of Slovenian texts does not allow the collection of the full set of documents. This is the main reason why there are some qualitative and quantitative differences among languages.

Annotations were performed using INCEpTION (Klie et al., 2018), a web-based platform which allows three levels of authorisations (administrator, curator, annotator) and is therefore particularly suitable to support large annotation efforts like ours. A screenshot of the interface is shown in Figure 1.

5 Quality control

We implement two quality control measures: 1) a web-based consistency checker, and 2) double annotation of a set of documents for each language to compute inter-annotator agreement and discuss difficult cases.

5.1 Quality Consistency Check

Given the complexity of the annotation process, which is carried out by multiple annotators for each of the six languages, it is important to ensure that the different annotations are consistent with the instructions provided in the guidelines.

To facilitate a consistency check, we developed a web-based tool to automatically find when annotations are not compliant with the guidelines. The tool takes an exported WebAnno file from INCEpTION as input and outputs a report describing

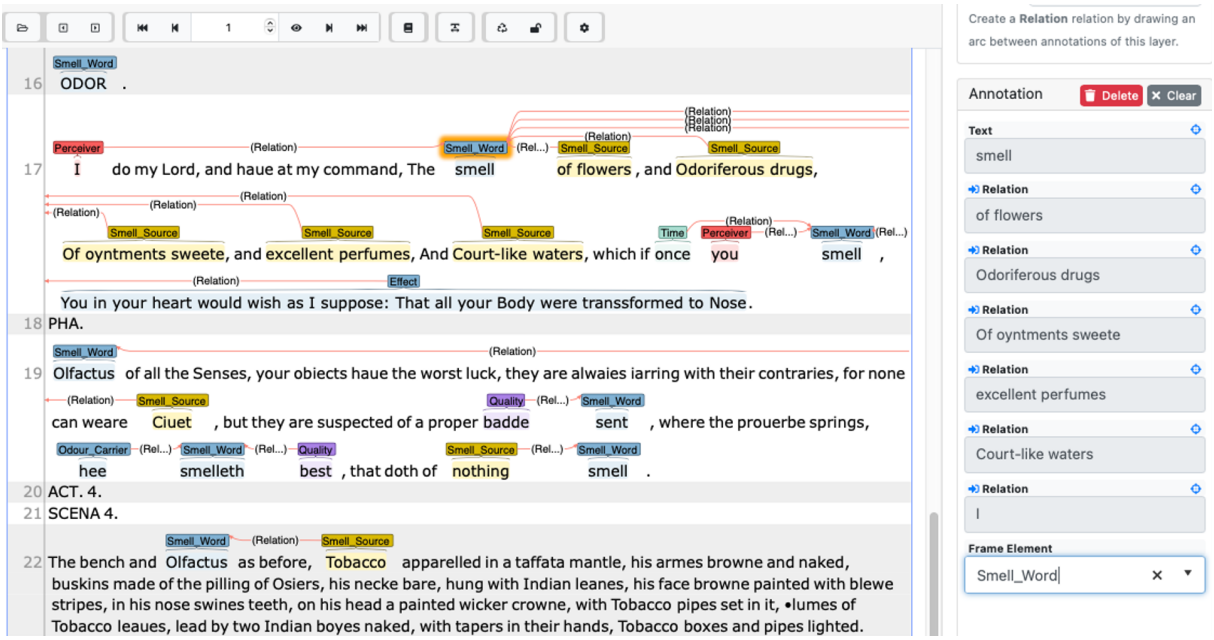

Figure 1: Screenshot of the INCEpTION annotation tool

which inconsistencies are found and where (with document ID, sentence number and string). This makes it straightforward to find the mistake and fix it quickly.

The inconsistencies identified in the files are related to both incorrect and missing annotations, focusing on the annotation procedure and not the content of the annotations. For instance, it checks if every frame element is properly connected to a smell word and if all selected spans have been assigned to a corresponding label. Operating at the level of labels and relations, that are the same for every language, and not considering the text content, the tool is language-independent.

After analysing the annotation output, the quality checker returns details about five error types:

- Spans that have been selected but not labeled;
- Smell words with double annotation, which have not been linked to themselves;[2]
- Frame elements that despite being annotated are not linked to any other element in text;
- A *Smell_Word* is the starting point of a relation instead of the ending point;
- Frame elements connected to something other than a *Smell_Word*.

Given the complexity of the annotation, for all languages involved the quality check step has been very useful to identify formal mistakes, allowing the removal of dozens of inconsistencies.

5.2 Inter-Annotator Agreement

Having at least two annotators for each language is necessary to obtain a double annotation of a subset of the benchmark and compute inter-annotator agreement, which is commonly considered a measure of annotation quality (Artstein and Poesio, 2008).

INCEpTION contains an integrated set of tools to compute inter-annotator agreement.[3] Among the proposed metrics, the most suitable for our task is Krippendorff's alpha (Krippendorff, 2011), as it supports more than two annotators (that is the case for some of the languages). This measure considers also partial overlaps, e.g. one annotator labelled only a noun while the other included also its article.

Inter-annotator agreement between two raters was computed, usually over a set of around 200 annotations (both FEs and smell words). In general, this was carried out after an extensive ini-

[2]There are instances where the same token can be at the same time a *Smell_Word* and another frame element related to the *Smell_Word* itself. For instance, 'odoriferous' may be both a *Smell_Word* and a *Quality*. In these cases, a relation should be set between the FE label and the smell word. This error notifies the absence of this relation.

[3]More details about this function are documented at https://inception-project.github.io/releases/20.2/docs/user-guide.html#sect_monitoring_agreement

	Dutch	English	French	German	Italian	Slovenian
Smell words	1,788	1,530	845	2,659	1,254	1,973
Total FEs	4,962	4,023	1,876	5,885	2,664	4,445
Source	1,922	1,313	710	2,297	952	1,638
Quality	1,071	1,084	450	1,730	707	936
Perceiver	336	362	140	399	153	266
Circumstances	399	248	88	274	202	228
Odour carrier	351	310	106	170	195	408
Effect	243	187	53	425	104	214
Evoked Odorant	228	91	103	258	74	285
Place	255	302	172	200	158	394
Time	127	126	49	131	119	75
Creator	30	0	5	1	0	1

Table 3: Overview of benchmark content for each language.

tial training of annotators. Agreement is 0.68 for English, 0.56 for Slovenian, 0.62 for French and 0.74 for Italian. For the other languages the process is still ongoing. In general, the major sources of disagreement are the extent of FE spans, a rather long distance between a FE and a smell word and possible different interpretations of some roles, in particular Location vs. Circumstances and Smell source vs. Odour carrier. While annotation guidelines were updated to make these distinctions clearer, some cases of disagreement are still very much dependent on annotators' preferences and interpretation.

6 Benchmark statistics

In this section, we detail the content of our benchmark in each language. Table 3 shows the number of occurrences of smell words and frame elements. Overall, for each language a good number of smell-related events and situations were annotated.

The average number of frame elements (FEs) associated with each smell event is between 2.1 and 2.7 for all languages, showing an interesting common feature. Furthermore, the most frequent FE is the *Smell Source*, followed by the *Quality* for all languages. This shows a pattern in the way smell situations and events are typically described, where the source and the quality are clearly core elements that are necessary to characterise the scene.

The FE element with the least annotations is instead 'Creator'. This is due to the fact that this role was added at a later stage in the annotation process, mainly to cover documents related to perfumery. It is therefore present only in the benchmarks that contain this kind of documents. For further discussion see Section 8.

In Figure 2, we report the number of documents per domain in each language-specific benchmark (see list of domains in Section 4). Overall, we observe a prevalence of literary texts (LIT), probably because this is the most represented domain in large repositories such as Wikisource and Project Gutenberg. Travel literature and medical texts are also well-represented in all languages. Despite the effort to have a balanced benchmark covering the same domains in all languages, however, results are mixed. For some languages, well-represented in large digital repositories, this balance was possible to some extent, with English being the only one covering all domains. For other languages, the benchmarks are affected by the limited variety of resources available in digital format, see for example Slovenian. Availability is a major obstacle when trying to create historical corpora that cover different domains.

In Figure 3, we report the temporal distribution of the documents present in the benchmark for each language. All languages overlap in the time period of interest, with the Dutch benchmark including some earlier texts but no data after 1880, and the Italian dataset going beyond 1930. Similar to the above remarks, also in this case we observe that, due to different data availability, not all time periods are covered equally.

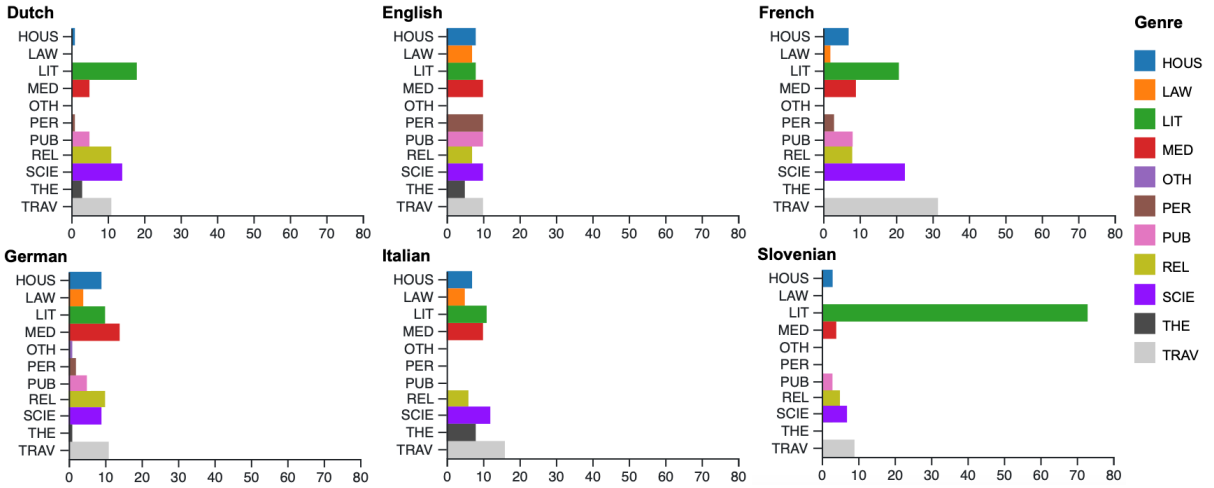

Figure 2: Number of documents per domain in each language-specific benchmark. HOUS = Household & Recipes, LAW = Law, LIT = Literature, MED = Medicine & Botany, OTH = Other, PER = Perfumes & Fashion, PUB = Public health, REL = Religion, SCIE = Science & Philosophy, THE = Theatre, TRAV = Travel & Ethnography.

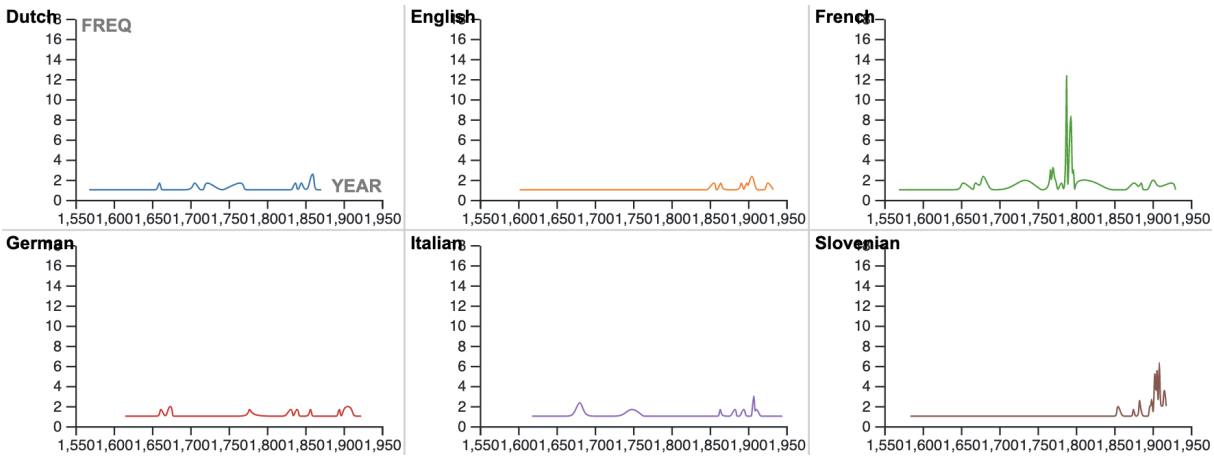

Figure 3: Temporal distribution of documents in each language-specific benchmark

7 Towards smell related information extraction

One of the goals of this benchmark is to enable temporal-aware information extraction tasks related to the olfactory domain. As a first step in this direction, we explore sentence classification using the English benchmark. Since our corpus consists of historical documents, we evaluate performance of a transformer model that is pre-trained using historical corpora, in light of Lai et al. (2021)'s proposal.

We focus on the task of classifying sentences as smell-related or not. Since the corpus is annotated at token level, we first label the sentences that contain any smell event annotation as smell-related, which are 897 out of the total 3,141 sentences. We randomly choose 650 (190 smell-related, 460 not smell-related) sentences as a held-out to measure the performance of fine-tuning on the remaining 2,491 sentences.

We compare the performance obtained using BERT base uncased with sequence length 128[4] (Devlin et al., 2019), RoBERTa base case-sensitive with sequence length 512[5] (Liu et al., 2019), and MacBERTh (Manjavacas and Fonteyn, 2021)[6] to identify sentences that are smell-related in English. MacBERTh is a BERT variant that is uncased with sequence length 128 and pre-trained from scratch using historical corpora. Each model was fine-tuned five times using five different ran-

[4] https://huggingface.co/bert-base-uncased, accessed on February 27, 2022

[5] https://huggingface.co/roberta-base, accessed on February 27, 2022

[6] https://www.github.com/emanjavacas/macberth-eval, accessed on February 27, 2022.

dom seeds (42, 43, 44, 45, 46) for all random aspects of the fine-tuning, batch size of 64, sequence length of 64, learning rate (2e-5), epochs (30), and random splitting for obtaining a development set from the training set (.15). Table 4 demonstrates the median performance of each fine-tuned model in terms of Matthews Correlation Coefficient (MCC), Precision, Recall, and F1-macro on the held-out dataset. We observe that macBERTh, which was pretrained using historical data, outperforms the base transformer models BERT and RoBERTa. This confirms the need to build models that are temporal-aware when dealing with historical corpora. Furthermore, the performance achieved by all models is above 0.90, showing that it is possible to yield good results in the task even if using relatively few training data.

Model	MCC	Precision	Recall	F1-macro
BERT	81.44	92.82	90.17	90.43
MacBERTh	85.66	94.08	91.91	92.72
RoBERTa	84.51	93.43	91.43	92.11

Table 4: Median scores in terms of Mathews Correlation coefficient (MCC) and macro precision, recall, and F1 over five runs

We analyzed the predictions of the best RoBERTa and MacBERTh models on 300 test sentences divided into two groups: the first one includes test sentences from documents published between 1619 and 1846, while the second covers the time period between 1847 and 1925. The F1-macro obtained with the MacBERTh model is 95.40 and 90.46 for the earlier (1619-1846) and later periods (1847-1925) respectively. The RoBERTa model achieves 92.46 and 91.42 F1-macro in the same setting. Although the MacBERTh model yields significantly better results for data published in the earlier period, the RoBERTa model yields a balanced performance across periods.

8 Discussion

During the creation of the benchmark, we have encountered two major issues related to working with historical data. The first, already mentioned in Section 6, is the limited availability of documents for some languages, domains and time spans. This has affected the possibility to create balanced benchmarks for all six languages, although a remarkable effort was put in manually looking for digital collections and selecting relevant documents.

Another major issue was the need to clean or correct some of the texts before the annotation, mostly due to the limits of OCR applied to old documents. Problematic transcriptions can be connected in part to stains or other imperfections in the paper, and in part to the evolution of language, with older documents presenting letters that have fallen into disuse in contemporary language. For instance, in French, Italian and English we found lost characters (e.g. long s "ſ", often confused with "f" as in “perfumes", misspelled as “persumes" in English), characters used differently (v instead of u, like in “vne" for French, or “vlcers" for English), changes in word spelling (“pourquoy" instead of “pourquoi" in French), and abandoned words.

Another interesting element is that annotation guidelines were adapted several times during the benchmark creation process, because it was not possible to foresee all potential issues we encountered during annotation. Indeed, domain specificity of some texts and the different use of language in historical documents made it difficult to straightforwardly follow annotation instructions. For example, frame element definitions have been adjusted and the ‘Creator’ element was added. Furthermore, the initial list of lexical units (Table 1) was extended in the process, enabling annotators to add new terms encountered during manual labelling.

9 Conclusion and Future Work

In this paper, we presented a multilingual benchmark annotated with smell-related information and covering six languages, which we make available to the research community. We have described the document selection rationale, the annotation process and the main challenges related to the creation of a multilingual benchmark containing historical documents. Annotation of Latin is in progress, and it will be added to the benchmark as soon as it is complete.

The benchmark is only a first step towards the analysis and extraction of olfactory information from historical documents. The work introduced in Section 7 will be extended to all six languages, using historical BERTs when available. Furthermore, we will go beyond simple sentence classification, training multilingual classifiers to iden-

tify lexical units and frame elements. Since the size of the benchmark is rather limited, we will try to expand it in the future but also explore semi-supervised, few-shot and cross-lingual approaches to olfactory information extraction.

Acknowledgements

This research has been supported by the European Union's Horizon 2020 program project ODEUROPA[7] under grant agreement number 101004469.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

C. Fillmore. 1976. Frame semantics and the nature of language *. *Annals of the New York Academy of Sciences*, 280.

Roxana Girju and Charlotte Lambert. 2021. Intersense: An investigation of sensory blending in fiction. *CoRR*, abs/2110.09710.

Łukasz Jędrzejowski and Przemysław Staniewski. 2021. *The Linguistics of Olfaction. Typological and Diachronic Approaches to Synchronic Diversity*. John Benjamins, Amsterdam.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

[7] https://odeuropa.eu/

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. https://repository.upenn.edu/asc_papers/43/.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event Extraction from Historical Texts: A New Dataset for Black Rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400, Online. Association for Computational Linguistics.

Els Lefever, Iris Hendrickx, Ilja Croijmans, Antal van den Bosch, and Asifa 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).

Francesca Strik Lievers and Chu-Ren Huang. 2016. A lexicon of perception for the identification of synaesthetic metaphors in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2270–2277, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Asifa Majid and Niclas Burenhult. 2014. Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2):266–270.

Asifa Majid, Niclas Burenhult, Marcus Stensmyr, Josje De Valk, and Bill S Hansson. 2018. Olfactory language and abstraction across cultures. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170139.

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 (NLP4DH)*.

Stephen McGregor and Barbara McGillivray. 2018. A distributional semantic methodology for enhanced search in historical records: A case study on smell. In *Proceedings of the 14th Conference on Natural Language Processing, KONVENS 2018, Vienna, Austria, September 19-21, 2018*, pages 1–11. Österreichische Akademie der Wissenschaften.

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. Framenet ii: Extended theory and practice. Working paper, International Computer Science Institute, Berkeley, CA.

Francesca Strik Lievers. 2021. Smelling over time. the lexicon of olfaction from latin to italian. In (Jędrzejowski and Staniewski, 2021), pages 369–397.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014a. A computational approach to generate a sensorial lexicon. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 114–125, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014b. 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, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

William Tullett. 2019. *Smell in Eighteenth-Century England: A Social Sense*. Oxford University Press.

Bodo Winter. 2019. *Sensory linguistics: Language, perception and metaphor*, volume 20. John Benjamins Publishing Company.