Multilingual Historical Corpora and Annotated Benchmarks

Deliverable D3.2

Version FINAL

Odeuropa

NEGOTIATING OLFACTORY AND SENSORY EXPERIENCES IN CULTURAL HERITAGE PRACTICE AND RESEARCH

The Odeuropa project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004469. This document has been produced by the Odeuropa project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.

Grant Agreement No.	101004469
Project Acronym	ODEUROPA
Project full title	Negotiating Olfactory and Sensory Experiences in Cultural Heritage Practice and Research
Funding Scheme	H2020-SC6-TRANSFORMATIONS-2020
Project website	http://odeuropa.eu/
Project Coordinator	Prof. Dr. Inger Leemans KNAW Humanities Cluster Email: inger.leemans@huc.knaw.nl
Document Number	Deliverable D3.2
Status & version	FINAL
Contractual date of delivery	September 2021
Date of delivery	31 October 2021
Type	Report and Data
Security (distribution level)	Public
Number of pages	16
WP contributing to the deliberable	WP3
WP responsible	WP3
EC Project Officer	Hinano Spreafico

Authors: Sara Tonelli,[1] Stefano Menini,[1] Marieke van Erp,[2] Raphaël Troncy,[3] Inna Novalija[4]

Internal reviewers: Pasquale Lisena,[3] Elisa Leonardelli[1]

Affiliations (1) FBK, (2) KNAW, (3) EURECOM, (4) JSI

Keywords: linguistic annotation, benchmark, corpora, multilinguality

Abstract: This document is a brief description of the content of the multilingual historical corpora we have collected within the Odeuropa project to study olfactory information in texts. It also contains an overview of the annotated benchmarks, which include excerpts of the historical corpora manually annotated with olfactory information, following the guidelines reported in D3.1. This document accompanies the release of the data, made available through the project GitHub page at https://github.com/Odeuropa/benchmarks_and_corpora, which represents the actual deliverable D3.2

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	Sept 10 2021	Definition of document structure and first draft	Sara Tonelli	All
0.1	Sept 10 2021	Added sections on English and Italian and on Quality control	Sara Tonelli, Stefano Menini	2,3
0.2	October 2021	Added sections on Dutch, Latin, and German	Marieke van Erp	2 & 3
0.3	October, 2021	Added sections on Slovene	Inna Novalija	2 & 3
0.4	October 2021	Added sections on French	Raphaël Troncy	2 & 3
0.5	October 2021	Added statistics on benchmark and Conclusions	Sara Tonelli and Stefano Menini	3 & 4
1.0	October 2021	Final check and approval by project manager	Marieke van Erp	-
1.1	February 2022	Minor typo fixes by project manager	Marieke van Erp	-

Executive Summary

Deliverable 3.2. contains multilingual historical corpora and annotated benchmarks in the form of a data release, which we make public through the Odeuropa GitHub repository at https://github.com/Odeuropa/benchmarks_and_corpora. This document provides an overview of the content of the deliverable and to describe the process leading to the collection and annotation of the datasets. The release comprises two multilingual datasets: the first set of documents was collected from various online sources in the seven project languages by retrieving freely available, copyright-free documents covering the period of interest for Odeuropa, i.e. from the 17th to 20th Century. For each document a minimal set of metadata including author, date of publication and source has was collected. These documents represent the sources that we will process throughout the project to automatically extract and model mentions of olfactory information.

The second set of documents includes a sample taken from the first corpus, which has been manually annotated with olfactory information following the Guidelines described in D3.1. The texts to be annotated were selected to cover ten domains defined by cultural historians and to represent the whole time span of interest for the project. The selection was furthermore guided by the presence of olfactory situations and events in the documents, which we analysed by identifying the text excerpts that contain most smell-related seed terms through automated scripts.

The process of corpus creation and data annotation will continue until M18, because our repository can be extended with new documents and annotation can be adjusted or expanded in parallel with the implementation of systems for olfactory information extraction. Therefore, this deliverable presents the first version of the data, which will be regularly updated in the GitHub repository and named according to the different corpus versions.

The release covers the seven project languages, namely English, Dutch, Italian, German, Slovenian, French and Latin. However, the content of the language-specific repositories may vary due to the different availability of digitised corpora. Nevertheless, for each language we were able to create a benchmark with a sufficient number of annotated smell-events and related frame elements (up to 6,523 in Slovenian). For Latin, the dimension of the extended corpus is not comparable with that of other languages, and the selection of texts to be annotated could not cover all ten domains of interest since after 16th Century Latin was used to communicate in specific contexts, mainly scientific ones. Nevertheless, also for this language a fair number of documents and an annotated benchmark with around 1,200 smell events were collected and made available through this release.

Contents

1 Introduction

In Deliverable 3.1. “Annotation Scheme and Multilingual Taxonomy” we described the annotation scheme we devised to model and identify olfactory information in texts. The proposed guidelines follow the FrameNet general approach [Ruppenhofer et al., 2006], starting the annotation from an event-evoking term, which in our case is an olfactory seed word (e.g. *perfume*, *smell*, *stink*) and then identifying the textual spans that describe any participant in the olfactory event. In this Deliverable, we go a step further by presenting the outcomes of the annotation obtained by applying the guidelines to documents in the seven project languages. In particular, we describe the benchmark created by manually annotating olfactory information in all languages, after manually selecting source documents from 1620 to 1920 and covering ten domains of interest defined by cultural historians. This document is meant to be an accompanying text to present the actual deliverable content, which is the multilingual benchmark.

The second component of this deliverable is the extended multilingual corpus, that contains free-from-copyright texts covering the same time period as the benchmark, and that was created for each language. Both components are described in detail in a spreadsheet with all metadata, released together with the documents.

The two corpora serve three main goals: first, they were used to carry out an exploratory investigation of the smell-related information that can be found in historical texts. Indeed, while historians have studied the relevance of specific domains or documents in terms of olfactory heritage, to date, no broad study was carried out aimed at assessing the relevance of olfactory information on a large scale, across different time periods, domains and languages. This practice, inspired by *distant reading* [Moretti, 2013] is made possible thanks to the availability of the large digital corpora collected within WP3.

Our second goal is to guide the development of the Odeuropa system for olfactory information extraction (Deliverables D3.3, D3.4 and D3.6). Indeed, while we do not foresee the implementation of a fully supervised system since it would require a much larger annotated corpus for training, the benchmark and the multilingual corpus make it possible to experiment with AI-based approaches that require less annotated data, such as semi-supervised learning, data augmentation and few-shot learning. The benchmark will be used as a gold standard for evaluating the developed system, or as a small set of examples to be expanded with automated approaches. Also cross-lingual annotation transfer [Tonelli and Pianta, 2008] and multilingual language models [Conneau et al., 2020] could be explored starting from the available benchmark.

The third goal that we will address is the annotation of emotions triggered by olfactory events in texts. To this purpose, having a first annotation layer enriched with high-quality smell information can represent a basis upon which emotions can be added, to guide the following development of a module for multilingual emotion recognition in smell-related texts (Deliverable D3.5).

This document is structured around two main parts: in Section 2 we describe the multilingual historical corpora collected for each of the seven project languages. In Section 3 the benchmarks are presented, describing not only the annotated information but also the workflow for quality control.

2 Description of Multilingual Historical Corpora

In this section, we detail the content of the multilingual historical corpus created for each project language by browsing freely available resources and collecting the ones that were published between 1620 and 1920. As expected, the dimensions of the different corpora variy greatly, depending on the availability of online sources and their quality. Furthermore, this is only the first version of the repository, which will grow throughout the project every time new relevant sources will become accessible. Instructions to access the corpora are available at: https://github.com/Odeuropa/benchmarks_and_corpora. This first set of documents will be used to study the evolution of olfactory terms, situations and events over time, as well as to carry out cross-language comparisons. No manual annotation is foreseen on these datasets.

2.1 Dutch

For Dutch, we first downloaded two main repositories, which are freely available and cover also other languages, i.e. Project Gutenberg[1] and Wikisource[2]. *Project Gutenberg* is a volunteer effort to digitize and archive cultural works, and contains different language-specific repositories, mainly in the literary domain. *Wikisource*, instead, covers a broader set of thematic categories, from arts to mathematics and natural sciences. It is a Wiki-based initiative with the goal to create a library of works either in the public domain or freely licensed.

Furthermore, the the following repositories were added:

- Digitale Bibliotheek voor de Nederlandse Letteren (DBNL):[3] literary, linguistic and cultural historical texts from the 12th to the 21st century;
- Census Nederlands Toneel (Ceneton):[4] Dutch 17th-19th century plays, maintained by Leiden University;
- Delpher:[5] Dutch National Library's digital collection, containing books, newspapers, magazines and radio bulletins. For Odeuropa, we will use the base collection containing ∼130,000 digitised books from the 17th to the 20th century concerning fiction as well as non-fictional subjects: architecture, topography, engineering, biodiversity of the Dutch East Indies, religion, political and societal manifestos, and legislature. We will also use the Google books collection accessible via Delpher containing ∼800,000 digitised out-of-copyright books;
- Early Dutch Books Online:[6] Books from the Amsterdam and Leiden Universities' libraries and the Dutch National Library's special collections dated 1780-1800;
- Amsterdam Notary Records:[7] Amsterdam notary deeds covering the period 1578-1915 concerning the lives of Amsterdam citizens. The deeds contain information about (international) trade, maritime events, slavery, personal possessions and inheritances as well as eyewitness reports of hardships aboard ocean faring ships, neighbourhood brawls, street fights and illegal gambling dens.

[1] https://www.gutenberg.org/

[2] https://en.wikisource.org/wiki/Main_Page

[3] https://www.dbnl.org/

[4] https://www.let.leidenuniv.nl/Dutch/Ceneton/

[5] https://delpher.nl

[6] https://www.kb.nl/organisatie/onderzoek-expertise/digitaliseringsprojecten-in-de-kb/afgeronde-projecten/early-dutch-books-online-edbo

[7] https://alleamsterdamseakten.nl/

2.2 English

For English, we followed the same process to corpus creation that we adopted for Dutch. We first downloaded the two large repositories Project Gutenberg[8] and Wikisource,[9] and then we added a set of curated repositories, whose domains are of interest to the Odeuropa text analysis efforts:

- London's Pulse:[10] Medical Officer of Health reports issued between 1848 and 1925, dealing with public health, sanitation and urban life;
- The Royal Society Corpus:[11] Scientific periodicals issued between 1665 and 1869;
- A small pre-processed version of the larger Old Bailey Papers dataset,[12] containing only the court proceedings published between 1720 and 1913. The complete collection will possibly be included at a later stage of the project;
- The British Library Digitised Books,[13] containing miscellaneous documents, with possible overlaps with other digital repositories (to be checked);
- The Hartlib papers:[14] scientific correspondence by Samuel Hartlib, written at the beginning of the 17th Century;
- Text Creation Partnerships books:[15] collection containing freely available transcripts of early English books, eighteenth Century collections and Evans Early American Imprints collection

2.3 French

Different repositories were consulted to obtain French documents. In particular, we built the French corpus using these freely available resources:

- The ARTFL Project:[16] a repository that contains several digitized French documents, collected in a joint effort between the University of Chicago and the French government. In particular, we focused on the ARTFL Public Databases available on the PhiloLogic4 application for free public use.[17] The available resources are organised in mono-thematic collections, spanning from philosophic essays to collections of French laws, from poetry to Diderot's Encyclopedia. It covers the period from the 16th to the 18th century. We extracted 92,895 texts, belonging to 15 mono-thematic collections;
- Project Gutenberg:[18] The French version of the Gutenberg project contains 4,486 texts, from different time periods and genres;
- Gallica:[19] This is the portal of the French national library providing access to over 200K digitized resources, covering the whole studied period. Resources are available through an API;
- roman18:[20] This is a collection of French Novels from 1750-1800. In total, 113 text document are available;

[8] https://www.gutenberg.org/

[9] https://nl.wikisource.org/

[10] https://wellcomelibrary.org/moh/about-the-reports/using-the-report-data/

[11] http://fedora.clarin-d.uni-saarland.de/rsc_v4/

[12] http://fedora.clarin-d.uni-saarland.de/oldbailey/downloads.html

[13] https://data.bl.uk/digbks/?_ga=2.232539935.1240845631.1613066619-211977171.1611608794

[14] https://www.dhi.ac.uk/data/download/hartlib

[15] https://textcreationpartnership.org/

[16] https://artfl-project.uchicago.edu/

[17] https://artfl-project.uchicago.edu/philologic4

[18] https://www.gutenberg.org/browse/languages/fr

[19] https://www.bnf.fr/en/gallica-bnf-digital-library

[20] https://github.com/MiMoText/roman18

- Finally, some documents from CORPUS17[21] (17th century) and CLEF-HIPE[22] (19th - 20th century) were selected.

2.4 German

For German, we first retrieved the documents published between 1620 and 1925 from Project Gutenberg.[23] and Wikisource[24] Then, we downloaded the following domain-specific corpora:

- Deutches Text Archiv:[25] a large cross-section of printed works in the modern New High German Language, ranging from ca. 1600 to 1900;
- Saarbrücken Cookbook Corpora:[26] a diachronic corpus of cooking recipes containing a historical and a contemporary subcorpus. The historical subcorpus spans 200 years (1569-1729) and includes 430 recipes from 14 cookbooks written in German;
- GeMi Corpus:[27] This corpus contains medical writings from 1500 to 1700;
- GerManC:[28] personal letters, sermons and fictional, scholarly (i.e., humanities), scientific and legal texts from 1650 to 1800.

2.5 Italian

For Italian, three main repositories of online data were consulted: Project Gutenberg,[29] LiberLiber [30] and Wikisource.[31] *LiberLiber* is an Italian initiative led by a non-profit organisation, whose goal is to enable the free circulation of cultural objects. In this framework, around 3,000 books, free from copyright, are shared online. To access them, FBK made a small donation to the association using institutional funds.

From the three repositories, we downloaded and cleaned the documents with a publication date between 1620 and 1925. All downloaded books have "date of publication" in the accompanying metadata, but for Project Gutenberg we noticed that this corresponds to the date of last publication. Therefore, we implemented a script to connect to Google Books APIs[32] and ask for the different versions of the book, and then select the earliest publication date. As additional corpora, we downloaded the following repositories:

- the De Gasperi corpus: [Tonelli et al., 2019][33] a freely available collection of political speeches by the Italian statesman, selecting the documents published till 1925 (1,165 in total);
- the Italian novel collection for ELTeC:[34] this is the European Literary Text Collection, produced by the COST Action Distant Reading for European Literary History;
- the ItaDraCor: corpus,[35] a corpus containing 139 original plays in Italian.

[21] https://github.com/e-ditiones/CORPUS17

[22] https://github.com/impresso/CLEF-HIPE-2020

[23] https://www.gutenberg.org/ebooks/bookshelf/38?sort_order=release_date

[24] https://wikisource.org/wiki/Category:Deutsch

[25] https://www.deutschestextarchiv.de/

[26] http://fedora.clarin-d.uni-saarland.de/sacoco/

[27] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2562

[28] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2544

[29] https://www.gutenberg.org/browse/languages/it

[30] https://www.liberliber.it/online/

[31] https://it.wikisource.org/wiki/Pagina_principale

[32] https://developers.google.com/books

[33] https://github.com/StefanoMenini/De-Gasperi-s-Corpus

[34] https://github.com/COST-ELTeC/ELTeC-ita

[35] https://github.com/dracor-org/itadracor

Overall, the first version of the Italian historical corpus contains around 5,600 books and documents As a next step, we plan first to remove possible duplicates from the corpus, because it is likely that LiberLiber, Project Gutenberg and Wikisource present some overlaps. Then, we will integrate all laws and regulations issued before 1925 and made available through the online portal *Normattiva*.[36]

2.6 Latin

Beside retrieving the documents from Project Gutenberg[37] and Wikisource[38], we download the following corpora for Latin:

- LatinISE corpus version 4:[39] this corpus consists of Latin texts from the 2nd century B.C. to the 21st century. Non-linguistic metadata include information on genre, title, century and specific date;
- The Latin Library:[40] a collection of public domain Latin texts. The Neolatin collection includes philosophy, physics and poetry;
- The Heinsius Collection:[41] a collection of Dutch Neolatin poems and prose texts maintained by Leiden University;
- The Perseus Digital Library:[42] an online collection of texts covering the history, literature and culture of the Greco-Roman world including humanist and Italian renaissance poetry.

2.7 Slovenian

The Slovenian collection contains documents from the IMP and DLib datasets. The IMP collection[43] provides documents from a variety of sources, including:

- WIKI: project Wikivir “Slovenska leposlovna klasika” (“Slovenian literary classics”), containing fiction books, news articles and manuscripts of Slovenian authors;
- FPG: the AHLib collection, with German books (from the period of 1848–1918) translated to Slovenian;
- NUK: older books prepared by NUK (the National and University Library of Slovenia) as part of the IMPACT project;
- KRN: selected excerpts of periodical magazine “Kmetijske in rokodelske novice” (“Agricultural and handicraft news”), NUK/IMPACT;
- ZRC: examples of three religious books (two of these books are the older books in the library), prepared by ZRC SAZU (Research Centre of the Slovenian Academy of Sciences and Arts).

The DLib dataset [44] is built up from the following sources:

- Books (monographs and dissertations);
- Periodicals (historical, scientific, general newspapers and journals);

[36] https://www.normattiva.it/

[37] https://www.gutenberg.org/browse/languages/la

[38] https://la.wikisource.org/wiki/Pagina_prima

[39] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-3170

[40] http://thelatinlibrary.com/neo.html

[41] https://www.let.leidenuniv.nl/Dutch/Latijn/Heinsius.html

[42] http://www.perseus.tufts.edu/hopper/

[43] http://nl.ijs.si/imp/#lexicon

[44] https://dlib.si/?language=eng

- Manuscripts (medieval codices and literary manuscripts);
- Images (photographs, postcards, posters);
- Music (musical scores and audio recordings);
- Maps (maps and atlases).

In the next project phase, we will focus mainly on the IMP and DLib subcorpora containing books and periodicals.

3 Description of Multilingual Benchmarks

3.1 Document selection

With the help of cultural historians on the project, we defined ten domains of interest, where we expected to find a high number of smell-related documents. These domains are: *Household & Recipes*, *Law*, *Literature*, *Medicine & Botany*, *Perfumes & Fashion*, *Public health*, *Religion*, *Science & Philosophy*, *Theatre*, *Travel & Ethnography*. The additional category *Other* was included in the list for the documents which are of interest for Odeuropa but do not fall within any of the previously mentioned categories.

Ideally, the benchmark should contain 10 documents for each category, distributed evently 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 information, while in some other cases a book could contain smell references scattered throughout the volume. Therefore, each annotation team 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 were 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 variety of Latin texts digitally available 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.

3.2 Annotation Setting

For each language, a team of annotators was selected. 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.

Each annotators' team underwent a training session by participating in a tutorial led by FBK, which was also recorded and made available to the consortium.[45] The tool adopted for annotation is INCEpTION [Klie et al., 2018], which was already introduced in Deliverable 3.1. FBK managed the tool centrally so that the task creation and the corresponding interface was the same for all languages. For each team, a curator account was created, enabling the upload of the language-specific documents for the benchmark and monitoring functionalities for the annotation process. Then, an annotator's account was created and assigned to each member of the teams, which worked in parallel at the benchmark creation. Periodical meetings among the members of each team were organised to exchange ideas, discuss doubts and suggest improvements to the annotation guidelines.

[45] Link here: https://drive.google.com/file/d/1d9fwEUCghfABjHJxGguKhbWGhgj1FcvE/view?usp=sharing

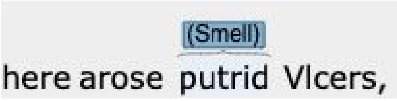
3.3 Quality control

As the annotation process is carried out independently by multiple annotators for each of the project languages, it is important to ensure that the different annotations are consistent. We therefore need a method to check if the way the annotations are made matches with the instructions provided in the guidelines.

To this end, we developed a web-based tool to automatically find when annotations are not compliant with the guidelines. The tool (available at https://dh-server.fbk.eu/odeuropa/) is complementary to the INCEpTION annotation tool, taking as input the exports from the annotation platform to process them.

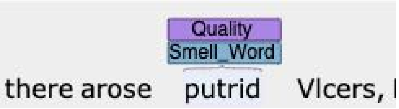
The tool identifies mistakes in the files related to both wrong and missing annotations. The script focuses on errors related to the annotation procedure and not on the content of the annotations. For instance, it checks if every frame element is properly connected to a smell word and if all selected 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 the following five error types, presented with an example on the right side:

Missing Annotation, probably (Smell): Detect spans that have been selected but not labeled. INCEpTION marks them with a generic *'(Smell)'* as label, that annotators should replace with the actual role of the selected text.



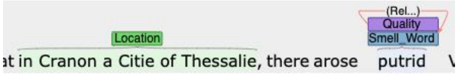
here arose putrid Vlcers,

Smell Word with a double annotation, not linked to itself: 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. In these cases, a relation should be set between the FE label and the smell word. This error notifies the absence of this relation.



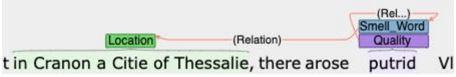
there arose putrid Vlcers,

FE not linked: This error reports frame elements that despite being annotated are not linked to any other element in text.



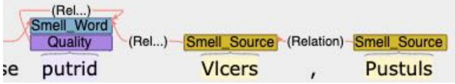
t in Cranon a Citie of Thessalie, there arose putrid

Wrong Relation Direction: The relations between a *Smell_Word* and other frame elements need to point to the *Smell_Word*. This error highlights when a *Smell_Word* is the starting point of a relation instead of the ending point.



t in Cranon a Citie of Thessalie, there arose putrid Vl

Relation error, not directed to Smell Word: According to the guidelines, all frame elements need to be connected to a *Smell_Word* with a relation (except for anaphoric and same_span relations that occurs between two FEs with the same label). This error warns for frame elements connected to something other than a *Smell_Word*.



se putrid Vlcers , Pustuls

After processing the annotated data, the tool returns a report file containing the errors found. Each error is associated with the document, the sentence number and the text string involved. This makes it easy for annotators to go back to the annotation interface, find the mistakes and fix them.

3.4 Benchmark statistics

Below, we report the content of the Odeuropa benchmark in each language. Table 1 shows the number of occurrences of smell words and frame elements. Note that the number of 'Creator' elements is low because this was added at a later stage in the annotations, when several annotators had remarked that they had come across an actor creating a scent but that the annotation guidelines did not facilitate marking this. This frame element is now included in the guidelines, and will be taken into account in future annotations.

	Dutch	English	French	German	Italian	Latin	Slovenian
Smell word	1,929	1,530	664	1,493	1,228	1,199	1,917
Quality	986	1,084	234	100	695	552	959
Source	1,422	1,313	349	17	933	772	1,713
Evoked Odorant	199	91	90	6	71	42	289
Perceiver	314	362	98	31	143	241	291
Effect	238	187	27	29	102	114	217
Creator	28	0	0	0	0	12	1
Circumstances	320	247	79	25	198	192	223
Time	116	126	35	3	119	108	72
Odour carrier	310	310	65	4	187	134	447
Place	215	302	109	3	147	111	394
Total FEs	4,148	4,022	1,086	218	2,595	2,278	4,606

Table 1: Overview of benchmark content for each language

The figures in the table show that, for each language, a good number of smell-related events and situations were annotated. Only the French benchmark contains fewer than 1,000 events, due to issues with the recruitment of annotators. Interestingly, the average number of frame elements (FEs) associated with each smell event varies greatly, going from $<$ 1 FE per olfactory event in German to 2.6 in English. For all languages except German, the most frequent FE is the *Smell Source*, followed by the *Quality*. This shows a trend in the description of olfactory situations that holds across languages and domains. In fact, the description of an olfactory situation tends to include at least where the smell comes from and what are the smell characteristics. The fact that in German, instead, the *Smell Source* is not frequently mentioned may be due to the presence of smell words that already include the source, such as 'Abgasgeruch' ('exhaust smell'), 'Zigarettenqualm' ('cigarette smoke'), 'Regengeruch' ('rain smell'), etc. This may explain also the low number of FEs on average, compared to the other languages.

We report in Figure 1 the number of documents per domain in each language-specific benchmark (see list of domains in Section 3.1). 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 except for Latin, which is characterised by a low variability in the data due to the specific contexts in which Neolatin was used.

In Figure 2 we report the temporal distribution of the documents present in the benchmark for each language. Latin is not displayed because the benchmark data are extracted from eight books only. All languages overlap in the time period of interest for Odeuropa, with the Dutch benchmark including some earlier texts but no data after 1880, and the Italian dataset going beyond 1930. We observe that, due to different data availability, not all time periods are well covered. In the future, we will work towards a better balance of the benchmark in terms both of domains and of decades.

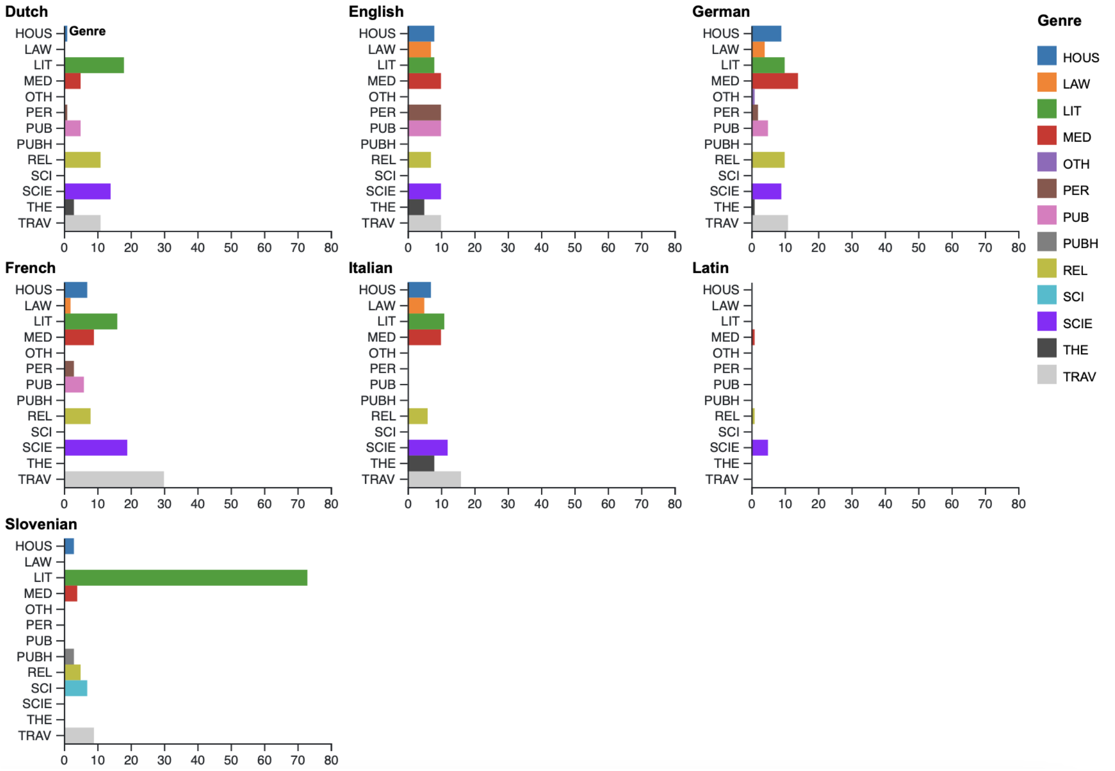


Figure 1: Number of documents per domain in each language-specific benchmark

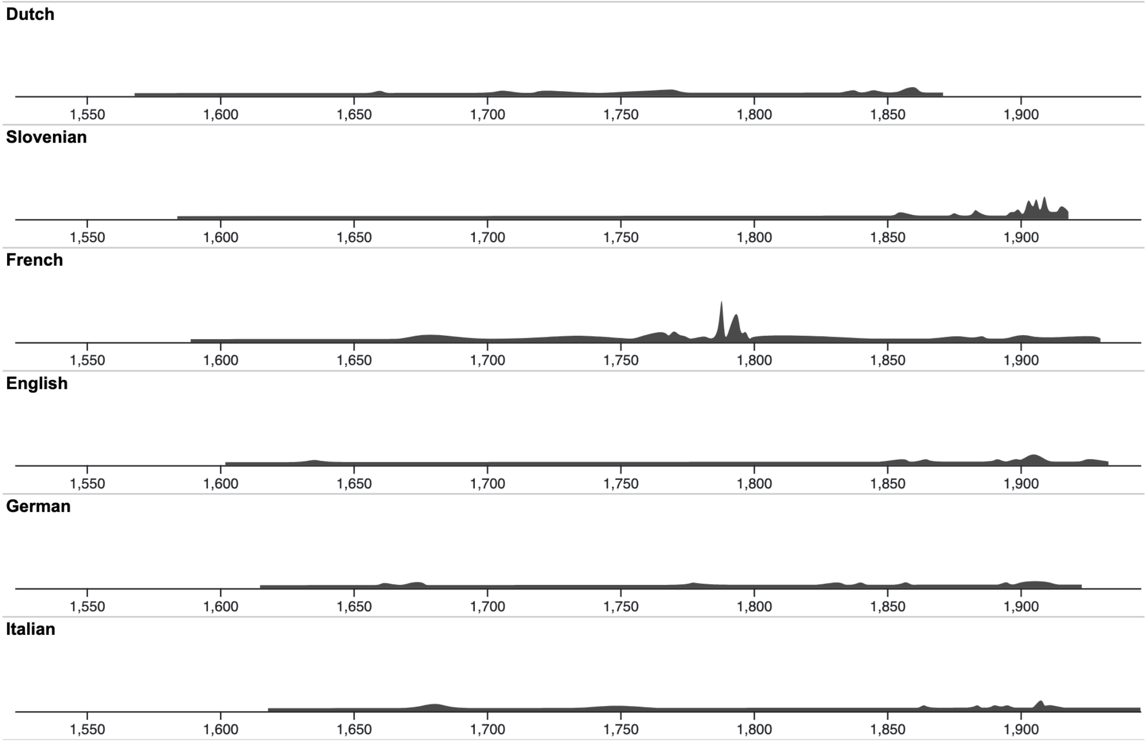


Figure 2: Temporal distribution of documents in each language-specific benchmark

4 Conclusions

In this document, we have presented the content of Deliverable 3.2., namely the multilingual corpora and the Odeuropa benchmark, which has been manually annotated in the seven project languages. The benchmark can be directly downloaded from the GitHub page https://github.com/Odeuropa/benchmarks_and_corpora. As regards the multilingual corpora, instead, they exceed the size limit allowed on GitHub and we have therefore created different repositories, which can be reached from the above GitHub page. We plan to merge them in a single repository, and to update the information on GitHub accordingly.

Both the corpora and the benchmark are likely to be extended in the future, given that novel digital collections will be added and that emotion annotation will be performed on top of the annotation of olfactory information. We will therefore adopt naming conventions so to highlight clearly the version number of the different data releases.

As a next step, we will perform an in-depth analysis of the benchmark content, attempting to automatically infer how olfactory situations have been described over time and what are the main differences among languages and genres. This will inform the future development of the multilingual system for olfactory information extraction. We will also compute inter-annotator agreement and improve quality control measures, to make sure that all annotations are consistent.

References

[Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[Klie et al., 2018] Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (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.

[Moretti, 2013] Moretti, F. (2013). *Distant reading*. Verso Books.

[Ruppenhofer et al., 2006] Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2006). Framenet ii: Extended theory and practice.

[Tonelli and Pianta, 2008] Tonelli, S. and Pianta, E. (2008). Frame information transfer from English to Italian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

[Tonelli et al., 2019] Tonelli, S., Sprugnoli, R., and Moretti, G. (2019). Prendo la parola in questo consesso mondiale: A multi-genre 20th century corpus in the political domain. In Bernardi, R., Navigli, R., and Semeraro, G., editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.