Annotation Scheme and Multilingual Taxonomy

Deliverable D3.1

Version FINAL

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 D3.1
Status & version	FINAL
Contractual date of delivery	June 2021
Date of delivery	30 June 2021
Type	Report
Security (distribution level)	Public
Number of pages	28
WP contributing to the deliverable	WP3
WP responsible	WP3
EC Project Officer	Hinano Spreafico
Authors: Sara Tonelli[1], Stefano Menini[1], Serra Sinem Tekiroğlu[1], Inna Novalija[2]	
Internal reviewers: Daniel Schwabe[2], Inger Leemans,[3] Marieke van Erp[3]	
Affiliations (1) FBK, (2) JSI, (3) KNAW	
Keywords: annotation guidelines, taxonomy building, multilingual resources	
Abstract: In this document, we describe the guidelines proposed for annotating olfactory information in Odeuropa multilingual benchmarks. The guidelines include both the conceptual framework developed for the annotation of olfactory situations, inspired by Fillmore's frame semantics [Fillmore and Baker, 2001], as well as details on how the annotation workflow has been implemented, starting from smell-related mentions and then extending the annotation to the involved participants. We also show how the INCEpTION tool has been set up to carry out manual annotation in all project languages. In the second part of the deliverable, we describe the process that led to the creation of the multilingual olfactory taxonomy, which starting from a core set of manually selected terms connected to WordNet has been expanded using n-grams and clustering via word embeddings.	

Table of Revisions

Version	Date	Description and reason	By	Affected sections
0.1	Mid May 2021	Definition of deliverable structure	Sara Tonelli	All
0.2	June 10	Description of taxonomy creation process	Stefano Menini, Serra Sinem Tekiroğlu	3
0.3	June 10	Description of annotation guidelines	Sara Tonelli	2
0.4	June 17	Guidelines for emotion annotation added	Inna Novalija	2
0.5	June 18	Finalisation of deliverable content, Executive summary added	Sara Tonelli	All
0.6	June 21	Deliverable revised, some sections restructured, typos fixed	Daniel Schwabe	All
0.7	June 23	Deliverable reviewed, some suggestions from history perspective made	Inger Leemans	All
0.8	June 23	Final version of deliverable revised and sent to project manager	Sara Tonelli	All
1.0	Wednesday 30th June, 2021	approval by project manager	Marieke van Erp	-

Executive Summary

The present document provides a description of the Odeuropa guidelines developed to manually annotate olfactory information in all the project languages, as well as of the process implemented to create the multilingual olfactory taxonomy. This work will guide the subsequent annotation task, which will lead to the creation of a multi-domain multilingual benchmark of historical texts annotated with olfactory information, due in month 9 and presented in D3.2. The guidelines and the benchmark represent also the backbone upon which the automated system for olfactory information extraction will be built.

The guidelines were inspired by frame semantics [Fillmore and Baker, 2001], a theory which has been implemented also through the FrameNet annotation project [Ruppenhofer et al., 2006], whose goal is to capture situations and events present in texts. These are modeled as a set of semantic roles or *frame elements*, which are typically the participants in the event, all connected to a *lexical unit*, i.e. the textual anchor that triggers the event or situation.

FrameNet aims at being a general-purpose resource, capturing all possible situations and events that may happen in real life. For Odeuropa, we adapt them to the olfactory domain by adopting the same structure based on lexical units and related frame elements, but we create domain-specific semantic roles that are the outcome of discussions with project partners who are experts in olfactory heritage and history. For example, we introduced the roles of *Smell source*, *Evoked odorant* and *Odour carrier*, while we borrowed from FrameNet some generic roles such as *Perceiver*, *Time*, *Place* and *Circumstances*.

With the help of domain experts, we also defined a basic list of seed terms, i.e. smell-related lexical units that evoke olfactory situations and events. These are first translated in all project languages (English, Dutch, Italian, French, German, Slovenian, Latin) and enriched with language-specific terms. Then, they are used both as a first list of lexical units, and as a core list of concepts upon which the multilingual olfactory taxonomy is built.

In this document, we report the list of seed terms and examples of annotated roles in multiple languages. We also detail some annotation conventions that we have adopted, deviating from the FrameNet standard. Additionally, we describe the use of the INCEpTION tool [Klie et al., 2018], which was selected to support the manual annotation process because it is highly flexible, supports multiple languages, enables an easy quality check and is extremely user-friendly.

Concerning the multilingual olfactory taxonomy, whose development is described in the second part of this document, it has been created following a semi-automatic approach: starting from the olfactory seed terms, they have been first mapped to WordNet synsets and manually checked, resulting in a first core version of the taxonomy relying on the availability of WordNet in multiple languages. In a subsequent step, an automatic expansion has been carried out by looking for terms that more frequently co-occur with the core concepts within n-grams, which are sequences of five words present in large collections made with extracted words from books and large digital archives. For English, Italian, French and German, we took advantage of the availability of the n-grams collections released by Google and extracted from Google Books, which also come with part-of-speech tags and the time period where the n-gram was found.

For Dutch, we resorted to other n-gram repositories extracted from newspapers published between 1600 and 1995 [De Goede et al., 2013], which however do not contain any part-of-speech information. For Slovenian, we used another repository of n-grams extracted from the IMP corpus of historical Slovenian [Dobrovoljc, 2018], which however does not contain publication dates for each n-gram nor part-of-speech information. For Latin, we could not find any set of available n-grams suitable for this task.

These differences among languages have led to taxonomies in the seven project languages with a highly variable coverage for each time period and different dimensions. Nevertheless, the fact that for all languages we started from WordNet guarantees that at least the core part of the taxonomy is consistent across the languages and of high quality.

The first version of the multilingual taxonomy described in this document is available at https://github.com/Odeuropa/multilingualTaxonomies, together with the scripts implemented to map and

extract information, as well as with links to n-grams repositories to replicate the process.

Since in the next months we will create large repositories of texts covering different domains and published between 1650 and 1925 to be processed with our information extraction pipeline, we will use them also to create additional domain-specific n-grams, so that the taxonomy expansion process will be repeated on cleaner and more comparable data. Subsequent versions of the taxonomy will also be shared on the same github page.

Contents

1 Introduction

One of the objectives of the Odeuropa project is to use Artificial Intelligence (AI) to extract olfactory information from large archives of texts and images and to build a knowledge graph starting from this information. For text processing, which is the main focus of WP3, this will be performed starting from large volumes of data in the 7 languages of the project, namely Dutch, English, German, French, Slovenian, Latin and Italian. We also aim to cover documents published between 1600 and 1925. This task poses a number of challenges that are not only technical, but also conceptual, because no framework to model olfactory information in texts and capture their mentions has been proposed so far.

To address this challenge, the first activity in this direction is the definition of *what* we aim to annotate in texts and *how* we are going to perform the task. This activity requires interdisciplinary knowledge because the expressions used to refer to smells, the way in which the related emotions are described and the olfactory situations change over time. Furthermore, differences among domains (e.g. medicine, science, fiction, travel writings, etc.) should be taken into account when defining the annotation scheme. We therefore involved the project partners with in-depth expertise in olfactory heritage and history (KNAW and ARU) in the definition of these guidelines from the beginning, starting from the identification of a set of smell-related core concepts to extensive discussions on the challenges of emotion annotation in olfactory situations.

Concerning the annotation guidelines, we rely on the linguistic framework called *frame semantics* [Fillmore and Baker, 2001], a theory which was implemented through the FrameNet annotation project [Ruppenhofer et al., 2006], whose goal is to capture situations and events mentioned in texts. These are modeled as a set of semantic roles or *frame elements*, which are typically the participants in the event, all connected to a *lexical unit*, i.e. the textual anchor that triggers the event or situation.

For Odeuropa, we propose an adaptation of FrameNet to the olfactory domain, where only situations related to smell are annotated and specific semantic roles connected to olfactory events are identified. In doing this, we pursue several objectives: we want our annotation guidelines to be generic enough to accommodate all possible smell-related events which may be mentioned in a text, and also to cover all languages, without the need to provide language-specific adaptations from a semantic point of view. We also want our framework to be flexible enough to deal with meaning change of terms over time, enabling annotators to add new smell-related terms to our initial list during the annotation process. Finally, we define semantic roles using labels that are self-explanatory and not ambiguous (e.g. Evoked odorant, Smell source), so to facilitate the role selection ensuring a good agreement among annotators.

The second contribution of this deliverable is the multilingual taxonomy of olfactory information, which represents the starting point for the self-supervised approach to information extraction from texts. This has been developed in parallel to the taxonomy of olfactory phenomena in images, created within WP2 and also due in month 6. Indeed, the multilingual taxonomy of WP3 has a textual focus, being centered around smell-related terms, and has been created starting from a core set of terms of WordNet [Miller, 1995] and then automatically expanding them via n-grams and word embeddings [Grave et al., 2018]. The fact that WordNet represents the core of the multilingual taxonomy makes the text-based and the image-based taxonomies easier to integrate, for example through ImageNet [Russakovsky et al., 2015], which is a widely-used taxonomy of images based on the WordNet conceptual structure and which is also used in WP2.

This deliverable is structured as follows. In Section 2, we describe the annotation guidelines, providing details on how to identify smell-related events and the corresponding semantic roles in texts in multiple languages. We also introduce our proposal for the annotation of emotions in texts, and we describe briefly the annotation tool we have adopted. Section 3, instead, is focused on the multilingual taxonomy. We first describe the process combining manual lemma selection and automated expansion, and then we provide details on the final taxonomies that we make publicly available. Finally, we draw some conclusions in Section 4.

2 Annotation Guidelines

In this section, we describe the two layers of annotations that we aim to add to texts in multiple languages, namely the *olfactory events* and the *emotions* they evoke. The two types of information are meant to be added in sequence: we will first ask annotators to identify olfactory events and their participants in texts, and then to mark whether the same text is connected with some emotions. We detail the two steps in Section 2.1 and 2.2 respectively.

2.1 Annotation of Olfactory Events

The scheme for the annotation of olfactory events and situations is inspired by FrameNet framework, whose goal is to annotate situations (or events) and its participants.[1] In FrameNet, events and situations are so-called *frames*, which from a cognitive point of view, are defined as components of the internal model of the world that a language-user has created by interpreting his/her environment [Fillmore, 1976]. Frames are used as synonyms for schemata, semantic memory or scenarios, and represent the perceptual base of our knowledge that is necessary to understand the meaning of words. For example, the predicate 'marry' refers to a scenario where two partners get involved in some kind of social relationship.

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. FrameNet includes few frames defined for smell-related lexical units. However, they make very fine-grained distinctions among similar situations, for example the *Perception active* and the *Perception experience* frames both include 'smell' as lexical unit, but distinguish between the intentional perception of an experience and the unintentional one. Since this distinction is not relevant for Odeuropa, we merge several smell-related frames into a single one, which we call *Olfactory event*.

2.1.1 Annotation of Lexical Units

The *Olfactory event* frame is typically evoked by 'smell' words, i.e. terms (different for each language) that unambiguously evoke or describe an odor-related situation or event. Using the same terminology of FrameNet, smell words correspond to so-called lexical units. These include nouns, verbs, adverbs and adjectives. To create a list of smell words, we asked Odeuropa domain experts to prepare a list of possible lexical units for each language, which is reported in Table 1. This list can be further expanded or reduced after the first annotation tests. The **Other** category covers (ambiguous) smell words, that should be annotated only if they refer to a smell experience or an olfactory situation.

Each olfactory situation can be evoked only by one smell word. If more smell-related terms are present, they have to be annotated as different frame instances. If a smell word can correspond also to a frame element, it has to be annotated as a smell word if no other smell words are found in the sentence. Consider for example the following sentence:

> The air in the room was mephitic.

The term 'mephitic' would be a quality, but since no other smell-word is present in the sentence, it is clearly this term that evokes the smell event, and it should therefore be marked as lexical unit (this is also part of the seed words listed in Table 1).

If a text clearly describes an olfactory situation but smell words from the pre-defined list cannot be found, other terms can be considered as lexical units if they are used as near-synonyms of

[1] FrameNet annotation guidelines: https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

'smell', even if out of context they may have another meaning. See for example 'composition', 'essence', 'perceive', etc.

2.1.2 Annotation of Frame Elements (FEs)

Each odor-related event or situation is evoked by a smell-word but can involve one or more participants in the event, each having a specific role. These olfactory frame elements (i.e. semantic roles) have to be identified and annotated as well. The list of participants or semantic roles pertaining to an odour-related situation is the following. Note that the smell term is underlined, while the frame element is reported between brackets:

Smell source: the person, object or place that has a specific smell. It can also refer to (non)human/ object that produces an odour (e.g. plant, animal, perfume, human). This FE (between brackets) is the entity or phenomenon that the perceiver experiences through his or her senses.

EN: The waiter smelled [something] foul in the kitchen
He carefully smelt [the soup] for any trace of poison
The peaceful odour [of Mrs. Dillon] was prevalent in the hall of the house.
The place was strongly scented [with lavender water].
ITA: Alfredo sentì improvvisamente un odor [di salami].
FR: O [fleurs] ! Que ton odeur éclate devant la bien-aimée
SLO: Janez je zavohal [nekaj] čudnega v kuhinji.
NL: [De kamer] stonk heel erg naar rook.
DE: Daß diese [Falschheit und Tod] an den falschen Zeugen gerochen werden müsse.

Odour carrier: this FE corresponds to the carrier of an odor, which can be either an object (e.g. pomander, bottle of perfume, handkerchief) or atmospheric elements like wind and air. Note that the Odor carrier has a different role from the Source, since the Source produces an odour, while the Carrier carries a smell that is produced by something else (possibly unknown). A Carrier should be annotated only when there is a clear distinction w.r.t. the Source. When this distinction cannot be understood or inferred from the text, a Source label should be selected. This means that when an odour is described as coming generically from an object or entity and it is not specified or clear from the context whether the object or entity actually produced the odour it should be annotated as Source.

EN: [An Egyptian gale] came rushing upon me, impregnated with pestilential vapours.
[The pomander] emitted a smell of musk.[2]
ITA: Il veleno esalava l'anima perversa nel profumo bizzarro che riempiva [l'aria]
FR: [L'air] était embaumé du parfum qu'elles exhalent sur le soir
SLO: [Veter] je po soseski širil vonj pokošene trave.
NL: [De bloemen] geurden heerlijk zoet in ons huis
DE: Wiederumb wenn [die Luft] unten frey durchstreichen kan / so führet sie die Dünste / welche aus dem Unflat von unten aufsteigen / und den Gestanck verursachen / mit sich weg.

Quality: This is a quality associated with a smell and used to describe it. For example rancid, fresh, etc. This is typically expressed by qualitative adjectives. It is often preceded by an intensifier such as "very, really". The intensifier has to be annotated with the related adjective in the same span. Qualities include intensity (not perceptible, weak, distinct, strong), volume / reach (far reaching), duration (lasting, permanent), state (old, deteriorated), character (humid, dry, garlicky,

[2] In this example 'of musk' would be Source

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, Reukorgaan, Reukstof, Reukwater, Reukwerk, Reukzin, Riecking, Ruiker, Snuf, Stank, Stinkbok , Stinker, Stinkerd, Stinkgat, Stinknest, Vunsheid, Waesem, Walm, Deodorisatie **Verbs**: Aromatiseren, Deodoriseren, Geuren, Meuren, Neuzen, Ontgeuren, Opsnuiven, Parfumeren, Rieken, Ruiken, 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, Weeïg, Welriekend, Zwavelig **Adverbs**: neusgierig, neuswijs, neuswijsheid, 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
Latin
Nouns: fetor, graveolentia, mephitis, putor, virus, fumus, nidor, odor, odoratio, odoratus, olfactus, spiritus, aura, fragrantia, beneolentia, suaveolentia, aroma, oleo, praeoleo **Verbs**: olfacio, odefacio, odoror, sentio, feteo, peroleo, puteo, fragro, oboleo, oleo, praeoleo, redoleo, sapio, olfacio, exhalo, halo, spiro, inodoro, odoro, suffio, vaporo **Adjectives**: fetidus, graveolens, rancidus, rancens, puter, putidus, virosus, olens, olidus, beneolens, fragrans, odorarius, odoratus, odorifer, odorus, radolens
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 project language

fruity, woody), hedonic characteristics (malodorous, aromatic, healthy). For specific cases where a perfume is described and is referred to with a proper name, such name is also annotated as Quality (see example 4 below):

EN: The coffee had a [pungent] smell
The waiter smelled something [foul] in the kitchen
[Pleasing to Western noses]
[Two Jubilaeum] scents
It was filled with an odour [hard to conceive]
ITA: Aleggiava nell'aria un profumo [tenue]
FR: Son parfum est en même tems si [fort] & si [agréable]
SLO: [Sveži vonj] vrtnic jo je spremljal, kamorkoli je šla.
NL: Durians hebben een [sterke], [ranke] geur die de buitenste schil doordringt.
DE: So z. B. erzählt Pausanias,nter den Phocäern die Ozolen, eingeborne Völker von Lokris, wegen der Eigenheit der Luft [durchaus übel] riechen.

Perceiver: The being that perceives an odour, who has a perceptual experience, not necessarily on purpose. The perceiver is mostly a person or an animate entity. The perceiver can also be expressed by mentioning the perceptive organ (e.g. nose, nostrils, nerves) used in the olfactory experience

EN: The olfactory [nerves of women of quality] are amazingly tender.
They have the old smell which [to me] would bring back Knole.
[The waiter] smelled the milk to see if it was fresh
ITA: [Alfredo] sentì improvvisamente un odor di salami.
FR: Car [je] sentais l'odeur de l'huile et du suif qui m'infectait
SLO: Puloverji iz omare so [mi] še vedno dišali po babici.
NL: [Diana] rook de rat.
DE: [Cap. Antonius] ziehet seine Armee weil [er] des Julius Caesars todt nicht gerochen.

Note for English annotation: Attention should be paid to the difference between "She smelled like a flower" and "She smelled an intense odour of wine". In the first case, "She" is the Smell source of the smell, in the second case "She" is the Perceiver. To distinguish between the two senses in English it is important to note the different constructions: PERSON smell like vs. PERSON smell OBJECT.

Evoked Odorant: This frame element describes the object, place or similar that is evoked by the odour, even if it is not visible in the scene. In English, this is often part of a comparison or similarity using the verb "to smell" and introduced by "like". Evoked Odorants also include situations, recollections or abstract concepts that are evoked in the Perceiver's mind by smelling an odour.

EN: He smells [like flowers].
I have no desire to reek [like the floor of a florist's stall]
Odour [of sanctity]
ITA: Andando alla posta, trovai una lettera profumata [come una scatola di canfora].
FR: ...cette odeur qu'il croit être, mais que cette manière d'être lui est occasionée par l'impression de quelque objet extérieur, de quelque [fleur], par exemple
SLO: Iz čajnika je dišalo [kot da bi bila na cvetočem vrtu].
NL: De geur [van roestig ijzer].
DE: Zu andern Eigenheiten meiner Natur gehört auch die, daß mein Fleisch bisweilen [wie Schwefel und Weihrauch] riecht.

In some cases, there is no linguistic evidence that a frame element is an *Evoked Odorant* and not an *Smell source*. The former is usually evoked in the mind of the perceiver, often as a

recollection of past experiences, while the latter is the actual entity emitting an odour. In these cases, annotators need to interpret the whole sentence and infer if the smell source is real or only evoked, as in the following example where 'nothing but painted paper and tinsel' would be Odour source:

> The horrid stench [of the leek] was composed of nothing but painted paper and tinsel.

Location: This frame element describes the location where the smell event takes place. Locations can include both named places (for example names of cities) and common nouns describing locations such as garden, street, kitchen, cliffs, promenade, neighborhoods, etc. Similar to the annotation of Odour Carrier, Locations are to be marked only when they are different from the Smell source. Otherwise, if it is not possible to distinguish whether a place produces a smell or is just impregnated by it, the more generic Smell source label is preferred.

> **EN**: [In Venice] the canals have an offensive smell.
> The peaceful odour of Mrs. Dillon was prevalent [in the hall of the house].
> [The place] was strongly scented with lavender water.
> **ITA**: [Sulla terrazza] alitava una frescura impregnata di selvaggi profumi campestri.
> **FR**: Une puanteur dangereuse qui s'exhaloit [des canaux dont la ville étoit traversée] en a fait diminuer le nombre
> **SLO**: Vonj po volnenih odejah je prevladoval [v vseh prostorih graščine.]
> **NL**: De bloemen geurden heerlijk zoet [in ons huis].
> **DE**: Dann je mehr die groben eingesaltzene Fische stincken (wie es denn offt wegen mangel des Saltzes / so sie daran sparen / geschiehet) je lieber jhn mancher käufft / dahero kan man [jhren Fischmarckt] ehe riechen / als man jhn sihet oder betritt.

Example of annotation of a place as Source:

> They kneel on the wet flags of this foetid [cave]$_{Source}$.

Time: an expression describing when the smelling event occurred. It includes expressions of duration, frequency and point(s) or period(s) of time.

> **EN**: [In Summer] the city was inundated with a pungent fish smell.
> [By day] they have little or no smell except in rainy weather, but [in the evening] they are delightfully fragrant.
> **ITA**: [Due sere dopo], sulla terrazza alitava una frescura impregnata di selvaggi profumi campestri.
> **FR**: L'air était embaumé du parfum qu'elles exhalent [sur le soir]
> **SLO**: [V času njegovega vladanja] se je od gradu širil vonj po razpadajočih truplih.
> **NL**: Die geur is precies hetzelfde als [pakweg veertig jaar gelee].
> **DE**: Wann der Zunder [Tag und Nacht] in Essig geleget / und dann wieder abtrocknet wird / soll er gar wenig riechen

Circumstances: This frame element describes the state of the world under which the smell event takes place. Note that this does not include places and temporal expressions, which should be annotated as Location and Time respectively. The role can describe causal implications that lead to or influence the smell event. Circumstances may also describe bigger events that include the smell event, for example historical (named) events. Annotators should first try to assign to the FE a Time or Location role and, only if it does not apply to the specific case, resort to Circumstances.

EN: [The alteration] it made in him would sometimes fill the room with a musty scent [At high heat] the smell of mud was pungent.
ITA: [Nel mio dolore], le soffiavo in volto una ondata di profumo.
FR: On l'y emploie à parfumer les habits, et même [dans les grandes occasions]
SLO: [Pod velikim pritiskom] začne tvarina oddajati toksični strup.
NL: [Zodra mijn vader terugkwam], rook hij die vreemde geur.
DE: Wann der Zunder Tag und Nacht [in Essig geleget / und dann wieder abtrocknet wird] / soll er gar wenig riechen.

Effect: This frame element describes an effect or reaction caused by the smell. This can include entire sentences or clauses describing another event, that is not necessarily a smell event. This can include also the description of emotions triggered in the Perceiver by the smell event.

EN: The smell of mud was in my nostrils, [the high stillness of primeval forest was before my eyes].
ITA: Al grave odore di solfo, ai densi volumi di fumo [le donne e i più timidi cominciarono a fuggire].
FR: les plus gros arbres ont été tellement perfumés que [les oiseaux les moins délicats ne s'y reposoient plus]
SLO: Zavohal sem sveži kruh [in pred očmi so se mi odvrteli prizori iz otroštva.]
NL: De indringende geur van bladaarde [verdreef de andere geuren].
DE: Ich habe einen Jacobiner-Mönch von Edelen Geschlechte zu Venedig gekennet / wenn er eine Rosen roch / oder von ferne ihr gewahr wurde / [ward ihme das Hertze matt / und fiel ohnmächtig zur Erden] / daß er / wie todt / blieb liegen.

Each semantic role may be present or not, or be expressed through multiple instances, if they appear separately in the text. For example, in the following sentence both "the excrements of animals" and "their sweat" are to be labeled as *Smell source*. However, if the elements corresponding to a FE are included in a list of consecutive elements, they can be annotated as a single frame element without splitting the elements in the list.

Usually, [the excrements of animals]$_{Smell\_source}$, and in particular [their sweat]$_{Smell\_source}$, are faetid.

For an olfactory event to be present, however, at least a smell word should be annotated.

2.1.3 Annotation Conventions

Following FrameNet annotation practice, we annotate whole constituents that realize frame elements relative to our smell words, rather than just tagging the head words of these constituents. That is, we work with a phrase structure grammar, rather than a dependency grammar. A consequence of this is that many frame element labels cover words that have no direct relation of their own to the target, but only to the head of their constituent. For instance, when a frame element is expressed by a noun which takes adjectival, prepositional or clausal complements or which is modified by such elements, these complements and modifiers are included in the frame element tag. This means also that, when a frame element is expressed by a noun phrase, also articles at the beginning of the phrase are labeled (see below example 1). The same holds with prepositions at the beginning of a prepositional phrase (example 2). When a frame element includes a relative clause, the latter is annotated as well.

Usually, [the excrements of animals]$_{Smell\_source}$ are faetid.
The peaceful odour of Mrs. Dillon was prevalent [in the hall of the house]$_{Location}$.

In some cases, the same FE label appears multiple times relative to a given target. This can apply to two cases: *i)* multiple separate instances of the same frame element, for example when several Smell sources are mentioned for the same smell event; *ii)* a single instance of a frame element, which is realized in two discontinuous pieces rather than as a single constituent. The cases of discontinuous FEs are particularly frequent in languages that foresee separable terms, for example German split verbs. Both segments have to be annotated and tagged with the same label, and a relation has to be specified going from the peripheral elements to the head of the constituent. For example, in case of separable verbs, a connection should go from the suffix to the main (conjugated) verb. For cases pertaining to *i)*, instead, the same FE label can be assigned to different text segments.

Concerning negated events or texts describing the lack of smell events, we annotate them as if they were standard smell events. This is because we are interested in understanding how smells (or a lack thereof) were described at scale and in extracting odor-related terminology, not so much in distinguishing whether a specific event description refers to the presence or absence of a smell. Therefore, the following examples would be annotated ignoring the presence of negation:

> Surprisingly, [the lady in question]$_{Perceiver}$ had not the [most powerful]$_{Quality}$ scent [of the onion]$_{Smell\_source}$ [in her delicate nostrils]$_{Perceiver}$.
> But the aroma cannot be described as [a floral emanation]$_{Evoked\_odorant}$.

Since we are also interested in metaphorical use of smell-related expressions, we include them in our annotation. This decision is in line with past annotation efforts related to emotions in Dutch texts from 1600 to 1800: no distinction was made between the references to body parts or bodily processes in a literal and a metaphorical sense because such a distinction is often quite difficult to make in early modern texts, where expressions which we now consider to be metaphorical often had a quite material basis in humoral theories of the passions [Leemans et al., 2017].

2.1.4 Deviations from FrameNet annotation

In FrameNet only one relation type is foreseen, that is the relation connecting a lexical unit and the different frame elements. We consider this relation type to hold also for Odeuropa as the standard (unmarked) relation, which should go from each annotated frame element to the related smell word. This relation should be explicitly marked also via the annotation tool, so as to avoid ambiguities when several smell words are present in the same text passage (see Section 2.3 for details). However, we introduce two additional relation types. One is used to mark discontinuous frame elements, roles that are expressed by two or more non contiguous strings. Through this relation, strings of text belonging to the same FE can be connected. We create a **Discontinuous** relation oriented towards the head or governor of the FE. For example, in the case of a separable verb, the relation will be directed from the prefix to the verb root.

The second relation type is **Anaphoric** and it is not present in FrameNet annotation. It was introduced to annotate cases where a frame element is a pronoun and its antecedent is mentioned in discourse (usually preceding the pronoun). In this case, we would annotate the pronoun as bearing a FE label (for example, *Smell source* or *Perceiver*) but we would also manually mark the string explicitly stating whom the pronoun refers to. To this purpose, the annotation would include the following steps:

1. Mark the pronoun with the FE label (for example *Perceiver*)
2. Mark the antecedent with the same FE label as the pronoun it resolves (for example, *Perceiver*)
3. Set a connection going from the antecedent to the pronoun and assign to the relation the *Anaphoric* label

This annotation is reported below with *Perceiver1* as the antecedent and *Perceiver2* as the pronoun. In the annotation tool, an arrow has to be set connecting the first to the second element, and the label *Anaphoric* should be selected.

There was no persuading [the lady in question]$_{Perceiver1}$ that [she]$_{Perceiver2}$ had the [most powerful]$_{Quality}$ scent.
He detected [persons of impure life]$_{Smell\_source1}$ by [their]$_{Smell\_source2}$ smell.

2.2 Annotation of Emotions

Within WP3, task T3.4 concerns multilingual emotion recognition for olfactory information and focuses on extracting emotions related to smells and scents. One of the goals for this task is to develop methods that would enable analysis of differences in odor-related emotions between languages and over time. The results of this task will be used to enrich the Odeuropa knowledge graph.

For this first version of the emotion annotation guidelines, we will focus on basic emotions found in past literature. However, other approaches based on the analysis of 'feelings' rather than 'valence' or 'emotions' [Delplanque et al., 2017] will be considered after the first annotation tests if the proposed repository of emotions turns out to insufficiently cover the olfactory domain or to yield low inter-annotator agreement.

In this task, the project partners combine a number of approaches and views on the emotion detection topic. In particular, we will rely on top-down approaches, based on opinions of historical and olfactory experts within the project, as well as a bottom-up methodology, based on data-driven results and machine learning. To enable the second approach, we first need to define an annotation scheme related to emotions, and then to manually annotate odor-related information in texts to validate such scheme and create a benchmark for the evaluation of automated emotion detection approaches.

The definition of annotation guidelines for this task is particularly challenging, because several frameworks to define emotion categories and hierarchies have been defined starting from Aristotle [Plamper, 2015], and spanning the philosophical, scientific and even medical domains. Below we present an expert overview of the related work in the domain of emotion classification, which would lead us to the formalisation of the Odeuropa emotions annotation guidelines.

Passion	Aristotle	Stoa	Descartes
Envy	Basic		
Fear	Basic		
Desire	Basic	Basic	Basic
Anger / wrath	Basic		
Sadness / pain		Basic	Basic
Hate	Basic		Basic
Love	Basic		Basic
Lust		Basic	
Pity	Basic		
Fear		Basic	
Joy	Basic		Basic

Table 2: List of basic passions

[Plamper, 2015] gives an introduction to the topic. Historical emotion classifications usually list: Aristotle, Plato, Stoa, Galen, Augustine, Aquinas, Descartes, Spinoza, Hobbes, Lock, Shaftesbury, Hutcheson, Hume, Rousseau, Kant, Freud, and then jump to more modern classifications in anthropology, biology, psychology, neuroscience and computational linguistics. The author attempts to summarise the different historical perspectives on emotions/passions. One major issue is the fact that *emotion* both as a word and a concept of strong, inner feeling triggering action, materialised in the 18th-19th Century. In the early modern period *passions* were also perceived to come from outward-inward, and were connected to physical states by humor theory. This is important for Odeuropa since it indicates that smell and emotions are strongly connected (therefore historians like to use the word *affect* since it entails both embodiment and feeling) [Leemans et al., 2017]. Another issue is that emotions have been perceived as cognitive

instruments. Whereas we in our modern world have learned to tie smell to emotion and frame motion as opposite to cognition, early-modern thinkers would disagree. They would be of the opinion that smell and sentiments are both "ways of knowing", of assessing and valuing the world around us.

To classify emotions, philosophers and moral theorists of pre-modernity have organised emotions according to *positive* / *negative* (e.g. pain / pleasure), following Plato, and *intensity* / *duration*, following Aristotle. They also made lists of basic passions, from which other passions and feelings were derived, see for example Table 2 where Aristotle's, Stoa's and Descartes' theories of basic passions are compared. Modern classification schemes have been developed along the same line, displaying the same kind of diversity. [Ortony and Turner, 1990], for example, collated a wide range of research on identification of basic emotions (Table 3).

Theorist	**Basic emotions**
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Arnold	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman, Friesen, and Ellsworth	Anger, disgust, fear, joy, sadness, surprise
Gray	Rage and terror, anxiety, joy
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
James	Fear, grief, love, rage
McDougall	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mowrer	Pain, pleasure
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Expectancy, fear, rage, panic
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love, rage
Weiner and Graham	Happiness, sadness
Frijda	Desire, happiness, interest, surprise, wonder, sorrow

Table 3: Modern classification of basic passions

Concerning data-driven approaches to emotion detection in texts, there have been several attempts to formalise them and to annotate texts accordingly. An overview of the most popular methods, highlighting the advances, challenges and opportunities, has been presented in [Acheampong et al., 2020]. The authors mention a number of datasets and models used in emotion detection, which we summarise in Table 4.

Taking into account the existing studies on emotion recognition and their historical perspective, Plutchik's wheel of emotions is investigated as an initial reference for emotion annotation in Odeuropa (Figure 1). [Plutchik, 1980] proposes a three-dimensional model that is a hybrid of both basic-complex categories and dimensional theories. It arranges emotions in circles where inner circles are more basic and outer circles more complex.

Our proposal for annotating Odeuropa texts would rely on the following main emotions, which are largely inspired by Plutchik's wheel but present some relevant changes:

- Joy (Ecstacy → Joy → Serenity)
- Fear (Terror → Fear → Apprehension)
- Surprise (Amazement → Surprise → Distraction)
- Sadness (Grief→ Sadness → Pensiveness)
- Disgust (Loathing → Disgust → Boredom)
- Anger (Rage → Anger → Annoyance)
- Desire
- Other

Dataset	Features	Emotion model
ISEAR	Obtained from cross-cultural studies in 37 countries and contains 7665 sentences annotated for joy, sadness, fear, anger, guilt, disgust and shame emotions	Discrete
SemEval-2017, task 4	Data contains 1250 texts obtained from Tweets, News headlines, Google News and other major newspapers. Annotated for Ekman's 6 basic emotions	Discrete
Emobank	News headline, essays, blogs, newspapers, fiction, letters and travel guides	Dimensional
WASSA-2017 Emotion Intensities(EmoInt)	Constructed from tweets and annotated for joy, sadness, fear, and anger emotions	Discrete
Cecilia Ovesdotter Alm's Affect data	Constructed from Tales and classified into angry, fearful, happy, sad, disgusted and surprised emotions	Discrete
DailyDialog	Contains 13,118 Dialogues extracted from conversations and annotated for happiness, sadness, anger, disgust, fear, surprise, and others	Discrete
CrowdFlower	Constructed from 39,740 tweets and annotated for thirteen emotions	Discrete
Grounded emotions	Data constructed from 2,557 tweets and annotated for happy and sad	Discrete
Emotion Stimulus	Data developed from FrameNets' annotated data for emotion lexical unit.Contains 1594 emotion-labeled sentences	Discrete
The Valence and Arousal dataset	Built from 2,895 Facebook Posts	Dimensional
MELD data	Obtained from dialogues and utterances in a Television Show called Friends	Discrete
Emotion Lines	Obtained from dialogue in Friends TV Show and Facebook messenger chats.	Discrete
SMILE dataset	Gathered from tweets about British Museum	Discrete
Dens Dataset	Data contains 9,710 passages extracted from online narratives on wattpad and literature on project Gutenberg and classified into joy, sadness, anger, fear, anticipation, surprise, love, disgust, neutral.	Discrete
Aman Emotion Dataset	Constructed from blogposts	Discrete

Table 4: Datasets Overview for Emotion Recognition Tasks from [Acheampong et al., 2020]

According to domain experts, *Trust* and *Anticipation* are less relevant to the time period considered in Odeuropa. For the same reason, we add *Desire* and *Other*. The latter, in particular, is considered an open field for all feelings that are affect-loaded but cannot be tagged with any of the above emotions. This will be used to revise the basic emotion list, possibly also with affects, after the first annotation tests.

The process first foresees the identification of olfactory events and their annotation following the guidelines presented in Section 2.1. Then, for each olfactory event or situation, annotators are asked to add information on whether emotions are mentioned or involved in the event by adding one or more emotions from our list inspired by Plutchik. Emotions that are not related to olfactory events should not be annotated. No constraints on the extension of the text span to be annotated are set: emotions can refer to single words or expressions, or even entire text passages. Also, multiple emotions can be associated with the same text span.

We expect that emotions are mostly expressed as 1) the description of Quality (i.e. adjectives referred to Smell sources , e.g. "a fresh smell"), 2) associated with the Effect role ("that smell made me feel young again"), but also 3) as inherent part of the smell event / verb "this sausage smells", "het riekt hier".

Below we report some examples of olfactory events annotated with one of Plutchik's emotions. Smell-words (lexical units) are underlined, while emotions are put in italics between squared brackets:

> The earthy smell of the dried leaves was balm to my sense *[serenity]* after the hateful odour of sea-weeds *[disgust]*.
> As I stept within, a pernicious scent assailed my senses, producing sickening qualms, which made their way to my very heart, while I felt my leg clasped, and a groan repeated by the person that held me *[fear, disgust]*.
> It should have little or no odour, and the odour should not be disagreeable, for dis-

Figure 1: Plutchik's wheel of emotions

eased meat has a sickly cadaverous smell *[disgust]*, and sometimes a smell of physic. *[disgust]*

2.3 Annotation Tool

To carry out annotation of olfactory information in texts following the above guidelines, we need a tool that can be easy customised, supports multiple languages and multiple annotators and, given the complexity of the task, enables a continuous quality control. After analysing several tools widely used by the natural language processing community for annotation tasks, we have selected INCEpTION[3] [Klie et al., 2018], a web-based text-annotation environment, which fulfills all the task requirements.

First of all, it provides a customisable environment and it is distributed under the Apache License v2.0. Second, it can be configured to run on a private server, in our case managed by FBK which is the WP3 leader, to avoid unnecessary exchanges of files among the partner and ensuring a better task control. Third, linguistic annotation can refer to different textual spans (characters, words or sentences) so that language variability and differences among languages can be accounted for. The set of labels to be attached to texts is fully customisable and it is possible to define custom relations between the annotated spans, supporting one-to-one, one-to-many and many-to-one relations. This was necessary in our case because we not only identify smell words and semantic roles, each having a different semantic label, but we also connect them using three types of relations (i.e. standard, anaphoric and discontinuous). Annotation can

[3] https://INCEpTION-project.github.io/

also be multilayered, i.e. the same text span can be labeled with two different information, so that smell-related information and emotions can both be added to the same text. The tool also supports several formats to export the annotations including xml and json.

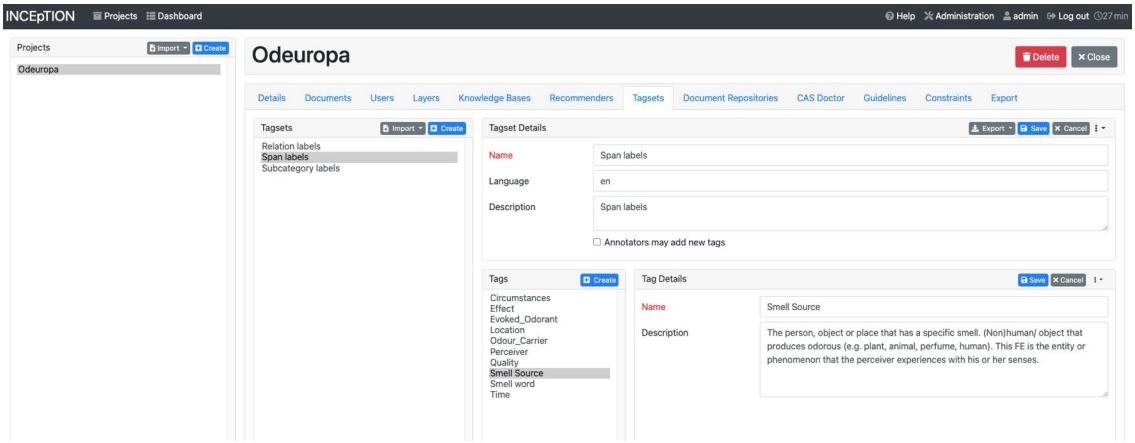

Figure 2: INCEpTION administrator panel for task definition

INCEpTION allows the creation of users with different levels of access: for each task we have an administrator, who is responsible for the design and management of the task, upload of the documents and export of the results. FBK will cover this role. Then, for each language a curator is defined, who is responsible for the quality control of texts annotated for a given language. Each partner involved in WP3 will be assigned a curator. Finally, annotators can be added, who do not have the possibility to modify the task but can only perform annotation. Curators have access to a panel where texts annotated by different users can be compared, correct them if needed or merge incomplete annotations.

Fig. 2 shows the panel managed by the administrator when the task is designed. In this view, the tagset has been defined and now the corresponding definitions for each tag are entered. Figure 3, in turn, shows the interface displayed to annotators. In this case, smell words have been marked (highlighted in light blue) and the related frame elements have been identified and connected to the smell word through a link (displayed as an arrow). INCEpTION includes also a suggestion function that, after a number of sentences have been manually annotated, starts to suggest annotations in unseen texts by replicating existing annotations. In preliminary tests, this function proved to speed up the process remarkably, in particular for the selection of smell words.

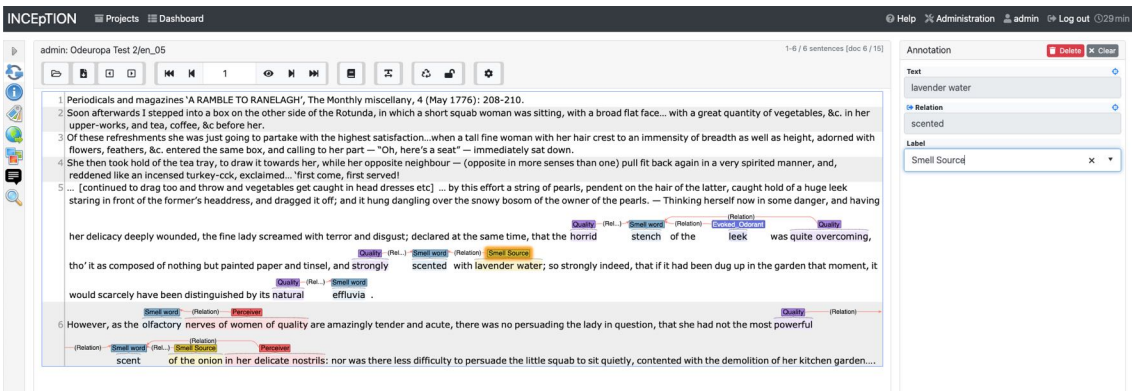

Figure 3: Annotation interface for olfactory information

3 Multilingual Taxonomy Creation

The second task to be carried out in WP3 within M6 is the creation of a multilingual taxonomy of olfactory information capturing domain-specific terms in the different project languages. This resource was created by taking into account knowledge from domain experts, by revising and merging existing olfactory lexicons and by taking advantage of statistical information related to word co-occurrences extracted from word n-grams and word embeddings. The taxonomy includes also temporal information related to the use of olfactory terms over time. The taxonomy creation process has been designed to be *i)* multilingual, making use of techniques and resources that are available for all project languages, and *ii)* modular, so that is can be incrementally improved throughout the whole project and single components can be easily updated.

An overview of the workflow for the taxonomy creation is shown in Figure 4. The development process starts from a set of so-called *seed terms*, i.e. words that are unambiguously related to the olfactory domain and that have been selected by domain experts for each project language. These are the same used as a reference for lexical units, reported in Table 1. Each term is looked up in the corresponding language-specific WordNet [Fellbaum, 1998, Miller, 1995], a cognitively-motivated database where terms (verbs, nouns, adjectives and adverbs) are organised into synsets, i.e. sets of synonyms. This first core set of synsets is then expanded using WordNet relations. The details of this step are described in Section 3.1.

Next, the core taxonomy is further expanded by using word sequences (called *n-grams*), extracted either from Google Books[4] or from other existing databases, to capture the terms that co-occur more frequently with the seed terms and that are likely to refer to the olfactory domain. Since n-grams are released together with information on their frequency and the year of publication of the book(s) where the n-gram was found, co-occurrence information can be analysed also over time. This step is detailed in Section 3.2. Since the number of terms co-occurring with seed terms can be very high, we introduce a last step, described in Section 3.3, where we cluster terms extracted from the n-grams trying to automatically assign them to smell categories (e.g. Emissions and traffic, Food and beverage, etc.) or scent families (e.g. Fragrant and fruity, Woody and earthy, etc.) via word embeddings [Grave et al., 2018].

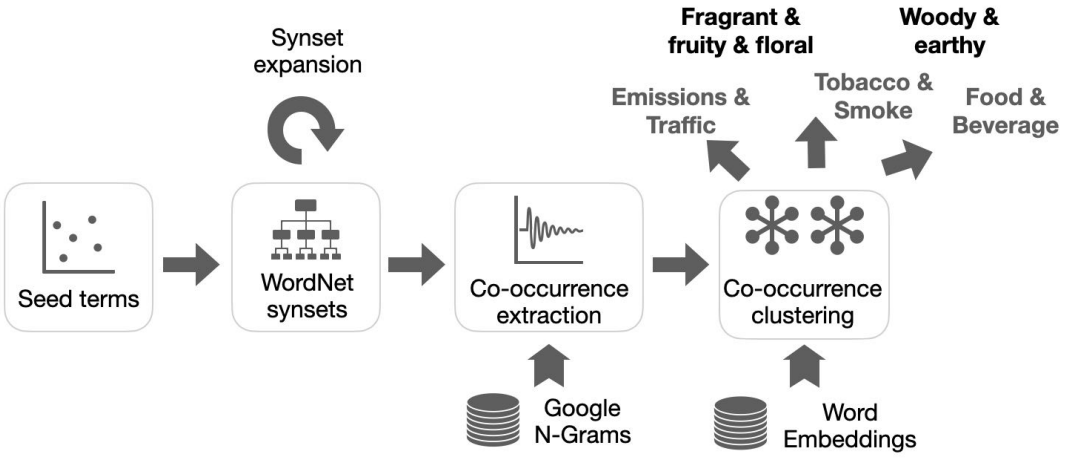

Figure 4: Workflow for multilingual taxonomy creation

[4] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

3.1 Core taxonomy creation from WordNet

To create a core taxonomy, we adopt a similar approach to [Tekiroğlu et al., 2014] that use a set of WordNet relations [Fellbaum, 1998, Miller, 1995] to expand a core set of seed words for five human senses. We also rely on the same intuition as [Kim and Hovy, 2004], that propose to use relations in WordNet to infer word polarity starting from a small set of synsets. WordNet contains nouns, verbs, adjectives and adverbs, which are grouped into sets of cognitive synonyms (synsets). Synsets are connected to each other through lexical semantic relations.

The core taxonomy is built starting from the seed words provided by the domain experts for each language included in Odeuropa and reported in Table 1. First, we have automatically mapped each word into a WordNet synset, after removing seed words from the list that may be ambiguous or apply also to other domains. The mapping was straightforward: each synset containing one of the seed terms was considered a candidate to be included in the taxonomy. Since the main objective while creating the core taxonomy for smell related words is precision, we have then conducted an annotation task on the obtained synsets using also their definitions, i.e., glosses, to remove the non-smell related ones. Also, not all seed words were found in WordNet, because its coverage for some languages is limited.

In the second step, we investigated all possible relations included in the WordNet of the given language to retrieve new smell related synsets. Also in this case, the outcome of the expansion step was manually revised to include only correct synsetsin the core taxonomy. For instance, for the noun seed *smell*, we expand the list with the hyponyms of its synset such as the nouns *bouquet, fragrance, fragrancy, redolence and sweetness*. The same process has been carried out in the seven project languages using the specific WordNets. Since their coverage and structure may vary, we adjusted the mapping and expansion steps as needed. The details related to the single WordNets are reported below:

- **English.** For English, we use Princeton WordNet [Fellbaum, 1998, Miller, 1995][5]. English WordNet is the most comprehensive among the other languages and includes the folllwing relations: synonyms, antonyms, derivationally_related_to lexical relations and hypernyms, hyponyms, instance_hypernyms, instance_hyponyms, also_sees, similar_tos, attributes, member_holonyms, substance_holonyms, part_holonyms, member_meronyms, substance_meronyms, part_meronyms, entailments, verb_groups, causes synset relations.

- **Italian.** We used MultiWordNet [6] [Pianta et al., 2002] for the Italian core taxonomy, which is strictly aligned with Princeton WordNet (PWN) and all PWN relations are represented in it. However, its coverage is much smaller compared to PWN.

- **French.** The WOLF [7](Wordnet Libre du Français) [Sagot and Fišer, 2008] is a semantic lexical resource for French. WOLF is also aligned with PWN, therefore we could utilize the same relations for the seed word expansion.

- **Slovenian.** sloWNet[8] [Fišer et al., 2012] is the lexical semantic database for Slovenian language. sloWNet is also based on PWN.

- **Dutch.** Open Source Dutch WordNet (odwn) [9] [Postma et al., 2016] is a Dutch lexical semantic database, derived from the Cornetto Database and PWN. We include all lexical semantic relations found both in OpenDutchWordNet and PWN.

- **German.** GermaNet[10] [Hamp and Feldweg, 1997] is a German lexical semantic database

[5] https://wordnet.princeton.edu/
[6] https://multiwordnet.fbk.eu/english/home.php
[7] http://pauillac.inria.fr/~sagot/index.html#wolf
[8] http://lojze.lugos.si/darja/research/slownet/
[9] http://www.cltl.nl/projects/current-projects/opensourcewordnet/
[10] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/

containing nouns, verbs, and adjectives. Relations that GermaNet contains can be found in https://uni-tuebingen.de/en/142846.

- **Latin.** For the Latin seed words, we also utilized MultiWordNet [Pianta et al., 2002]. Latin is the only language for which the expansion has been carried out without a manual revision of the outcome. This has been delayed due to problems in hiring annotators for this language.

Language	Mapped Synsets	Unique lemma_PoS	Expanded Synsets	Unique lemma_PoS
English	49	76	121	268
Italian	22	58	38	90
French	32	75	48	88
Slovenian	16	54	26	42
Dutch	41	67	56	106
German	18	35	86	123
Latin	61	152	-	-

Table 5: Core Taxonomy statistics: number of retrieved synsets by mapping to WordNet after manual correction (except for Latin), number of unique lemmas from mapped synsets, number of synsets after 1 step of expansion and manual correction, and total number of unique lemmas extracted from the expansion.

Table 5 shows the statistics about the core taxonomy. As shown in the table, we were able to create a core taxonomy for each of the project languages, and to manually revise its content. The final lists are different due to the fact that the initial lists of seed words have different sizes, and the language-specific WordNets have different coverage. English is the language with the most lemmas in the taxonomy, followed by German and Dutch. Slovenian, on the other hand, has the fewest lemmas, which event decreased after the first expansion. This is due to the fact that several terms were discarded because of vagueness or ambiguity, and also because different synsets may contain the same lemmas. Note that the multilingual taxonomy is a work in progress, and that the taxonomy creation process was designed so as to make updates easy and fast. Therefore, the content of these lists may change while manual annotation progresses.

3.2 N-Gram Based Expansion

Starting from the lemmas obtained from WordNet synsets, we further expand the taxonomy following a data driven approach. The idea is to enrich the resource by looking into actual texts (from books and newspapers) for terms that although not being smell words are strictly connected to smells, expressing possible sources (e.g. *smoke*, *bread*), qualities (e.g *fruity*) or information that can be use to reconstruct the presence of specific smells and the way in which they were perceived. For this purpose, we exploited existing textual resources released in the form of n-grams (contiguous sequences of a fixed number of tokens extracted from texts).The source of the n-grams varies according to what is available for the different languages included in the taxonomy.

Google Ngrams[11] [Michel et al., 2011] covers four of the languages in the taxonomy (English, Italian, German and French). These n-grams are extracted from the documents in the Google Books collection, covering the time period from circa 1525 up to 2009. Dutch is covered in the Newspaper n-gram collection[12] [De Goede et al., 2013] containing n-grams from Dutch newspapers from 1600 to 1995. Other Dutch n-grams have been created by the Digitale Bibliotheek voor de Nederlandse Letteren, but they are currently not available for download.[13]

[11] https://storage.googleapis.com/books/ngrams/books/datasetsv3.html
[12] https://lab.kb.nl/dataset/newspaper-ngram-collection
[13] https://www.dbnl.org/ngram-viewer/

For Slovenian we rely on the n-grams extracted from the IMP corpus of historical Slovenian [14][15] [Dobrovoljc, 2018] that contains books and newspapers from the end of the 16th century to 1918. For Latin there are no available repositories of n-grams, therefore we will carry out this step once we have collected a large number of documents by extracting our domain-specific n-grams. For all the other languages, the expansion was done using 5-grams, looking for words related to smells in spans of 5 tokens.

Before looking for smell-related terms in the N-Grams, we manually extend the list of lemmas by adding all their inflections. Indeed, N-Grams are not lemmatised, so that searching for all possible forms of a word increases the possibility to find some occurrences. We tried different alternatives to automatically generate word forms starting from a lemma, but the output was generally not accurate enough, so we decided to perform this task manually with the help of native speakers. We first compare two different expansions: one starting from the unique lemmas extracted from the mapped synsets (Column 3 in Table 5) and the other from the more extensive list obtained after the expansion (last Column in Table 5). This preliminary comparison was done on the English set. We observed that through the expansion, the second list would retrieve many more co-occurring lemmas, but less related to smells and the olfactory domain, while the first list would lead to more pertinent terms. Therefore, we carried out the n-gram based expansion starting from the first list also for the other languages.

The expansion process foresees the following steps:

1. For each lemma in the list, create all the inflected word forms
2. Look for each word form in the n-grams of the corresponding language with a date included in the time period between 1650 and 1925
3. Discard the n-grams that occur only once
4. For each retrieved n-gram, compute pointwise mutual information (PMI) between the smell-related word and each term in the surrounding context. PMI [Church and Hanks, 1989] is a measure of association between pairs of words indicating whether two terms to co-occur more frequently than usual and are therefore related
5. If N-Grams contain PoS information, keep only co-occurring terms that are a *noun*, *verb*, *adjective* or *adverb* (applies only to English, Italian, German and French)
6. Discard terms that co-occur only once with the smell-related term
7. Discard co-occurring terms with a PMI ≤ 0 (i.e. indicating that the two words are are independent).

For the languages were the n-grams are associated with a date (all but Slovenian) all smell-related terms and the extracted co-occurrences were grouped into 6 time spans, namely *1650–1700*, *1701–1750*, *1751–1800*, *1801–1850*, *1851–1900*, *1901–1925*. In this way, we are able also to compute statistics on the frequency of use of the terms over time.

3.3 Term Categorisation

Since the co-occurring terms extracted by computing PMI on n-grams can total up to several thousands for each language (see last column of Table 8), we decided to organise them into categories by applying an automatic clustering algorithm. While clustering can be fully unsupervised, the categories to be included in the taxonomy were defined starting from existing works on odour classification.

In the literature, several approaches to odour classification have been proposed (for a summary see [Kaeppler and Mueller, 2013]), adopting different perspectives, for example focusing on

[14] https://www.clarin.si/repository/xmlui/handle/11356/1194
[15] http://nl.ijs.si/imp/

the functions of odour receptors, or on the study of molecules. For Odeuropa, we adopt a categorisation that is more related to odour descriptions, since they can be more easily connected to texts. For this reason, in this first stage, we try to consider smell descriptors in their mere nature of words and not as related to precise chemical compounds or perceptual experiences.

The way odours are described depends on two main factors: the *source* of the odour, namely what emits the odour that is perceived, and the *evaluation* of odours, which admits different levels of interpretation, such as intensity, hedonic tone, affect, memory and quality. This dichotomous differentiation allows us to classify smell descriptors in terms of lexical entities, so that nouns represent smell-sources and adjectives describe the odour evaluation. We aim at merging different existing resources along these two dimensions. The lexicons we consider are [Lynott and Connell, 2013], in which nouns and adjectives are rated in terms of their association with the five perceptual modalities, a selection from Sensicon [Tekiroğlu et al., 2014], an automatically generated sensorial lexicon that associates words with senses; the olfactory lexicon by [Lievers, 2015]; the smell vocabularies available at https://sensorymaps.com/?projects=comparative-smell-vocabularies, the urban smell dictionary by [Quercia et al., 2016] and that of [Henshaw, 2013].

We also look for existing taxonomies to cover the different domains relevant to Odeuropa, namely travel literature, scientific texts, and medical records. We therefore choose the taxonomy of Linnaeus[16], belonging to the field of botany; the perfume wheel of [Edwards, 2018], first released in 1992, which classifies perfumes and fragrances; the odor wheel of historic books by [Bembibre and Strlič, 2017], which classifies smell descriptors for books' odours; and the classification of [Castro et al., 2013], which presents the attempt to identify the so-called primary odours.

In the lexicons we work with, we first select nouns and adjectives, and subsequently remove from the list those that are strictly smell words, i.e. synonyms or near-synonyms of perfume, smell, odour, because they are neither smell sources nor evaluations. We then perform a third selection by manually eliminating human referents and people in general (policeman, janitor etc.) and some specific terms which are not useful for our purposes (e.g., scientific names of rare animals).

The final harmonisation of these classifications has led to eight categories for qualities and nine for smell-sources, in which we distributed the lists of adjectives and nouns previously collected, starting from the smell descriptors already classified in the original taxonomies. With respect to categorisation of smell-sources, we focus on the studies carried out on urban smells by [Henshaw, 2013, Quercia et al., 2016], since people participating in these studies tried to identify the origin of the smell they perceive referring very frequently to objects. For what concerns qualities instead, we mainly refer to the other taxonomies described above, since in these cases researchers are also interested in a description of the effects produced in the perceiver and then of the qualities of the perceived smell. The final list of categories identified for smell sources and for qualities is reported in Table 6.

Smell Sources	**Qualities**
emissions,traffic, fuel, dust	fragrant, fruity, floral
industry	woody, earthy, mouldy
food, beverage	chemical, hydro-carbons, synthetic
tobacco, smoke	fresh, cool
cleaning, medicinal	sweet, spicy
synthetic	smoky, toasted, burnt, fatty
waste, garbage, pee, vomit, excrement, rotten	decayed
animal, people	pungent
nature, flowers, plant, tree, soil	

Table 6: Categories identified to classify smell sources and qualities

A list of words extracted from existing taxonomies was manually assigned to each of the above

[16] https://psychology.wikia.org/wiki/Linnaeus%27s_classification_of_smell

categories. Overall, we collect 347 English words (nouns) as smell sources and 94 adjectives as qualities, for a total of 441 words. We assume that the categories are language-independent, and we use them for all project languages, and consequently the terms were manually translated into Italian, French, German, Slovenian and Dutch.

Since our goal is to assign to a category the co-occurrence of terms extracted from the n-grams as described in Section 3.2, we proceed as follows:

1. We represent each of the 441 categorised terms as a word embedding using fastText [17] [Grave et al., 2018] vector space. fastText embeddings cover 157 languages, including the ones used in Odeuropa, trained on Common Crawl and Wikipedia.
2. Each category reported in Table 6 is represented as a cluster of embeddings
3. Each co-occuring term t extracted from the n-grams is represented as a word embedding in the same multidimensional space, to be assigned to one category
4. If t is a noun, then we try to assign it to one of the categories for smell sources; if t is an adjective, we assign it to qualities
5. The assignment is performed by estimating the probability of belonging to one of the categories of smell sources/qualities by mean of proximity with each cluster, with the distance represented as the cosine distance between the term embedding and the centroid of the cluster.
6. The term is assigned to the category of the cluster with the highest cosine similarity, with a minimum threshold of 0.4 for smell sources and 0.5 for qualities.
7. If no category reaches this minimum similarity threshold, the term is included in the taxonomy but is not assigned any category label. The same for co-occurring terms that are neither nouns nor adjectives.

For example, the terms 'trash' and 'toilet' were assigned to the *waste* & *garbage* category. 'Incense', 'opium' and 'cigar' fell in the *tobacco* & *smoke* category, 'humid' was assigned to the *woody, earthy and mouldy* category and 'disgusting' to the *decayed* one.

In Table 7, we report the number of terms assigned to different categories of smell sources and of qualities through clustering. This task involved only the four languages whose n-grams contain PoS-tag information (EN, IT, DE, FR), since it is necessary to distinguish between nouns and adjectives to assign them to the correct categories.

3.4 Multilingual Taxonomy v.1

The first version of the taxonomy and the scripts to perform the different mapping and extraction steps are available at https://github.com/Odeuropa/multilingualTaxonomies. For each language, we release two files in tab-separated format. The first file contains the following columns:

1. entry: term listed in the taxonomy
2. source: whether the term comes from the WordNet-based core taxonomy or has been obtained through n-gram co-occurrences.
3. synset: if it comes from WordNet, which is the synset unique identifier
4. first appearance: if it comes from co-occurrences, in which year it appeared first (if n-grams contain temporal information)
5. time periods: for each time period between 1650 and 1925 (spans of 50 years), whether the term is mentioned or not

[17] https://fasttext.cc/docs/en/crawl-vectors.html

Smell-Source	**EN**	**IT**	**DE**	**FR**
animal, people	147	42	41	241
cleaning, medicinal	87	32	43	195
emissions, traffic, fuel, dust	35	21	20	114
food, beverage	200	42	20	187
industry	67	24	14	194
nature, flowers, plant, tree, soil	248	71	55	274
synthetic	17	4	0	2
tobacco, smoke	27	5	6	26
waste, garbage, excrement, rotten	83	23	13	63
Quality	**EN**	**IT**	**DE**	**FR**
chemical, hydro-carbons, synthetic	16	10	3	38
decayed	34	15	1	28
fragrant, fruity, floral	42	26	13	71
fresh	8	5	2	1
pungent	42	10	13	33
smoky, toasted, burnt, fatty	16	4	2	8
sweet, spicy	19	5	0	20
woody, earthy, mouldy	42	4	9	22

Table 7: Number of terms for each language automatically assigned to the categories included in 'Smell source' and 'Quality'. The former are all nouns, the latter all adjectives.

6. smell-source: for nouns in English, Italian, German and French, to which category of smell sources it was assigned (see Section 3.3)

7. quality: for adjectives in English, Italian, German and French, to which category of qualities it was assigned (see Section 3.3)

The second file contains information about pairs of co-occurring terms (i.e. seed word + co-occurring term with high PMI) extracted from n-grams. Each file in tsv format contains the following columns:

1. seed word + part of speech

2. co-occurring term, extracted following the process described in Section 3.2

3. time span when the two terms were found (if found in multiple time spans, the row is repeated with different values)

4. frequency: number of times the two terms were co-occurring in the given time span

5. total tokens: number of overall tokens present in the n-grams for the given time period, to be used as a reference or to normalise the frequency

Table 8 displays for every language the number of unique seeds that have been found in every time span and their total, as well as the number of of co-occurrences related to these seeds.

4 Conclusions

In this document, we describe the Odeuropa annotation scheme for capturing information on olfactory events and related emotions, as well as the tool that will be used for this task. We give a detailed description of the annotation scheme and we provide a motivation of the choices made,

	1650-1700		1701-1750		1751-1800		1801-1850		1851-1900		1901-1925		Total	
Lang	Seeds	Coocs	Seeds	Coocs	Seeds	Coocs	Seeds	Coocs	Seeds	Coocs	Seeds	Coocs	Seeds	Coocs
EN	29	316	51	1,989	64	4,805	68	8,519	73	9,963	74	9,581	76	10,365
IT	3	9	5	17	19	165	35	636	46	1,367	46	1,603	48	1,783
DE	2	3	4	7	12	321	22	1,111	30	1,451	34	1,568	35	1,692
FR	26	420	32	941	48	3,335	54	6,029	56	6,878	55	6,270	57	7,776
NL	-	-	-	-	-	-	26	905	49	3,904	47	3,878	57	6,579
SL	-	-	-	-	-	-	-	-	-	-	-	-	48	4,897

Table 8: Number of seeds and co-occurrences found in every time span for each language. For Slovenian n-grams, temporal information was not available.

adapting the existing FrameNet standard with project-specific extensions. In additon, for emotions, we present several past approaches to emotion classification, and for now select Plutchik's emotion wheel as an initial reference.

Some minor adjustments of the annotation schemes are possible, and will be based on the outcome of upcoming pilot annotations, analysing in particular whether inter-annotator agreement reaches sufficient levels. Language-specific adaptations will also be implemented if needed, even if the guidelines definition has been as much language-independent as possible. By month 9, the annotated benchmark will be presented (D3.2), for which a set of documents for each of the seven project languages is manually annotated following these guidelines.

Concerning the multilingual taxonomy, we presented the process implemented to obtain a first version of the resource for each project language, taking advantage of the fact that for each of them a WordNet version is available. This first version of the multilingual taxonomy has variable content for each language: whereas the Latin one is composed only of the core set of synsets related to smells, for the other languages more terms have been extracted by applying PMI to n-grams. Also in this case, we observe differences in terms of coverage and content: for English, Italian, German and French, co-occurrences have also been classified into categories of Smell sources and Qualities, All languages but Slovenian present also historical information (i.e. frequency per time period). Furthermore, some taxonomies include more terms than others, in particular the English one is larger. This depends on the variable number of n-grams available for each language, with the English repository containing more than eight billion unique n-grams (68 billion occurrences). For comparison, the Dutch repository contains only 29 million unique n-grams and the Slovenian one 10 million.

As a next step, we will work towards improving consistency across different languages and try to extend those with limited coverage. We will also integrate WP3 taxonomy with WP2 image-based one, so that both will be easily included in Odeuropa knowledge graph (WP4).

References

[Acheampong et al., 2020] Acheampong, F. A., Wenyu, C., and Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189.

[Bembibre and Strlič, 2017] Bembibre, C. and Strlič, M. (2017). Smell of heritage: a framework for the identification, analysis and archival of historic odours. *Heritage Science*, 5:1–11.

[Castro et al., 2013] Castro, J. B., Ramanathan, A., and Chennubhotla, C. (2013). Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization. *PLoS ONE*, 8.

[Church and Hanks, 1989] Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

[De Goede et al., 2013] De Goede, B., Van Wees, J., Marx, M., and Reinanda, R. (2013). PoliticalMashup Ngramviewer. In *International Conference on Theory and Practice of Digital Libraries*, pages 446–449. Springer.

[Delplanque et al., 2017] Delplanque, S., Coppin, G., and Sander, D. (2017). *Odor and Emotion*. Springer Handbook of Odor. Springer International Publishing. ID: unige:92488.

[Dobrovoljc, 2018] Dobrovoljc, K. (2018). IMP corpus n-grams 2.0. Slovenian language resource repository CLARIN.SI.

[Edwards, 2018] Edwards, M. (2018). *Fragrances of the World*.

[Fellbaum, 1998] Fellbaum, C. (1998). Wordnet. *An Electronic Lexical Database (Language, Speech and Communication)*.

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

[Fillmore and Baker, 2001] Fillmore, C. J. and Baker, C. F. (2001). Frame semantics for text understanding. In *In Proceedings of WordNet and Other Lexical Resources Workshop*.

[Fišer et al., 2012] Fišer, D., Novak, J., and Erjavec, T. (2012). slownet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117.

[Grave et al., 2018] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[Hamp and Feldweg, 1997] Hamp, B. and Feldweg, H. (1997). Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.

[Henshaw, 2013] Henshaw, V. (2013). *Urban Smellscapes: Understanding and Designing City Smell Environments*. Routledge.

[Kaeppler and Mueller, 2013] Kaeppler, K. and Mueller, F. (2013). Odor classification: a review of factors influencing perception-based odor arrangements. *Chemical senses*, 38 3:189–209.

[Kim and Hovy, 2004] Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373.

[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.

[Leemans et al., 2017] Leemans, I., Zwaan, J. V. D., Maks, I., Kuijpers, E., and Steenbergh, K. (2017). Mining embodied emotions: A comparative analysis of sentiment and emotion in dutch texts, 1600-1800. *Digital Humanities Quarterly*, 11.

[Lievers, 2015] Lievers, F. S. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of Language*, 22:69–95.

[Lynott and Connell, 2013] Lynott, D. and Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, 45:516–526.

[Michel et al., 2011] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

[Miller, 1995] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

[Ortony and Turner, 1990] Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97 3:315–331.

[Pianta et al., 2002] Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

[Plamper, 2015] Plamper, J. (2015). *The History of Emotions: An Introduction*. The University of Chicago Press.

[Plutchik, 1980] Plutchik, R. (1980). *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, 1*. New York: Academic.

[Postma et al., 2016] Postma, M., van Miltenburg, E., Segers, R., Schoen, A., and Vossen, P. (2016). Open dutch wordnet. In *Proceedings of the Eight Global Wordnet Conference, Bucharest, Romania*.

[Quercia et al., 2016] Quercia, D., Aiello, L. M., and Schifanella, R. (2016). The emotional and chromatic layers of urban smells. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1).

[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.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

[Sagot and Fišer, 2008] Sagot, B. and Fišer, D. (2008). Building a free french wordnet from multilingual resources. In *OntoLex*.

[Tekiroğlu et al., 2014] Tekiroğlu, S. S., Özbal, G., and Strapparava, C. (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.