Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4030–4039
Marseille, 20-25 June 2022

Building a Multilingual Taxonomy of Olfactory Terms with Timestamps

Stefano Menini[1], Teresa Paccosi[1,2], Serra Sinem Tekiroğlu[1], Sara Tonelli[1]
[1] Fondazione Bruno Kessler – Via Sommarive 18, Trento, Italy
[2] Università di Trento – Corso Bettini 84, Rovereto, Italy
{menini, tpaccosi, tekiroglu, satonelli}@fbk.eu

Abstract

Olfactory references play a crucial role in our memory and, more generally, in our experiences, since researchers have shown that smell is the sense that is most directly connected with emotions. Nevertheless, only few works in NLP have tried to capture this sensory dimension from a computational perspective. One of the main challenges is the lack of a systematic and consistent taxonomy of olfactory information, where concepts are organised also in a multi-lingual perspective. WordNet represents a valuable starting point in this direction, which can be semi-automatically extended taking advantage of Google n-grams and of existing language models. In this work we describe the process that has led to the semi-automatic development of a taxonomy for olfactory information in four languages (English, French, German and Italian), detailing the different steps and the intermediate evaluations. Along with being multi-lingual, the taxonomy also encloses temporal marks for olfactory terms thus making it a valuable resource for historical content analysis. The resource has been released and is freely available.

Keywords: olfactory information, taxonomies, multilingual resources

1. Introduction

In the last two decades, the attention of scholars in the humanities and social sciences has shifted away from the visual and textual dimensions to a multi-sensory perspective, following a so-called ‘sensorial revolution’ (Howes, 2006; Classen, 1999). For example, recent works by history scholars have dealt with the meaning of odours in particular places and times (Dugan, 2011), or with the role of smells in shaping identity and otherness in the past (Smith, 2006; Tullett, 2016). This turn, however, has had a limited impact on NLP research. Indeed, while the visual dimension is prevalent in texts, linguists have showed that in Western languages some senses such as taste and smell are less represented and are expressed with a more limited and ambiguous vocabulary than the visual one (Winter, 2019; Majid and Burenhult, 2014). This may explain why multi-sensory studies are still a niche area in the computational linguistics community. However, being able to automatically identify how the different sensory experiences are described would be very relevant to accurately carry out different tasks, from emotion detection to metaphor identification. Furthermore, accounting for differences in sensory vocabulary across languages and over time would be of great interest for cultural studies, digital humanities and historical content analysis.

In this work, we present a taxonomy of olfactory-related terms that we have semi-automatically created. Our goal was to make the process rather fast, taking advantage of existing resources such as WordNet (Miller, 1995) and Google n-grams, while ensuring a high-quality outcome. The resource covers olfactory terms in English, French, German and Italian, and, while its core structure is based on synsets, it has been extended using co-occurrence information from n-grams, word embeddings and a final classification step for qualities and smell sources. Our approach aims at integrating domain information and existing resources top-down with distributional and similarity information acquired bottom-up. Furthermore, the terms in the taxonomy are enriched with temporal information about their appearance over time, so that it is possible to track their usage in relation to smell situations in the past centuries. The taxonomy has been released and is freely available at this link: `https://github.com/Odeuropa/multilingualTaxonomies`.

2. Related Work

Various studies report findings supporting that sensory language triggers cognition in specific ways. Stevenson and Case (2005) state that humans respond in a similar manner when they imagine a smell and when they actually perceive a smell. Rodriguez-Esteban and Rzhetsky (2008) report that using words related to senses in a text could clarify the meaning of an abstract concept by facilitating a more concrete imagination. The readability and understandability of text could also be affected by the use of sensory words (Rodriguez-Esteban and Rzhetsky, 2008). Additionally, sensory words affect private psychology by inducing a positive or negative sentiment (Majid and Levinson, 2011). For instance, de Araujo et al. (2005) show that the pleasantness level of the same odour can be altered by labeling it as *body odour* or *cheddar cheese*. Sensorial information in the lexical form draws attention from various disciplines. (Lynott and Connell, 2009; Lynott and Connell, 2013) collect the modality norms for 423 prenominal adjectives which are considered as concept properties and for 400 nouns. Following their work, Winter (2016) generate a verb dataset with modality norms. In order to detect synaesthetic metaphors in the language, Lievers (2015) assemble a list of directly sensorial words both for English and for Italian. We rely on some of these resources to create a list of categories for smell sources and qualities (see Section 6.1).

Concerning the computational analysis of olfactory terms and the extraction of related information, only few works have addressed this topic within the NLP community. Most works have focused on the creation of structured resources to capture the sensory domain, automatically deriving them from WordNet (Tekiroğlu et al., 2014). In particular, Tekiroğlu et al. (2015) propose a novel technique to automatically discover human sense-word associations from a dependency-parsed corpus. Other works have dealt with the automated analysis of texts related to specific domains like wine reviews, where olfaction plays a central role (Lefever et al., 2018). Other studies have focused on synaesthetic aspects of language, starting from a controlled lexicon of perception (Lievers and Huang, 2016). Tonelli and Menini (2021) introduce an annotation scheme inspired by FrameNet to capture smell events in texts, while Brate et al. (2020) propose both a simple annotation framework to capture smelly experiences and two semi-supervised approaches to automatically replicate this annotation. Another line of research has addressed so-called urban smellscapes, i.e. how modern cities can be described from an olfactory point of view. More specifically, Quercia et al. (2015) and Quercia et al. (2016) obtain descriptions of different urban areas by asking annotators to walk around cities and take note of the smell characterising different places. Such descriptions are then combined with social media posts about the same places, allowing the authors to build an olfactory representation of different cities and categorise urban smells into odour wheels.

Our contribution is novel in that *i)* it presents a pipeline for olfactory taxonomy creation integrating manually curated resources and distributional information, *ii)* evaluates this approach on four languages, with the goal to create a multilingual taxonomy, and *iii)* takes temporal information into account, integrating timestamps in the final resource.

3. Overview of Multilingual Taxonomy Creation

In this paper we detail the creation of a multilingual taxonomy of olfactory information, capturing domain-specific terms in four different languages and enriching them with timestamps. This resource has been 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 n-grams and word embeddings. The taxonomy creation process has been designed to be *i)* multilingual, making use of techniques and resources that are available for different languages, and *ii)* modular, so that it can be incrementally improved and single components can be easily updated.

In order to address language change over time, not only in terms of changing spelling and grammar conventions but also with respect to changes in meaning, we also aim at including the diachronic dimension of this resource by associating each term with information on the time period where it occurred. We focus on a time span between the 17th to the 20th Century for three main reasons:

- **Availability** of open access data: Since our goal is also to use the taxonomy as a starting point for information extraction, connecting the terms to existing corpora, we focus on the years before 1920 because there are more datasets in the public domain published before that period;
- **Feasibility**: Going back in time before 17th Century would mean have a limited amount of data, often affected by OCR problems;
- **Cultural historical context**: In sensory history, the 19th Century was a watershed moment in European history, characterized by rapid industrialisation and a significant shift in the status of olfaction (Tullett, 2019). Considering a time period between the 17th to the 20th Century enables the observation of changes before and after this important turn.

An overview of the workflow for taxonomy creation is reported in Figure 1. 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 language of interest. 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 4. Next, the core taxonomy is further expanded by using word sequences (*n-grams*), extracted from Google Books[1], in order 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 and enriched with the corresponding timestamps. This step is detailed in Section 5. Since the number of terms co-occurring with seed terms can be very high, we introduce a clustering

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

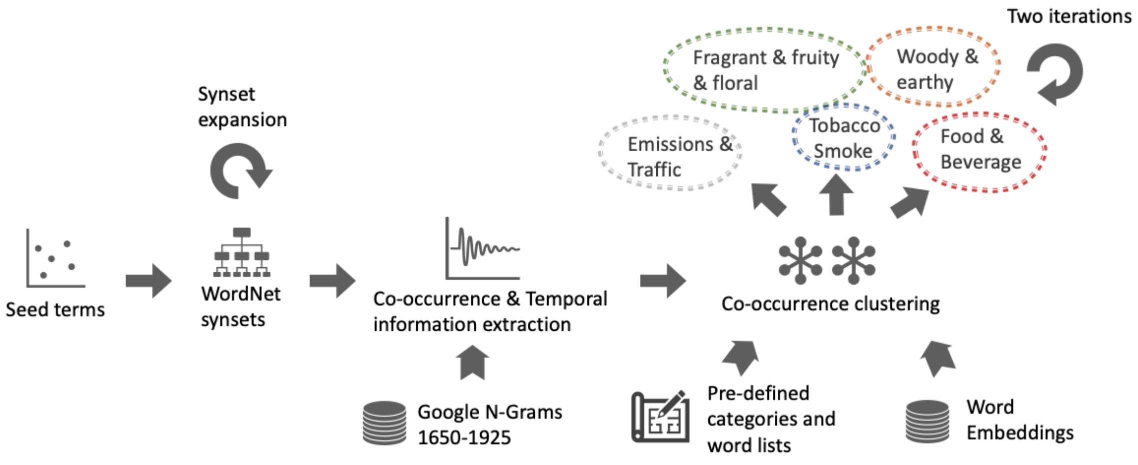


Figure 1: Workflow for multilingual taxonomy creation

step, described in Section 6, where we group nominal and adjectival terms extracted from the n-grams trying to automatically assign them to smell categories (e.g. Emissions and traffic, Food and beverage, etc.) or types (e.g. Fragrant and Fruity, Woody and Earthy, etc.) via word embeddings (Grave et al., 2018). The clustering is performed in two steps. In the second step, we use the output of the first one to increase the quality of the clusters and the number of terms assigned to a category. Each step is followed by a manual evaluation and correction. Finally the terms from the n-grams not associated with a category are discarded from the taxonomy to retain only those that are strictly related to the smell domain.

English: aroma.n, bouquet.n, essence.n, aromatic.a, fetid.a, inodorous.a, reek.v, smell.v, sniff.v, stink.v, whiff.v [...]
Italian: aromatico.a, fetido.a, fragrante.a, olfattivo.a, aroma.n, esalazione.n, annusare.v, emanare.v, puzzare.v [...]
French: aromatique.a, fétide.a, méphitique.a, arôme.n, effluve.n, flatulence.n, fumet, empester.v, parfumer.v [...]
German: duftend.a, stinkend.a, Bluetenduft.n, Knoblauchgeruch.n, Riechstoff.n, duften.v, riechen.v [...]

Table 1: Sample of seed words for each language

4. Core Taxonomy Creation

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 a list of terms which unambiguously refer to the olfactory domain (called *seed words*) which have been defined by three scholars experts in olfactory studies related to history and cultural heritage. A sample of seed words for each language is reported in Table 1. First, we have automatically mapped each word into a WordNet synset. 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 have 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 has been manually revised, to include only correct synsets in 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 English, French, German and Italian 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)[2]. English WordNet is the most comprehensive among the other languages and includes various lexical and semantic relations. In this study we have used all possible synset level relations that

[2] https://wordnet.princeton.edu/

can be retrieved through the Natural Language Toolkit:WordNet package (Bird et al., 2009)[3], i.e. 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, and the following lexical relations: synonyms, antonyms, derivationally_related_to.

- **Italian.** We have used MultiWordNet [4] (Pianta et al., 2002) for the Italian core taxonomy, which is strictly aligned with Princeton WordNet (PWN) and includes the same relations as PWN. However, its coverage is much smaller compared to PWN.
- **French.** The WOLF [5](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.
- **German.** GermaNet[6] (Hamp and Feldweg, 1997) is a German lexical semantic database containing nouns, verbs, and adjectives. Relations that GermaNet contains can be found in `https://uni-tuebingen.de/en/142846`. Since this resource is not aligned with PWN, we considered all possible relations.

Lang.	Mapped Synsets	Unique lemma	Expanded Synsets	Expanded Unique
EN	49	76	121	268
IT	22	58	38	90
FR	32	75	48	88
DE	18	35	86	123

Table 2: Core Taxonomy statistics: number of retrieved synsets by mapping to WordNet after manual correction, number of unique lemmas from mapped synsets, number of synsets after 1 step of expansion through relations and manual correction, and total number of unique lemmas extracted from the expansion.

Table 2 shows the statistics about the core taxonomy after the manual revision. The final lists are different in size across the languages due to the fact that the initial lists of seed words are different. In fact, seed words should include words that are unambiguously smell-related, which change a lot across languages and make translations sometimes not possible. Furthermore, language-specific WordNets have different coverage. English is the language with most lemmas in the taxonomy, because it was the most represented in the initial seed list. The other three lists have a comparable size, with the most lemmas in German probably due to the presence of compound words which correspond to single entries in the list (e.g. *Tabakgeruch*, *Alkoholgeruch* vs. *smell of tobacco* and *smell of alcohol*).

5. N-Gram Based Expansion

Starting from the list of lemmas reported in Table 2, 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 used to reconstruct the presence of specific smells and the way in which they were perceived. To this purpose we exploit existing textual resources released in the form of n-grams (contiguous sequences of a fixed number of tokens extracted from texts). We rely on Google Ngrams[7] (Michel et al., 2011) since it has an extensive coverage for all the four languages of interest. These N-Grams are extracted from the documents in the Google Books collection, covering the time period from the 16^{th} to the 21^{st} Century, which enables also the extraction of time anchors for the different terms. The expansion was done using 5-grams, the maximum size available, looking for words related to smells in spans of 5 tokens.

Before searching 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 compare two lists of lemmas by adding to the seed terms the ones from the two different WordNet expansion steps: one starting from the unique lemmas extracted from the mapped synsets (Column 3 in Table 2) and the other from the more extensive list obtained after the expansion (last Column in Table 2). 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 smell and the olfactory domain, while the

[3] `https://www.nltk.org/_modules/nltk/corpus/reader/wordnet.html`

[4] `https://multiwordnet.fbk.eu/english/home.php`

[5] `http://pauillac.inria.fr/~sagot/index.html#wolf`

[6] `https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/lexica/germanet-1/`

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

first list would lead to more pertinent terms. Therefore, we decided to carry 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 related to smell, create a list of all its 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 co-occur more frequently than usual and are therefore related.
5. Since Google N-Grams contain PoS information, keep only co-occurring terms that are a *noun* or *adjective*
6. Discard terms that co-occur only once with the smell-related word and with a PMI ≤ 0 (i.e. indicating that the two words are independent or they co-occur less frequently than expected).

From the extraction we obtain 15,856 unique co-occurrences for English, 1,992 for Italian, 7,099 for French and 1,284 for German. This value is influenced both by the number of unique lemmas used for the extraction and by the dimension of the N-Grams for each language.

Since the n-grams are associated with a date, 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 also able to compute statistics on the frequency of the usage of the terms over time.

6. Term Categorisation

Since the co-occurring terms extracted by computing PMI on n-grams can total up to several thousands for each language, 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.

6.1. Category Definition

In the literature there have been several proposals related to odour classification (for a summary see (Kaeppler and Mueller, 2013)), adopting different perspectives, such as focusing on the functions of odour receptors, or on the study of molecules. For this work we adopt a categorisation that is more related to odour descriptions, since they can be more easily connected to texts.

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 are generally used to 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)'s, 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` and the urban smell dictionary by Quercia et al. (2016) and that of Henshaw (2013).

We also look for existing taxonomies to cover different domains relevant to olfactory experiences, namely travel literature, scientific texts, and medical records. We therefore choose the taxonomy of Linnaeus,[8] belonging to the field of botany; the perfume wheel of Edwards (2018), first released in 1992, which classifies perfumes and fragrances; the odour 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, and because they would be present in the initial seed word list anyway. 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 or plants).

The final harmonization 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. The list of categories is reported in Table 3

With respect to categorisation of smell-sources, we focus on the studies carried out on urban smells by Henshaw (2013) and 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 descrip-

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

tion of the effects produced in the perceiver and then of the qualities of the perceived smell.

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

Table 3: Categories identified to classify smell sources and qualities

6.2. Term Clustering

The goal of this step is to assign to a category the co-occurrence of terms extracted from the n-grams as described in Section 5. Overall, in the previous step detailed in Section 6.1 we manually collect from past literature 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 4 languages, and consequently the terms were manually translated into Italian, French and German. We then proceed as follows:

1. We represent each of the 441 categorised terms as a word embedding using fastText[9] (Grave et al., 2018) vectorial space. fastText embeddings cover 157 languages, trained on Common Crawl and Wikipedia.

2. Each category reported in Table 3 is represented as a cluster of embeddings.

3. Each co-occurring 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.

Lang	Iter.	Threshold		Words	Accuracy		
		Noun	Adj		Noun	Adj	Total
en	1	0.65	0.75	46	94.44	100.00	95.65
	2	0.55	0.65	305	88.84	79.37	86.89
	Total	-	-	**351**	89.57	82.19	**88.03**
it	1	0.55	0.65	66	94.00	93.75	93.94
	2	0.5	0.6	76	81.03	66.67	77.63
	Total	-	-	**142**	87.04	79.41	**85.21**
fr	1	0.65	0.75	36	100.00	100.00	100.00
	2	0.55	0.65	246	78.67	60.00	76.02
	Total	-	-	**282**	81.48	64.10	**79.08**
de	1	0.6	0.7	21	94.12	100.00	95.24
	2	0.5	0.6	76	93.44	86.67	92.11
	Total	-	-	**97**	93.59	89.47	**92.78**

Table 4: This table reports, for every language, the threshold selected for each of the two iterations, the number of words classified in each step and their accuracy. We also report the accuracy of the output after the two iterations.

6. The term is assigned to the category of the cluster with the highest cosine similarity, by setting a minimum threshold to improve the accuracy of the labels.

7. If no category reaches this minimum similarity threshold, the term is not included in the taxonomy. The same for co-occurring terms that are neither nouns nor adjectives.

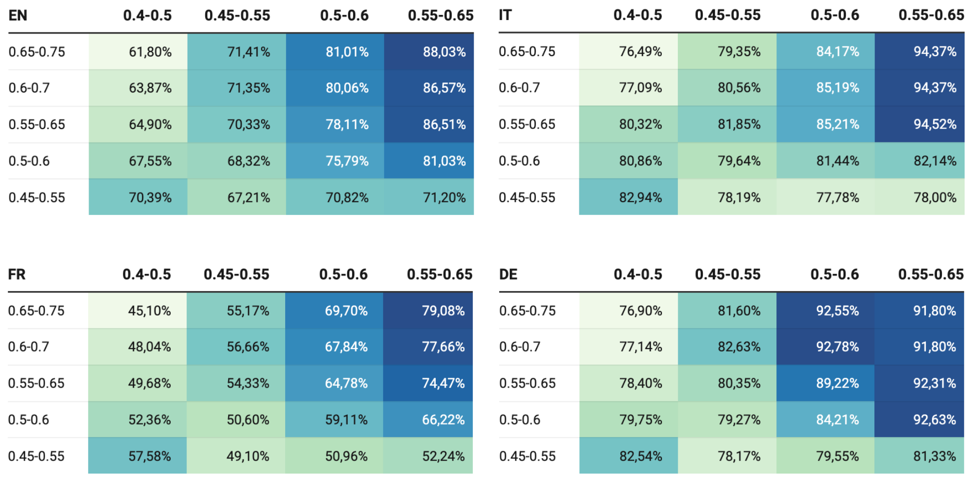
The process is repeated twice, so to extend the clusters used for the second iteration with the terms acquired during the first one. Each iteration is manually evaluated post-hoc by two domain experts in order to find the best trade off between the number of words assigned to a category and the accuracy in doing that. This manual check allows the identification of the best threshold value (Step 6).

We found this approach performing best by using different thresholds for smell sources and qualities with the cosine similarity threshold for qualities set 0.1 higher than the one for smell sources.

In Figure 2 we report the accuracy of the possible combinations of thresholds for the first and second iterations, testing thresholds from 0.4 to 0.75 with steps of 0.05. The best configuration is selected by taking into consideration both the accuracy and the number of words classified. Higher thresholds and accuracy result in fewer words classified. Table 4 reports, for every language, the threshold selected for the two iterations with the number of words classified and their accuracy.

As an example of the output, the terms 'trash', 'urine' 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,*

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



EN	0.4-0.5	0.45-0.55	0.5-0.6	0.55-0.65
0.65-0.75	61,80%	71,41%	81,01%	88,03%
0.6-0.7	63,87%	71,35%	80,06%	86,57%
0.55-0.65	64,90%	70,33%	78,11%	86,51%
0.5-0.6	67,55%	68,32%	75,79%	81,03%
0.45-0.55	70,39%	67,21%	70,82%	71,20%

IT	0.4-0.5	0.45-0.55	0.5-0.6	0.55-0.65
0.65-0.75	76,49%	79,35%	84,17%	94,37%
0.6-0.7	77,09%	80,56%	85,19%	94,37%
0.55-0.65	80,32%	81,85%	85,21%	94,52%
0.5-0.6	80,86%	79,64%	81,44%	82,14%
0.45-0.55	82,94%	78,19%	77,78%	78,00%

FR	0.4-0.5	0.45-0.55	0.5-0.6	0.55-0.65
0.65-0.75	45,10%	55,17%	69,70%	79,08%
0.6-0.7	48,04%	56,66%	67,84%	77,66%
0.55-0.65	49,68%	54,33%	64,78%	74,47%
0.5-0.6	52,36%	50,60%	59,11%	66,22%
0.45-0.55	57,58%	49,10%	50,96%	52,24%

DE	0.4-0.5	0.45-0.55	0.5-0.6	0.55-0.65
0.65-0.75	76,90%	81,60%	92,55%	91,80%
0.6-0.7	77,14%	82,63%	92,78%	91,80%
0.55-0.65	78,40%	80,35%	89,22%	92,31%
0.5-0.6	79,75%	79,27%	84,21%	92,63%
0.45-0.55	82,54%	78,17%	79,55%	81,33%

Figure 2: Accuracy for each combination of thresholds over 4 languages. The Y axis contains the thresholds tested in the first iteration, while the X axis represents the thresholds used in the second one.

Smell Source	**EN**	**IT**	**FR**	**DE**
animal, people	143	64	99	66
cleaning, medicinal	86	44	83	40
emissions, traffic, fuel, dust	42	27	51	37
food, beverage	298	171	222	169
industry	60	40	62	28
nature, flowers, plant, tree, soil	243	116	166	111
synthetic	18	8	23	13
tobacco, smoke	30	12	21	15
waste, garbage, excrement, rotten	100	49	66	42
Sources total	1020	531	793	521
Quality	**EN**	**IT**	**FR**	**DE**
fragrant, fruity, floral	68	29	34	24
chemical, hydro-carbons, synthetic	16	19	29	5
decayed	51	30	35	9
fresh	9	8	6	5
pungent	52	11	19	18
smoky, toasted, burnt, fatty	17	10	8	6
sweet, spicy	32	18	20	17
woody, earthy, mouldy	43	15	20	19
Qualities total	288	140	171	103
Not classified	113	33	55	55
Total	**1421**	**704**	**1019**	**679**

Table 5: Number of terms for each language assigned to the categories included in 'Smell source' and 'Quality'. The former are all nouns, the latter all adjectives. Terms related to smell without a category are also reported.

earthy and mouldy category and 'disgusting' to the *decayed* one.

7. Multilingual Taxonomy

The final taxonomy contains and integrates data from different sources: the lists of seed terms, the WordNet synsets, the co-occurring terms (+ associated category) and the terms used in the literature to define the categories and to initialise the clusters. A summary of the taxonomy content is reported in Table 5. We display the number of terms in each language assigned to different categories of smell sources and of qualities through clustering and the number of terms (from the seed list or WordNet expansion) without a category.

The taxonomy is available at `https://github.com/Odeuropa/multilingualTaxonomies`. For each language, we release a table containing 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

6. for nouns, to which category of smell sources it was assigned (see Section 6)

7. for adjectives, to which category of qualities it was assigned (see Section 6)

Figure 3 provides a representation of the distribution of the categories over different time periods. For each of the four languages analysed, the graph displays whether a specific category, represented by the terms associated with it in the taxonomy, is present in that time span or not. If a term belonging to a specific category, for instance *decayed*, is mentioned in 1801-1850 and in 1851-1900, it is counted in both the respective bars in the graphs. This kind of analysis provides a characterization of the progress of linguistic variety

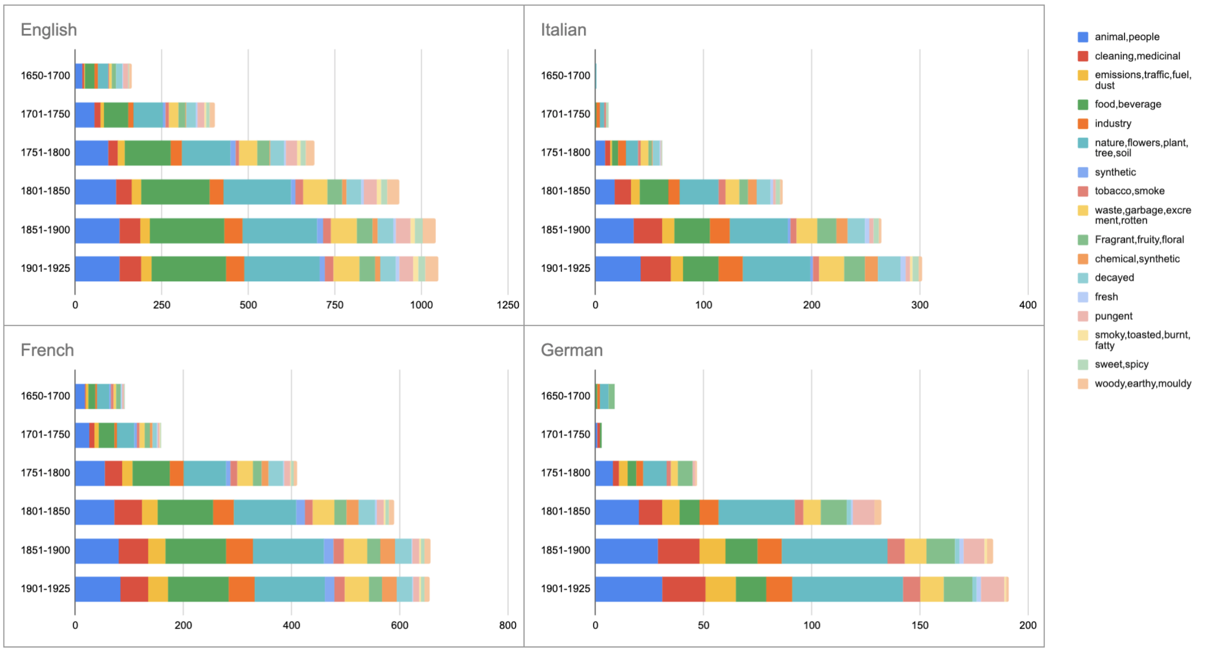


Figure 3: Distribution of content categories for each language in different time spans

used in odour descriptions, showing the presence of a specific category in smell-related mentions (considering n-grams as a proxy of the written texts published in each time span). As we can see from the chart, the proportion of the most relevant categories across languages is similar, with the exception of nature and flowers that are more present in German than in other languages starting from 1800. This category, together with food and beverage, is the most represented in texts. Also, terms describing smell sources are more present than smell qualities.

The bars grow over time because more n-grams are available, but also because the number of terms for each category increases. Future applications include the possibility to use this categorization as a general tool for the selection of specific olfactory descriptors in texts, including historical ones.

8. Conclusion

In this work we detail the semi-automatic development of a multilingual taxonomy about smell. The olfactory terms have been obtained from different sources, from structured resources to n-grams, and enriched whenever possible with information about their occurrence over time. Also, the outcome of each step aimed at adding more terms to the taxonomy has been manually validated. We release the taxonomy for future research, providing an example of usage related to tracking olfactory vocabulary over time. In the future, we plan to extend the taxonomy with more languages. Indeed, a similar work is already in progress for Slovenian, Dutch and Latin. However, the fact that for these three languages Google N-Gram is not available is a main limitation, which we plan to overcome by creating time-stamped n-grams from existing digital repositories. We also plan to integrate this work with ongoing creation of benchmarks with multilingual olfactory information (Menini et al., 2022), to assess whether merging taxonomy-based information with smell-related semantic roles can benefit the development of systems for olfactory information extraction.

9. Acknowledgements

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

10. Bibliographical References

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.

Bird, S., Loper, E., and Klein, E., (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

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

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, 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, June. Association for Computational Linguistics.

Classen, C. (1999). Other Ways to Wisdom: Learning Through the Senses Across Cultures. *International Review of Education*, 45(3-4):269–280, May.

de Araujo, I. E., Rolls, E. T., Velazco, M. I., Margot, C., and Cayeux, I. (2005). Cognitive modulation of olfactory processing. *Neuron*, 46(4):671–679.

Dugan, H. (2011). *The Ephemeral History of Perfume: Scent and Sense in Early Modern England*. Baltimore: Johns Hopkins University Press.

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

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)*.

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

Howes, D. (2006). Charting the sensorial revolution. *The Senses and Society*, 1(1):113–128.

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

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.

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

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

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

Lynott, D. and Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41(2):558–564.

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.

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

Majid, A. and Levinson, S. C. (2011). The senses in language and culture. *The Senses and Society*, 6(1):5–18.

Menini, S., Paccosi, T., Tonelli, S., Erp, M. V., Leemans, I., Lisena, P., Troncy, R., Tullett, W., Hürriyetoğlu, A., Dijkstra, G., Gordijn, F., Jürgens, E., Koopman, J., Ouwerkerk, A., Steen, S., Novalija, I., Brank, J., Mladenić, D., and Zidar, A. (2022). A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the Third Workshop on Computational Approaches to Historical Language Change (LChange'22)*.

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.

Quercia, D., Schifanella, R., Aiello, L., and McLean, K. (2015). Smelly maps: The digital life of urban smellscapes. In *Proceedings of ICWSM*.

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), Mar.

Rodriguez-Esteban, R. and Rzhetsky, A. (2008). Six senses in the literature. The bleak sensory landscape of biomedical texts. *EMBO reports*, 9(3):212–215, March.

Smith, M. M. (2006). *How race is made: Slavery, segregation, and the senses*. Chapel Hill : The University of North Carolina Press.

Stevenson, R. J. and Case, T. I. (2005). Olfactory imagery: a review. *Psychonomic Bulletin & Review*, 12(2):244–264.

Tekiroğlu, S. S., Özbal, G., and Strapparava, C. (2015). Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado, June. Association for Computational Linguistics.

Tonelli, S. and Menini, S. (2021). FrameNet-like annotation of olfactory information in texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 11–20, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Tullett, W. (2016). Grease and sweat: Race and smell in Eighteenth-Century English culture. *Cultural and Social History*.

Tullett, W., (2019). *Smell in Eighteenth-Century England*. Oxford University Press.

Winter, B. (2016). Taste and smell words form an affectively loaded and emotionally flexible part of the english lexicon. *Language, Cognition and Neuroscience*, 31(8):975–988.

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

11. Language Resource References

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

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*.

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

Pianta, E., Bentivogli, L., and Girardi, C. (2002). MultiWordNet: Developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Sagot, B. and Fišer, D. (2008). Building a free French WordNet from multilingual resources. In *OntoLex*.

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, October. Association for Computational Linguistics.