Building a Taxonomy of Olfactory Terms with Timestamps

Stefano Menini[1], Teresa Paccosi[1,2], Serra Sinem Tekiroğlu[1], Sara Tonelli[1]
[1]Fondazione Bruno Kessler, [2]Università di Trento
{menini, tpaccosi, tekiroglu, satonelli}@fbk.eu

FONDAZIONE BRUNO KESSLER

UNITRENTO

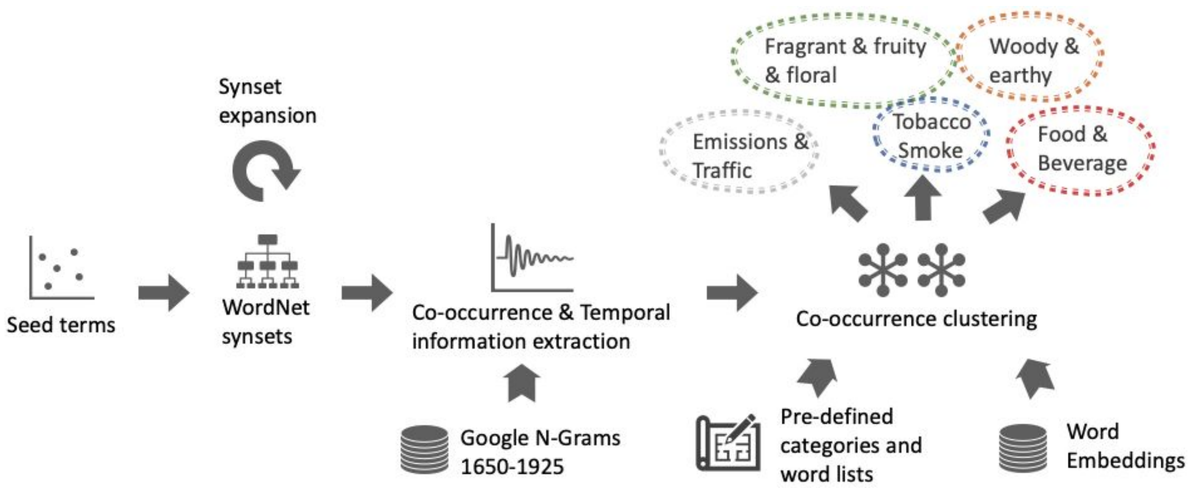
Goal:

- Creation of an **olfactory taxonomy**. Timespan f**rom 1650 to 1925**.
- Automatically identify how smell experiences are described.
- Accounting for differences in sensory vocabulary across languages and over time (of great interest for cultural studies, digital humanities and historical content analysis).

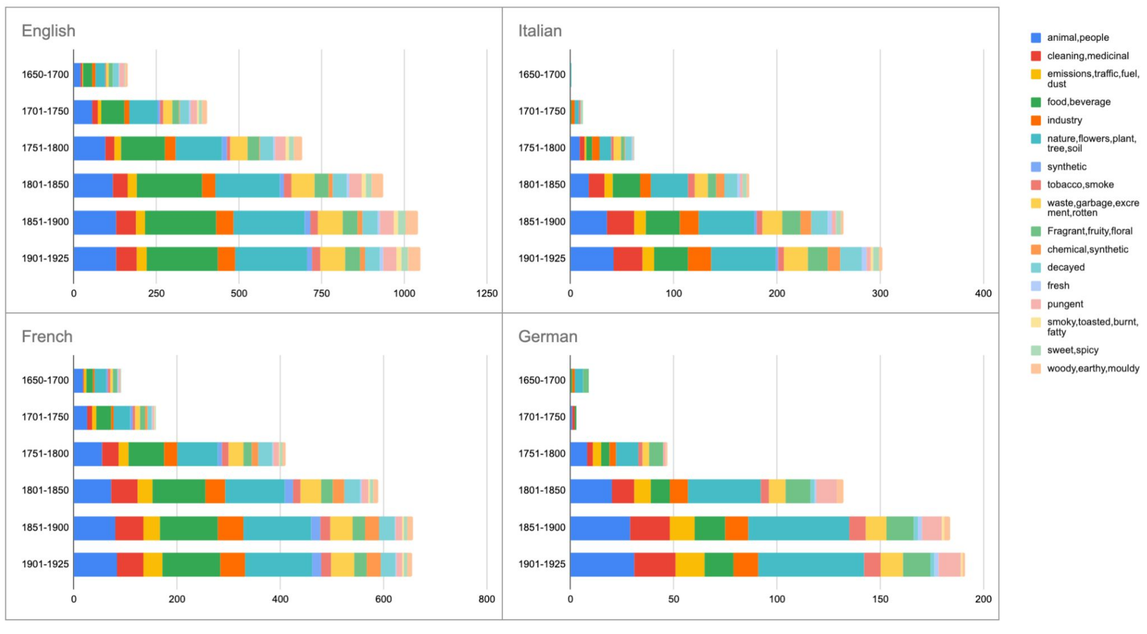
Languages

- English
- Italian
- French
- German

Pipeline:



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
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**



Seed List

- Words (lemmas) unambiguously related to the olfactory domain
- Selected by domain experts for each language of interest

Google N-grams

- For each seed term, we extract the terms that are most associated with it (**co-occurrences**)
- Words in Google n-grams are annotated with **PoS**
- **Temporal information** (i.e. date of occurrence)

WordNet

- We extract all synsets containing a seed terms
- Manual cleaning to retain only **synsets** actually **related to smell** (around 50 for each language)
- Automatic expansion through WordNet relations

Creation of Categorie and Lists

- Manual harmonisation of existing lists & taxonomies
- Lists of **Qualities** and **Smell Sources**
- We manually translate these lists to **Italian**, **French** and **German**.

Term Clustering

Assign the co-occurrence from the n-grams and WordNet to one of the Source or Quality Categories

- **Terms** in categories as **word embeddings** (fastText)
- Categories as **clusters** of embeddings.
- Assignment based of words to clusters based on cosine distance between the term embedding and the centroid of the cluster (minimum threshold required).
- **Nouns** are Smell Sources, **adjectives** are Qualities.
- Clustering step is repeated twice
- Manual post-hoc evaluation of the clusters using different similarity thresholds.

Taxonomy available at:

https://github.com/Odeuropa/multilingualTaxonomies

Co-funded by the Horizon 2020 programme of the European Union

Odeuropa

DH DIGITAL HUMANITIES