Multilingual Text Collection and Annotation

Stefano Menini & Sara Tonelli
Fondazione Bruno Kessler, Trento (Italy)
ODEUROPA Text Analysis

Goal: Extract and process relevant olfactory information from unstructured texts in multiple languages:

• Which contexts, places and events were mentioned in relation to olfactory information
• Which terminology was used
• How this information changed over time and across languages
Creating a system for Olfactory Information Extraction

- Based on **supervised** approach: system learns what to extract from existing examples

  Manual annotation!
  Collect and annotate a large number of text passages to train an information extraction system
Olfactory Information Extraction

• **What texts** should be annotated?
  Which time period, domains, languages

• **What information** should be annotated?
  What characterises olfactory events
Olfactory Information Extraction

• **What texts** should be annotated?
  Which time period, domains, languages

**Domains of interest:** Household & Recipes, Law and Regulations, Literature, Medicine & Botany, Perfumes & Fashion, Public health, Religion, Science & Philosophy, Theatre, Travel & Ethnography

~10 documents for each category, evenly distributed between 1620 and 1920, for a total of ~100 documents for each language.
Olfactory Information Extraction: Benchmark
Olfactory Information Extraction: Benchmark
### Olfactory Information Annotation: Smell words

**English:** aroma.n, bouquet.n, essence.n, aromatic.a, fetid.a, inodorous.a, reek.v, smell.v, sniff.v, stink.v, [...]

**Italian:** aromatico.a, fetido.a, fragrante.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 [...]

**Dutch:** reuk.n, geur.n, aromatiseren.v, stinken.v, ontgeuren.v, balsemachtig.a, reukstof.n [...]

**Slovenian:** vonj.n, smrad.n, dišava.n, smrdeti.v, zaudarjati.v, zavonjati.v, vohati.v, smrdljiv.a [...]


Olfactory Information Annotation: Semantic roles

**Smell Source:** The person, object or place that has a specific smell.

**Quality:** A quality associated with a smell and used to describe it.
*The {odour} [of tar] and [pitch] was [so strong].*

**Odour Carrier:** The carrier of an odour, either an object (e.g. handkerchief) or atmospheric elements (wind, air)
*The unpleasant {smell} [of the vapour] of linseed oil extended for a considerable distance.*

**Perceiver:** The being that perceives an odour, who has a perceptual experience, not necessarily on purpose.
*The {scent} is described by [Dr. Muller] as delicious.*

**Location:** The location where the smell event takes place.

**Time:** An expression describing when the smelling event occurred.
*And, particularly, [at the foot of the garden], the {smell} [at night] was so offensive that has sickened him.*

**Effect:** An effect or reaction caused by the smell.
*An ill {smell} [gives a nauseousness].*
I do my Lord, and haue at my command, The smell of flowers, and Odoriferous drugs,

Of oyntments sweete, and excellent perfumes, And Court-like waters, which if once you smell,

You in your heart would wish as I suppose: That all your Body were transsformed to Nose.
One morning the Artist had paused a moment to make a rough sketch of a man who [...] was selling segments of round flat cheeses of goat's milk; vile smelling compounds that, judged from their outer coating of withered leaves, straw, and dirt, would appear to have been made in a stable and dried on a rubbish heap.

In its pure state, civet has, to nearly all persons, a most disgusting odor; but when diluted to an infinitesimal portion, its perfume is agreeable.
In general, Cabbages are thought to allay fumes, and prevent intoxication: But some will have them noxious to the Sight

Fumes | Cabbages | noxious
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101004469.