Tracing Taste over Time: Automatic Extraction and Diachronic Analysis of Gustatory Language in English

Teresa Paccosi*
DHLab, KNAW Humanities Cluster

Sara Tonelli**
Fondazione Bruno Kessler

In this paper, we present a benchmark of texts manually annotated with gustatory information, following a FrameNet-like approach previously applied to olfactory language and here adapted to capture taste-related events. We explore the benchmark to illustrate the possible insights this approach can offer, focusing in particular on the expression of emotional valence across different textual genres. Building on this resource, we train a supervised system for the automatic extraction of gustatory information from both historical and contemporary texts. The system is then applied to a variety of corpora, and we provide a publicly available notebook for exploring the extracted data, along with an analysis of the system's output in the literary domain.

1. Introduction

Despite the central role of nutrition in our lives, taste has been often classified as an inferior sense in the Western philosophical tradition. This downplayed role is reflected in the vocabulary used to describe the gustatory experience, which, together with smell, is characterized by a scarcity of domain-specific terms (Winter 2019). The difficulty in capturing the semantics of taste could help explain why there are few works in the fields of Natural Language Processing (NLP) and Digital Humanities (DH) that deal with this sense and, in particular, with the language used to describe its experience. While there has been renewed interest in the automatic extraction of nutrients and ingredients from texts for health and medicinal purpose (Bagler and Goel 2024), less attention has been devoted to the development of tools and models focused on capturing the semantics of sensory experiences, especially in a diachronic fashion. In this paper, we present an English benchmark for the semantics of gustatory language, a supervised system for the automatic extraction of taste-related events trained on this resource, and a preliminary large-scale analysis of automatically extracted data. The benchmark was developed as the gustatory counterpart to the olfactory resource introduced in (Menini et al. 2022), with the aim of enabling historical and contemporary comparative analyses between these two sensory domains. Our approach is grounded in Frame Semantics (Fillmore 1976), with the system designed to identify lexical units and the semantic roles involved in the construction of gustatory events. We report the results of classification experiments and explore the benchmark data to demonstrate the value of this semantic approach for cross-modal comparisons. To illustrate the applicability of the system beyond the benchmark, we apply it to several English corpora, extracting gustatory

* DHLab, KNAW Humanities Cluster, Oudezijds Achterburgwal 185, 1012 DK Amsterdam, the Netherlands. E-mail: teresa.paccosi@dh.huc.knaw.nl
** Fondazione Bruno Kessler, Via Sommarive, 18, 38123 Povo TN, Italia. E-mail: satonelli@fbk.eu

Table 1
List of Gustatory Frame Elements

Frame Element	Definition
Taste_Source	The food items that are ingested
Quality	Any property used to describe the taste (usually adjectives)
Taste_Carrier	Anything that can contain the taste source
Taster	The person/animal who ingests the food
Evoked_Taste	The taste that is evoked but it is not present (e.g., it tastes like onions)
Location	The place in which the food is tasted
Taste_Modifier	An ingredient that can modify the perception of the taste of a taste source
Circumstances	The condition or circumstance in which the taste event occurs
Effect	Any effect provoked by the tasting experience

information at scale. A case study based on literary texts highlights the system's potential to support sensory research with quantitative insights. Finally, we adapt the system for use through a web interface, making it accessible to scholars without a technical background and facilitating its application to historical studies.

2. Related work

In recent years, there has been a growing interest within the NLP community in developing resources designed to capture the sensory content of language (Tekiroğlu, Özbal, and Strapparava 2014). In particular, in the framework of the three-year European Project "Odeuropa"[1] aimed at preserving intangible cultural heritage, several works have focused on analyzing smell descriptions (Brate, Groth, and van Erp 2020) and extracting olfactory information from texts. For instance, (Menini et al. 2022) created a manually annotated benchmark with smell events, which has been subsequently used to train a system for olfactory information extraction (Menini et al. 2023; Menini 2024). The benchmark focuses on the language used to describe olfactory experiences and covers a period of four centuries (1600-1900), making it useful for historical research. An extension in this direction is SENSE-LM (Boscher et al. 2024), a system for extracting sensory information from texts, which shows that combining language models with lexical resource-based approaches yields better results in extracting sensory references from texts compared to systems that do not integrate these two components. The authors were the first to combine sensorimotor representations with the textual features of language models for the task of sensory information extraction. Even if they propose the system for all the five senses, they only tested it on olfactory and auditory language, using respectively the benchmark of (Menini et al. 2022) and an artificial dataset they generated with GPT-4 (AI 2023). Most existing work on food representation in the field of NLP focuses on health-related applications. A notable exception adopting a linguistic focus is (Magnini et al. 2018), where the authors deal with identifying noun-compound head-nouns for developing conversational agents in the e-commerce domain. They propose a supervised approach based on a neural sequence-to-sequence model to identify the most informative token in Italian food compound-nouns, obtaining promising results despite the complexity of the task. Taste has been also addressed from a diachronic point of view in (Jurafsky 2014), in which the author reconstructs the evolution of food

1 https://odeuropa.eu/

language focusing on the history of some dishes and ingredients across continents using computational linguistic tools. To our knowledge, however, no model has yet been specifically developed to address the language of taste, particularly with a focus on its diachronic dimension. The vast majority of studies in the emerging field of *computational gastronomy* (Goel and Bagler 2022; Bagler and Goel 2024) has focused on the development of Named-Entity Recognition (NER) models to automatically extract food entities for medicinal purposes and food science applications (Cenikj et al. 2020; Stojanov et al. 2021). These efforts often involve creating domain-specific corpora by sourcing data from culinary websites and online recipe books (Popovski, Seljak, and Eftimov 2019; Wróblewska et al. 2022). While most work on computational gastronomy focuses on entity extraction, the diachronic variation of the dimension of taste has received little attention. In contrast, in the realm of diachronic computational studies, semantic change detection occupies a central role. The study of variation in meaning over time has been extensively explored with a variety of methodologies and benchmarks for analyzing how word meanings evolve over time. For example, the DURel framework (Schlechtweg et al. 2024) measures semantic proximity and sense clustering through pairwise comparisons of contextualized instances. These resources have been widely used to evaluate models for lexical semantic change detection (Cassotti et al. 2023), as well as approaches based on definition generation that aim to obtain interpretable sense representations (Giulianelli et al. 2023). While these studies primarily focus on changes in individual lexical items, our work adopts a complementary perspective by representing structured sensory events. Although the resource is not specifically designed for traditional semantic change detection, the developed model enables the collection of large amounts of fine-grained data capturing the full structure of gustatory events, as well as domain-specific data that can be used to train specialized embeddings. This makes it possible to conduct quantitative analyses of how such events are represented over time. Moreover, the event-based annotation allows researchers to tailor their investigations to specific historical interests, focusing on particular elements of the event structure (e.g., the taste source or the taster). In this way, the resource supports targeted diachronic analyses of gustatory language change. A first version of this work was introduced in (Paccosi and Tonelli 2024a). In this extension, we present some novel contributions. We begin by applying our system to extract taste-related data from a collection of historical English corpora. A subset of this data is then selected for analysis, demonstrating the potential of our tools to support research in the humanities. As a case study, we examine the representation of *tea* in literary texts from the 18th to the 20th century (Section 7). Finally, we introduce a novel web interface (Section 8) designed for scholars without a technical background, aimed at facilitating qualitative historical research through quantitative insights.

3. Benchmark for Taste

The training data we use for the models in this paper is a benchmark created according to the annotation guidelines presented in (Paccosi and Tonelli 2024b). The formalization adopted to annotate the benchmark is inspired by Frame Semantics (Fillmore 1976) and its implementation through the FrameNet annotation project (Ruppenhofer et al. 2016). In FrameNet, events and situations are constructed as *frames*, structures that represent the knowledge necessary to understand the meaning of words. Frames include two main components, namely **lexical units**, domain-specific words or expressions that trigger the frame, and **frame elements**, domain-specific semantic roles usually attached as dependents to the lexical unit. In our case, taste events are captured through a so-

Table 2
Lexical units for Taste

Part of Speech	Lexical Units
Nouns	Acidity, aftertaste, aroma, bitterness, dainty, delicacy, disgust, distaste, flavor, flavour, flavorful, flavourful, flavoring, flavouring, flavorsome, flavoursome, flavorous, flavourous, gustation, insipidity, mistaste, over-eating, palatableness, piquancy, pungency, rancidity, relish, rellish (obsolete), saltness, sapidity, sapor, savor, savoriness, savour, sharpness, smack, smatch, sourness, sowreness (archaic form of sourness), sweetness, tang, tarage, tartness, tast (obsolete), taste, tastelessness, tasting, unsavoriness, unsavouriness
Adjectives	Acid, acidic, appetizing, appetizing, bitter, bitter-sweet, bland, dainty, delectable, delicious, delightsom(e), disgusting, flavorless, flavorful, flavourful, flavourless, flavoursome, gamy, indigestible, insipid, juicy, mellow, palatable, piquant, pungent, racy, rancid, rank, salt/salty, sapid, savory, savoury, savourly, seasoned, sharp, sour, soured, sower (archaic form of sour), spicy, stale, sweet, tangy, tart, tasteless, tasty, toothsome, unpalatable, unsavor, unsavour, unsavoury, unsavory, unseasoned, unsweet, unsweetened, wearish, wersh, yummy
Verbs	Drink (up), drinking (up), drank (up), drunk (up), eat (up), ate (up), eateth (archaic), eaten (up), eating (up), distaste, distasting, distasted, mistaste, mistasted, mistasting, partake, partaking, partook, partaken, relish, relisheth (archaic), relishing, relished, season, seasoning, seasoned, smack, smacking, smacked, smatch (obsolete), sweeten, sweetening, sweetened, taste, tasting, tasted
Adverbs	Sweetly, sourly, tastefully, bitterly, tastingly, unsavourily, unsavourly, insipidly, savourously, savourily, flavourfully

called *Gustatory Frame*, which is triggered in a document by Taste_Words (i.e., taste-specific lexical units). Each lexical unit is annotated in the benchmark together with the frame elements associated with it, which the taste extraction system should then identify automatically. For instance, in the sentence "[Slimy milk]$_{Taste\_Source}$ has an [unpleasant]$_{Quality}$ taste", the system has to identify the Taste_Word ('taste'), and then the possible frame elements (in this case, Taste_Source and Quality). A list of the possible frame elements and their definition is provided in Table 1, while in Table 2, we display the list of lexical units (or taste words) presented in (Paccosi and Tonelli 2024b).

The documents annotated in the benchmark cover five different domains or genres, almost evenly distributed with three or four documents for century in every domain for a total of 72 documents. The genres are: *Literature*, *Science & Philosophy*, *Household & Recipes*, *Travel & Ethnography*, and *Medicine & Botany*. To select the documents we automatically search for texts presenting a greater density of lexical units (taste words) spanning through several English corpora and taste-related websites. The corpora from

which we extract the documents we annotate are: (1) *Early English Books Online (EEBO)*,[2] a collection of documents published between 1475 and 1700 covering different domains such as literature, philosophy, politics, religion, geography, history, politics, and mathematics; (2) *Project Gutenberg*,[3] a digitized archive of cultural works, containing different repositories, mainly in the literary domain; (3) *medievalcookery.com*,[4] a list of texts freely available online relating to medieval food and ancient cooking recipes; (4) *foodsofengland.co.uk*,[5] an online library which holds the complete texts of several cook books from 1390 to 1974; (5) *Wikisource*,[6] an online digital library of free-content textual sources managed by the Wikimedia Foundation; (6) *British Library*,[7] a collection of 65,227 digitized volumes from the 16th to the 19th Century; (7) *London Pulse Medical Reports*,[8] a collection of 5,800 Medical Officer of Health reports from the Greater London area from 1848 to 1972.

In Table 3 we report the statistics of the annotated benchmark (note that in (Paccosi and Tonelli 2024b) we presented only a preliminary version of the benchmark containing around 1,400 Taste_Words). The most frequent frame element is the Taste_Source, followed by Quality and Taste_Modifier, which represent the core and most representative frame elements, while the rest of the frame elements are much sparser. Even if the distribution of the frame elements is not balanced, the system is trained to extract the taste words and all the 9 frame elements. Two expert linguists, trained on (Paccosi and Tonelli 2024b)'s guidelines, annotated three documents from 1670, 1720, and 1920 to assess Inter Annotator Agreement (IAA). The Krippendorff's alpha score (Krippendorff 2011) at span level was 0.70. While this score suggests a reasonable level of consensus, there remains potential for improvement. However, upon closer examination of the discrepancies, it is observed that there is a general agreement regarding the choice of labels. The disagreement arises from inconsistencies in the selection of spans, particularly in the exact number of tokens encompassed within those spans, as seen in the following instance, where the label `Taste_Source` is correct but the tokens are selected in a different way:

- Season it with savoury spice, [a little minced sage]$Taste\_Source$, and sweet herbs
- Season it with savoury spice, a [little minced sage]$Taste\_Source$, and sweet herbs

We believe that, as a result, the agreement is actually higher than the one indicated by Krippendorff's alpha. In the future, we plan to manually correct these types of discrepancies or to consider alternative ways to evaluate text spans beside exact match, for instance by computing the cosine similarity between correct and incorrect instances.

2 https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/
3 https://www.gutenberg.org/
4 https://www.medievalcookery.com/etexts.html?England
5 http://www.foodsofengland.co.uk/references.htm
6 https://en.wikisource.org/wiki/Main_Page
7 https://data.bl.uk/digbks/
8 https://wellcomelibrary.org/moh/about-the-reports/about-the-medical-officer-of-health-reports/

Table 3
Statistics of the Taste Benchmark

Frame Elements (FEs)	1500	1600	1700	1800	1900	Overall
Taste_Words	440	2,417	500	1,498	803	5,648
Taste_Source	372	1,627	375	1,081	599	4,393
Quality	197	1,495	255	881	489	1,732
Taste_Modifier	135	142	66	154	78	1,357
Taster	65	173	85	185	100	638
Evoked_Taste	20	127	31	53	16	247
Location	11	44	12	24	16	116
Taste_Carrier	9	38	9	26	12	98
Circumstances	19	206	38	228	82	656
Effect	24	56	32	34	31	174

4. Exploration of olfactory and gustatory benchmarks

To demonstrate the potential of a frame-based analysis for cross-domain comparisons across sensory modalities, we conduct a comparative study of positive and negative contents conveyed by smell and taste-related language. It has been observed indeed that words used to describe olfactory and gustatory experiences tend to appear more frequently in emotionally charged contexts and carry a stronger evaluative content compared to words related to other senses (Winter 2016). By *evaluative content*, we refer in this paper to the concept of *emotional valence*, which is defined as "the pleasantness of a word in terms of positive and negative meaning" ((Winter 2019), p. 201). We therefore conducted an exploration of the gustatory benchmark to investigate the positive and negative connotations of gustatory events across different text genres. We perform the same analysis for olfactory events, using the olfactory benchmark of (Menini et al. 2022) in order to compare the outcome for the two senses.

This benchmark contains 85 documents from 16th to 19th century and covers several topics. As reported in Section 3, the genres represented in the gustatory benchmark are: Literature, Science & Philosophy, Household & Recipes, Travel & Ethnography, Medicine & Botany. In the olfactory benchmark presented in (Menini et al. 2022), there are instead ten different genres: Household & Recipes, Law & Regulations, Literature, Medicine & Botany, Perfumes & Fashion, Public health, Religion, Science & Philosophy, Theatre, Travel & Ethnography. For the present analysis, we decide to select and compare the representation of taste and smell in the genres which are shared by the two senses. The olfactory benchmark was annotated with smell-related information based on Olfactory events, which were defined in a similar way as the Gustatory frame. Similarly to the Gustatory Frame, the Olfactory event is indeed evoked by olfactory-specific terms such as "smell", "odour", or "perfume", to which various semantic roles can be assigned. This configuration allows us to compare and analyze the elements that contribute to the realization of these sensory events in a consistent manner, focusing our investigation on frame elements that can be considered counterparts across the two senses, such as Quality, which in both cases refers to the property used to describe the sensory experience.

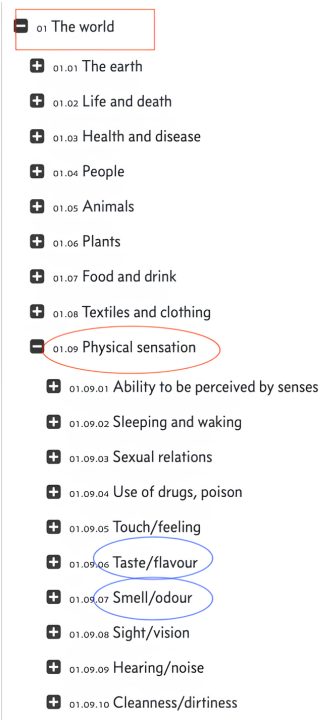

Figure 1
Thesaurus hierarchy for smell and taste

To perform this analysis, we first divide Taste_Words (i.e. lexical units from the Gustatory frame) and Smell_Words (i.e. lexical units from the Olfactory frame) into *positive* and *negative*.

To this purpose, we rely on the the Historical Thesaurus of English, which documents the history of each word, including its first recorded use, when they fall into disuse, and its synonyms across various time periods. In this resource, words are organized hierarchically by meaning, enabling users to explore the evolution of English vocabulary across different semantic categories. For our analysis, we select categories related to the semantic fields of taste and smell. These two senses appear as separate branches within the same hierarchy, under The World and then Physical Sensation classes (see Fig. 1). Within the olfactory domain, the category *Smell/Odour* is divided into three subcategories: Stench, Fragrance, and Inodorousness. The gustatory domain, *Taste/Flavour*, includes five subcategories: Unsavouriness, Savouriness, Insipidity, Sweetness, and Sourness/Acidity (see Fig. 2). For this study, we treat Unsavouriness as the negative category for taste, Savouriness as positive, and Insipidity as representing a lack of taste. To focus on positive and negative connotations, we extend the Savouriness category to encompass all terms in the Sweetness category, while grouping all terms from Sourness/Acidity under Unsavouriness. These categories reflect indeed positive and negative connotations, as demonstrated by their frequent use in conceptual metaphors related to taste expressing emotional states (Bagli 2021). For example, in the conceptual metaphor `EMOTIONAL FEELINGS ARE TASTE`, sweetness refers to pleasure, whereas bitterness and sourness symbolize emotional suffering.

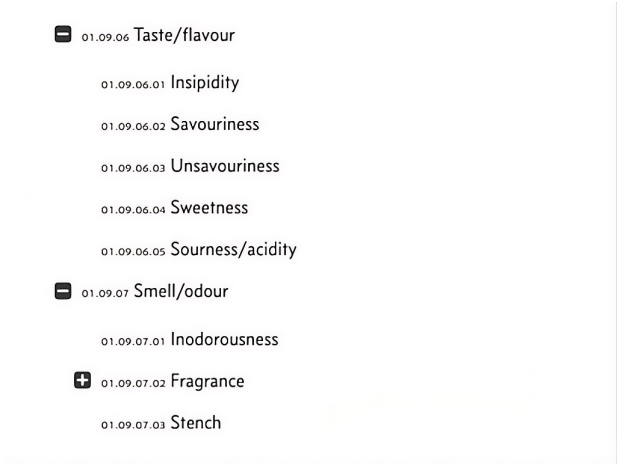

Figure 2
Detailed Thesaurus hierarchy for smell and taste

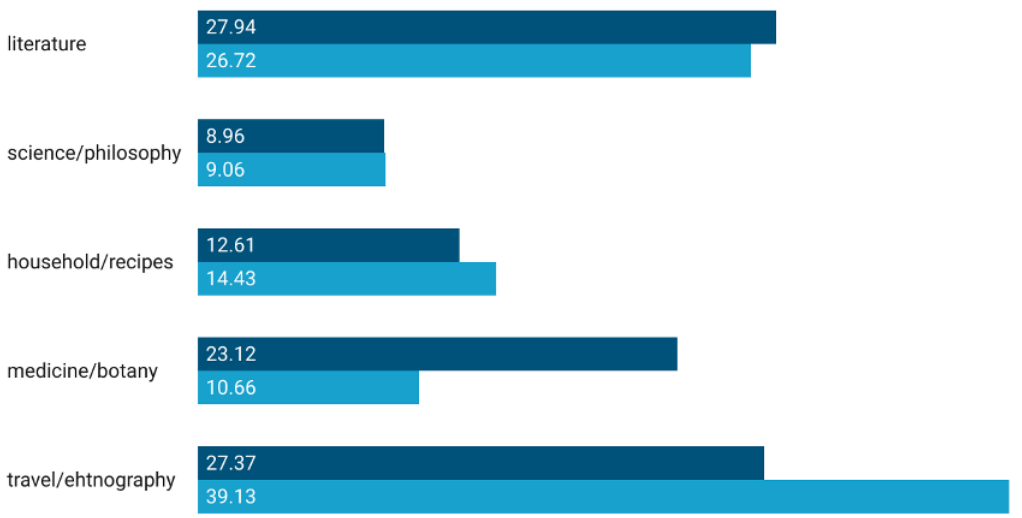

Figure 3
`Savoury` (dark blue) and `Unsavoury` (light blue) frequencies of taste words in genres

In the Thesaurus, every category of the hierarchy is divided per part of speech (PoS). For our analysis, we manually select all the nouns, adjectives and adverbs used in the period we cover with our documents, namely from 16th century to 20th century, in the Taste and Smell categories. We then assign the words labeled as Taste_Words and Smell_Words in the documents to one of the two categories (positive or negative) and calculate the normalized frequency of each category across different text genres.

We display the output of this analyses in Fig. 3 (for taste words) and Fig. 4 (for smell words), aimed at showing which emotional valence prevails in each genre for the two senses. We observe that two genres exhibit opposite tendencies: `medicine/botany` shows a more negative orientation in the smell benchmark and a more positive one

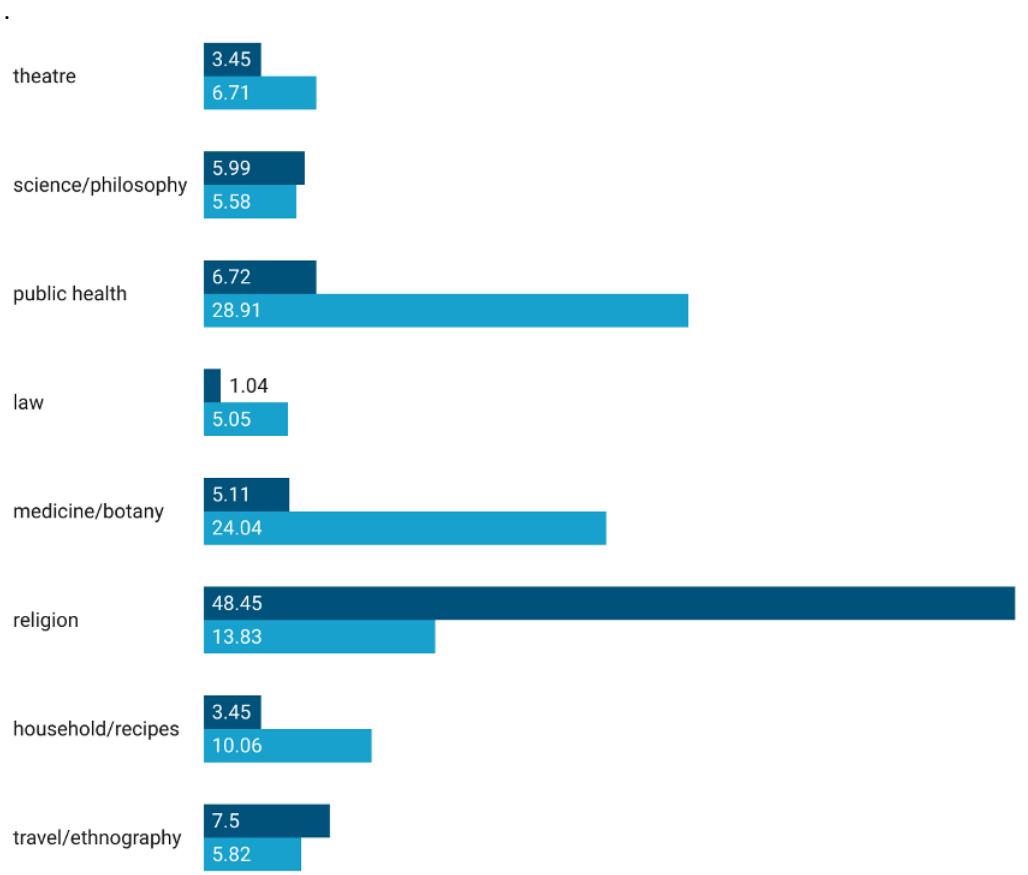

Figure 4
`Fragrant/Fragrance` (dark blue) and `Stench` (light blue) frequencies of smell words in genres

in the taste benchmark, whereas `travel/ethnography` is more positive concerning smell and more negative for taste (see Fig. 3 and Fig. 4, where the light blue refers to negative valencies and the dark blue to positive ones). We analyze the most frequent smell / taste sources in the two selected genres to motivate why they exhibit such difference in emotional valence. We notice that smell sources in `medicine/botany` tend to be common to hospital and disease-related domains having words such as "urine" and "fetid bronchitis", while taste sources more easily belong to the realm of common food, with words such as "almonds" and "apples". For what concerns `travel/ethnography` instead, among the most frequently described taste sources there are exotic and rare foods such as "coconut" and "plantain", likely resulting unpleasant to the palates of foreign travelers. Smell sources tend to refer instead to plants, like "flowers" or "roots", hence usually pleasant or neutral to the noses of the writers. This analysis of categories and sources' distribution in the genres underlines the importance of a frame-base analysis for understanding and comparing sensory descriptions, in particular their emotional valence.

Table 4
Results (F_1) of the classifiers on the lexical unit (T_Word) and 9 frame elements with multitask (first) and single (second) configurations, with the latter in italics. The results are the average of the F_1 results of each label across the 5 folds. The labels are reported as follow: Taste_Word (1), Taste_Source (2), Quality (3), Circumstances (4), Effect (5), Evoked_Taste (6), Location (7), Taste_Carrier (8), Taste_Modifier (9), Taster (10), Overall (11).

Model	1	2	3	4	5	6	7	8	9	10	11
BERT	0.917	0.537	0.780	0.413	0.196	0.457	0.379	0.111	0.781	0.518	0.509
BERT	*0.903*	*0.530*	*0.712*	*0.308*	*0.019*	*0.254*	*0.206*	*0.0*	*0.681*	*0.434*	*0.405*
mBERT	0.919	0.554	0.784	0.402	0.180	0.466	0.357	0.087	0.763	0.511	0.502
mBERT	*0.910*	*0.557*	*0.740*	*0.284*	*0.0*	*0.304*	*0.162*	*0.0*	*0.694*	*0.434*	*0.409*
MacBERTh	**0.943**	0.580	0.799	0.444	**0.285**	**0.501**	0.338	0.093	0.783	0.512	0.528
MacBERTh	*0.909*	*0.548*	*0.720*	*0.366*	*0.021*	*0.226*	*0.242*	*0.0*	*0.688*	*0.455*	*0.418*
RoBERTa	0.913	0.558	0.786	0.414	0.219	0.473	0.406	0.094	0.772	0.508	0.514
RoBERTa	*0.891*	*0.553*	*0.726*	*0.343*	*0.0*	*0.33*	*0.228*	*0.0*	*0.726*	*0.5*	*0.430*
RoB.-xlm	0.932	**0.587**	**0.817**	**0.452**	0.279	0.497	**0.416**	**0.105**	**0.784**	**0.563**	**0.543**
RoB.- xlm	*0.903*	*0.601*	*0.777*	*0.4*	*0.021*	*0.409*	*0.25*	*0.0*	*0.743*	*0.539*	*0.464*

5. System for Gustatory Information Extraction

The benchmark introduced in the previous sections is used to train a classifier whose goal is to detect gustatory information in English texts. The system is based on multi-task learning (Section 5.1), and is then compared with a "single task" classifier, which we consider our baseline (Section 5.2).

5.1 Multitask configuration

To build our system for gustatory information extraction, we adopt a multitask learning approach (Caruana 1993, 1997), a configuration successfully tested for olfactory information extraction in (Menini et al. 2023; Menini 2024). This approach treats the classification of lexical units and each frame element as different tasks. Additionally, we explore a "single task" classification approach, where both lexical units and frame elements are classified within a multiclass token classification task. The results of these experiments serve as a baseline for evaluating the effectiveness of the multitask approach. In both configurations, we employ a transformer-based model fine-tuned for a token classification task (Vaswani et al. 2017). This methodology has proved effective across various NLP tasks, including olfactory information extraction (Menini 2024) and the extraction of food-related ingredients (Stojanov et al. 2021). We fine-tune each model using the same data, maintaining identical training, validation, and test splits, and evaluate them using 5-fold cross-validation. Each fold contains 80% of the lexical units and their related frame elements for training, 10% for validation, and 10% for testing. These splits are consistent across all configurations and stratified to ensure a balanced distribution of frame elements in every run. For labeling the data, we adopt the IOB (Inside-Outside-Beginning) labeling format, as used in (Menini et al. 2023; Menini 2024). This method facilitates a comprehensive analysis of sentences and lexical expressions by labeling each token with either "Inside', "Outside", or "Begin" labels as appropriate. To fine-tune the models, we use MaChAmp (Van Der Goot et al. 2021), a specialized toolkit designed for multi-task fine-tuning scenarios. In this approach, each label classification is treated as a distinct task. This setup ensures that simpler tasks, such as recognizing lexical units, contribute as auxiliary tasks to more complex

Table 5
Hyperparameter value used for the experiments which yield the best results

Hyperparameter	Value
$\beta 1$, $\beta 2$	0.9, 0.99
Dropout	0.2
Epochs	20
Batch Size	32
Learning Rate (LR)	0.0001
Decay Factor	0.38
Cut Fraction	0.3
All tasks loss weight	1

label classifications like Circumstances or Effect, which include entire sentences rather than individual words. MaChAmp enables the choice of different parameters, such as loss weight, epochs and batch size, and we tested different configurations.[9] The results in Table 4 for the multitask approach share the configuration which yield the best results. The configuration is the same for all the models and the hyperparameter setting for all our models is presented in Table 5. The setting is the default MaChAmp's hyperparameter values, with the addition of loss weights at 1, and 20 epochs of training.

5.2 'Single Task' configuration as Baseline

Similar to the system for smell information extraction presented in (Menini 2024), we design our baseline approach as a single-task multiclass classification, where the model assigns one of 21 possible labels to each token. These labels include 20 representing either "Begin" or "Inside" of each lexical unit and frame element, and 1 label representing "Outside". As we did for the multitask approach, each model is fine-tuned with a token classification head on top.[10] During the training of each model, hyperparameter search was conducted on the first fold of our data. The search space included learning rates $[1e-5, 2e-5, 3e-5, 4e-5, 5e-5]$, batch sizes $[8, 16, 32]$, and training epochs up to 20, with warmup applied for 10% of the training steps. After determining the optimal hyperparameters for each model, it is fine-tuned five times, each time with a different data fold, and the average scores were computed. We present the results of for the single task approach of each model in italics in Table 4.

5.3 Models Description

We experiment the two configurations with monolingual and multilingual versions of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019) and RoBERTa (Robustly Optimised BERT Pretraining Approach) (Liu et al. 2019), as well as a monolingual historical model, MacBERTh (Manjavacas Arévalo and Fonteyn 2021). The models we use are listed below:

9 Loss weight with different combinations over the labels $[1, 0.75]$, epochs $[10, 20, 30]$, and batch size $[16, 32]$.
10 `https://huggingface.co/docs/transformers/tasks/token_classification`

- **English BERT**: bert-base-cased[11] (Devlin et al. 2019)
- **Multilingual BERT (mBERT)**: bert-base-multilingual-cased[12] (Devlin et al. 2019)
- **English historical model**: MacBERTh[13] (Manjavacas Arévalo and Fonteyn 2021)
- **English RoBERTa**: roberta-base[14] (Liu et al. 2019)
- **Multilingual RoBERTa (RoBERTa xlm)**: xlm-roberta-base[15] (Conneau et al. 2019)

All models are transformer-based encoders that generate contextually-aware token representations through bidirectional attention mechanisms. BERT is pre-trained on BooksCorpus and English Wikipedia (∼16 GB), while RoBERTa extends this with a substantially larger and more diverse corpus (∼160 GB), including Common Crawl News, OpenWebText, and STORIES. A key architectural difference lies in the masking strategy: BERT applies static masking during pre-training, whereas RoBERTa adopts dynamic masking, varying the masked tokens across training steps to improve generalisation. To ensure comparable model sizes, we opt for the base versions when available and in all cases they consist of 12 transformer layers with 768-dimensional token representations (∼110M parameters), except for RoBERTa-xlm which uses a large configuration (∼270M parameters). MacBERTh shares BERT architecture but is pre-trained on a curated corpus of historical English texts (1450-1950), making it particularly suited for diachronic language tasks. For our experiments, we select these models based on their strong performance on information extraction benchmarks, expecting RoBERTa-based models to achieve the highest overall scores given their more robust pre-training setup.

5.4 Results

The results reported in Table 4 show consistent and substantial improvements of the multitask configuration over the single-task baseline across all models and Frame Elements, with Overall F_1 gains ranging from ∼8 to ∼10 points. This confirms that treating each Frame Element as a distinct auxiliary task allows the models to better leverage shared linguistic representations, particularly for rare and syntactically complex categories such as Effect, Location, and Taste_Carrier, which the single-task approach largely fails to detect, scoring between 0.00 and 0.416 F_1. RoBERTa-xlm consistently outperforms all other models on the majority of frame elements and achieves the best Overall F_1 (0.543 in the multitask setting), benefiting from its large-scale multilingual pre-training and greater capacity to model extended contextual spans, a critical factor for Frame Elements such as Circumstances and Effect, which can span whole sentences. Across all configurations, we observe significant performance variations across different Frame Elements, with the best results achieved for Quality and Taste_Modifier. This is likely due to the fact that their syntactic realization tends to be consistent across documents, with Quality mainly expressed by adjectives and Taste_Modifier by preposi-

11 https://huggingface.co/google-bert/bert-base-cased
12 https://huggingface.co/google-bert/bert-base-multilingual-cased
13 https://huggingface.co/emanjavacas/MacBERTh
14 https://huggingface.co/FacebookAI/roberta-base
15 https://huggingface.co/FacebookAI/xlm-roberta-base

tional phrases introduced by *with*. On the contrary, classification results for Taste_Source are quite low despite it being the most frequent FE in the training set, probably because it can be expressed by many different role fillers and syntactic constructions. Upon reviewing the test and prediction results, we find that most mistakes concerning Taste_Source are due to a wrong span extent, for instance, the system predicts "the taste of [lollipop]" while the gold standard is "the taste [of lollipop]". As discussed in Section 3, this issue is also reflected in the inter-annotator agreement (IAA) scores of the benchmark. It is worth noting that MacBERTh outperforms RoBERTa-xlm on Taste_Word detection by a full point (F_1 = 0.943 vs. 0.932), a result that is particularly interesting as it suggests that pre-training on historical texts may give the model an advantage in capturing the lexical variation characteristic of historical gustatory language, ultimately benefiting the identification of taste-related lexical units. However, based on the overall results, we select RoBERTa-xlm in the multitask configuration as our model of choice for future research on gustatory language, as it achieves the best performance across the majority of target labels.

6. Data Collection

We run our RoBERTa-xlm model for taste-related terms, configured in the multitask learning setup. The model is applied to a selection of seven different English corpora, selected among the corpora from which we took the documents for the benchmark:

- *Project Gutenberg*[16]
- *Early English Books Online (EEBO)* [17]
- *London Pulse Medical Reports (from the Wellcome Library)*[18]
- *Wikisource*[19]
- *Eighteenth-Century Collections Online (ECCO)*[20]
- *British Library*[21]
- A pre-processed subset of the *Old Bailey Papers* dataset[22]

The corpora were chosen to represent a variety of domains and genres, as well as to provide coverage across different historical periods. Each corpus is first converted into a format compatible with BERT and RoBERTa models, in this case in the multitask configuration (see Fig. 5). The script takes as input a folder containing the raw texts, from which taste-related information is to be extracted, retaining only those text snippets that contain at least one lexical unit (i.e., taste word).

We then automatically extract the taste words and gustatory frame elements using the RoBERTa-xlm model, which achieved the best performance on frame element extrac-

16 https://www.gutenberg.org/
17 https://textcreationpartnership.org/\%20tcp-texts/eebo-tcp-early-english-books-online/
18 https://wellcomelibrary.org/moh/about-the-reports/about-the-medical-officer-of-health-reports/
19 https://en.wikisource.org/wiki/Main_Page
20 https://quod.lib.umich.edu/e/ecco/
21 https://data.bl.uk/digbks/
22 http://fedora.clarin-d.uni-saarland.de/oldbailey/downloads.html

4995	eebo44506	551-65	-	I	O	O	O	O	O	O	O	O	O	
4996	eebo44506	551-66	-	cannot	O	O	O	O	O	O	O	O	O	O
4997	eebo44506	551-67	-	taste	O	O	O	O	O	O	O	O	O	O
4998	eebo44506	551-68	-	such	O	O	O	O	O	O	O	O	O	O
4999	eebo44506	551-69	-	grosse	O	O	O	O	O	O	O	O	O	O
5000	eebo44506	551-70	-	meate	O	O	O	O	O	O	O	O	O	O
5001	eebo44506	551-71	-	;	O	O	O	O	O	O	O	O	O	

Figure 5
Converted Text

Table 6
Extracted data statistics

LUs or FEs	Token Count
Taste_Word	265,361
Taste_Source	189,059
Quality	104,192
Taste_Carrier	6,503
Evoked_Taste	5,270
Location	9,522
Taster	92,301
Taste_Modifier	31,242
Circumstances	10,439
Effect	8,977

tion among the models we tested. Subsequently, we integrate the corpus metadata by adding the publication year of each book and the full sentence from which the various frame elements were extracted. The final extraction statistics are presented in Table 6. As can be seen from the table, the majority of the extracted data is represented by the labels that form the core of the tasting event, both semantically and syntactically, namely taste word, taste source, and quality. For this reason, we focus our exploration on these elements.

7. An exploration of the representation of *tea* in the literary domain

Given the central role of food in shaping cultural identity (Jurafsky 2014), we aim to demonstrate the potential of the system for gustatory information extraction to analyze how food is described and how its perception has evolved over time. This approach is based on the assumption that the way certain foods and drinks are represented in text may reflect their historical and cultural significance. As highlighted in our benchmark analysis, genre appears to play a key role in shaping food discourse. For this reason, we focus our analysis on a single literary-domain corpus, the Project Gutenberg corpus. We first isolate the data sourced from Project Gutenberg, narrowing our scope to texts spanning the 18^{th}, 19^{th}, and 20^{th} centuries. The final statistics for each century are presented in Figure 6. Our interest is in the descriptions of what we ingest, how the object of the taste experience is described and how it changes over time. We thus center our analysis on the elements that in our frame-based schema are labeled as *taste source*. We first calculate the most frequent taste sources in each century, distinguishing between

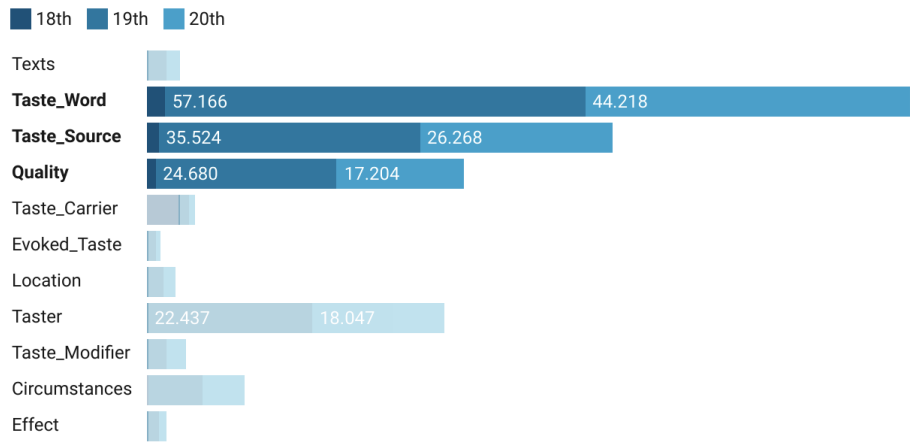

Figure 6
Gutenberg statistics

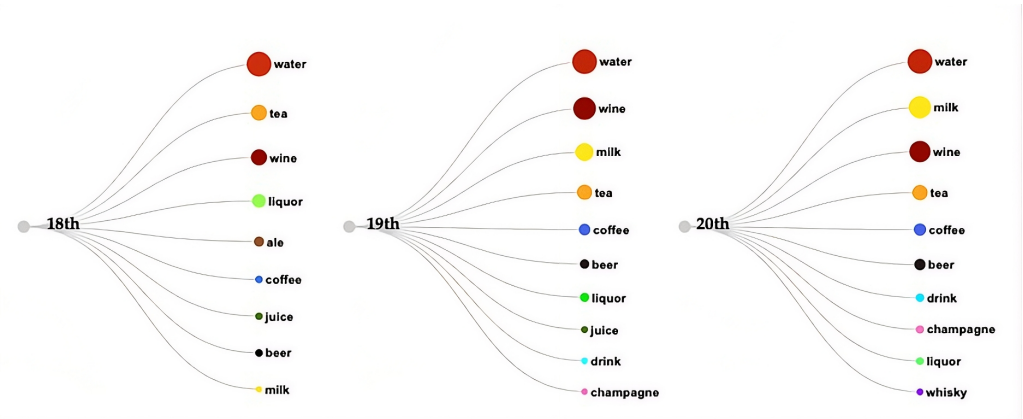

Figure 7
Most frequent drinks across centuries

food and drinks. In both cases, the most commonly described foods and drinks remain largely the same across the centuries, with just slight variations, as we can see in Fig. 7, 8. For example, we observe that in the 18th century there is still a distinction between *beer* and *ale*, which were traditionally considered two different drinks produced with different ingredients (ale flavors came from herbs and spices while beers were hopped brews). However, starting from the 19th century this distinction blurred and the terms were used as synonyms, with *beer* being more frequently used.

To better analyze possible differences across these sources and to demonstrate the potential of a similar approach for quantifying such changes, we select a drink whose perception has been shown to evolve over time, as documented in previous historical studies, i.e. **tea** (Smith 2014; Menini et al. 2023). We first have a look at the distribution of this source in the Project Gutenberg corpus, calculating its temporal frequency, using a 21-year moving average (±10 years). This smoothing technique offers a more accurate representation of a target word's distribution over time than raw frequency counts, as

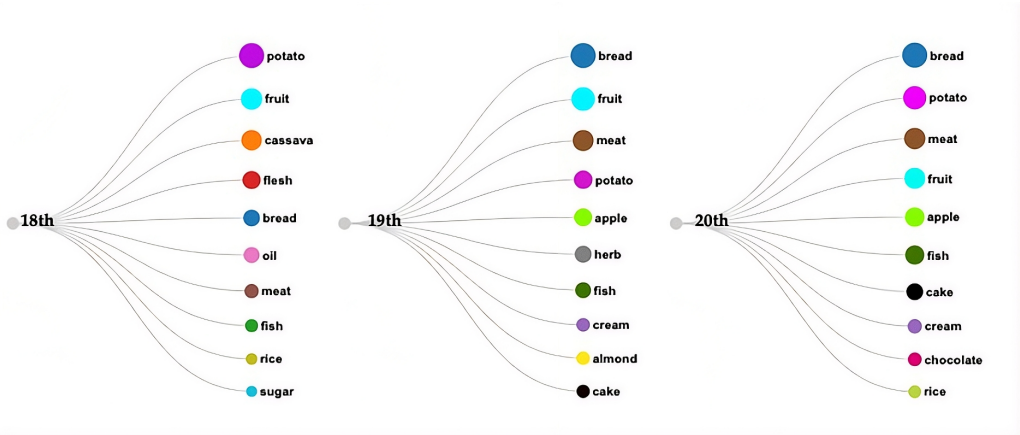

Figure 8
Most frequent food across centuries

Figure 9
Moving average of *tea* in the Gutenberg corpus

it reduces short-term fluctuations and noise that might otherwise obscure meaningful long-term patterns. The representation of this target taste source is presented in Fig. 9. This plot clearly reflects the real-life spread of tea consumption in Europe. From the mid-18th century onward, the increasing frequency of the word *tea* in our corpus mirrors its emergence as a widely consumed product, aligning with historical accounts of its growing popularity during this period (Leung Kin 2007). Although the visualization of tea's distribution in the corpus offers insights into its increasing integration into the everyday lives of English people, it does not allow us to trace potential shifts in how the beverage

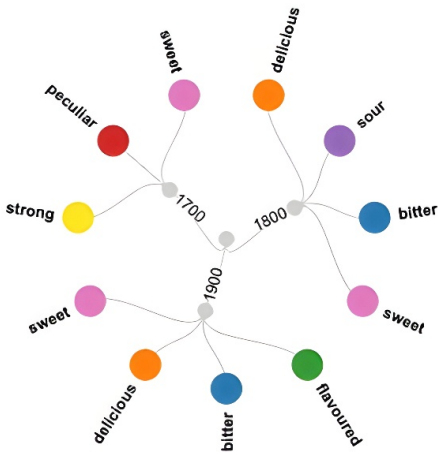

Figure 10
Highest PMI scores for *tea*'s qualities

was perceived over time. To address this limitation, we investigate the qualities associated with tea across different centuries. As noted above, the frame element Quality refers to the taste characteristics of specific substances, typically expressed through attributes that are either qualitative, such as delicious or disgusting, or descriptive, indicating aspects like type or origin. These qualities offer possible insights into shifts in perception as well as evolving patterns of use and consumption over time. To identify the qualities most commonly used to describe tea across different centuries, we employ Pointwise Mutual Information (PMI), a widely recognized measure of association that has proven effective in analyzing shifts in the perception of specific smell objects over time (Paccosi et al. 2023). We calculate the PMI between a given smell source (w_1) and its corresponding qualities (w_2) to measure their strength of association as follows:

$$\text{PMI}(w_1, w_2) = \log 2 \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)}$$

Here, $P(w_1, w_2)$ represents the probability of a smell source and a quality to co-occur, while $P(w_1)$ and $P(w_2)$ indicate their individual probabilities. For each century, we compute the PMI scores between tea and the words labeled as Quality during that period, selecting the top ten qualities with the highest association scores and a frequency of association greater then ten (> 10). The results are illustrated in Fig. 10. Overall, the qualities remain generally stable, with a slight tendency toward more positive adjectives. In the 18th century, terms such as *peculiar* and *strong* mirrored tea's exotic status upon its introduction. While the description of tea's bitterness has remained consistent over time, the adjectives used to describe tea tend increasingly towards positive connotations. Given the smaller amount of data for the 18th century compared to the later periods, we conducted an additional co-occurrence analysis to better explore potential contextual differences in tea descriptions. The co-occurrence of terms was calculated in a five-term window of the sentences from which at least a taste word is extracted (see Fig. 11). We isolated the most frequent adjectives occurring with tea (excluding those

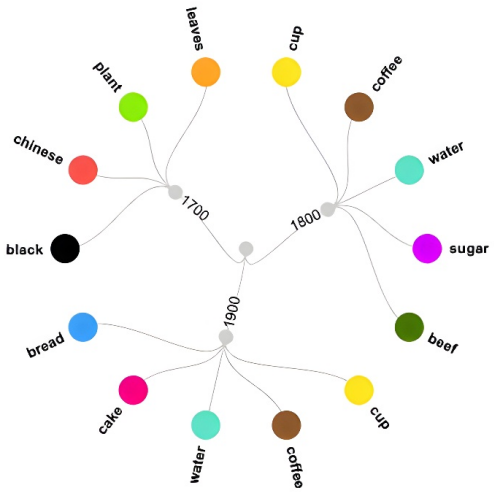

Figure 11
Noun and adjectives co-occurrences for *tea*

already identified by the PMI analysis) as well as the most frequent nouns. In the first century under investigation (1700s), it is interesting to note that among the adjectives most frequently associated with tea are *black* and *Chinese*, which again highlight the exotic and specific nature of this beverage in 18th century England. This is supported by the nouns *plant* and *leaves*, which refer to a form of tea that becomes less mentioned in later centuries within literary sources, namely its identity as a botanical plant whose leaves are infused. Early references to tea often focused on the botanical origin and preparation method, which later shifted towards its consumption as a beverage. In the 1800s and 1900s, tea starts to be seen as a domestic rather than an exotic good and is mostly mentioned in association with words such as *cup*, *water*, *sugar*, and *cake* mirroring its modern consumption. Similarly, *coffee* appears with comparable frequency and in similar social contexts. Indeed, in the 1700s, coffee appears mainly in production-related contexts, reflecting its main status as a commodity. By the 1800s, it became a consumed beverage, associated with taste, and meals, and by the 1900s it was firmly integrated into culinary and social contexts, often alongside snacks and desserts (see Fig. 12)

A particularly interesting word associated with tea is *beef*, which is frequent only in the 1800s. This refers to the historical use of beef tea, which was a common restorative or medicinal beverage in Europe during tea's early introduction (Talea 2025). In this period, beef tea was often recommended medically alongside or as an alternative to tea, which was itself initially used as a medicinal beverage. Beef tea is a clarified beef broth, made by boiling beef (often without bones and with connective tissue) to extract flavor and nutrients, very popular in Europe in the 19th centuries. It was considered an easily digestible food, mainly used as a tonic for convalescents (Griffith 1856). As shown in Fig. 13, *beef* was not generally more frequent in this period; its prominence is restricted

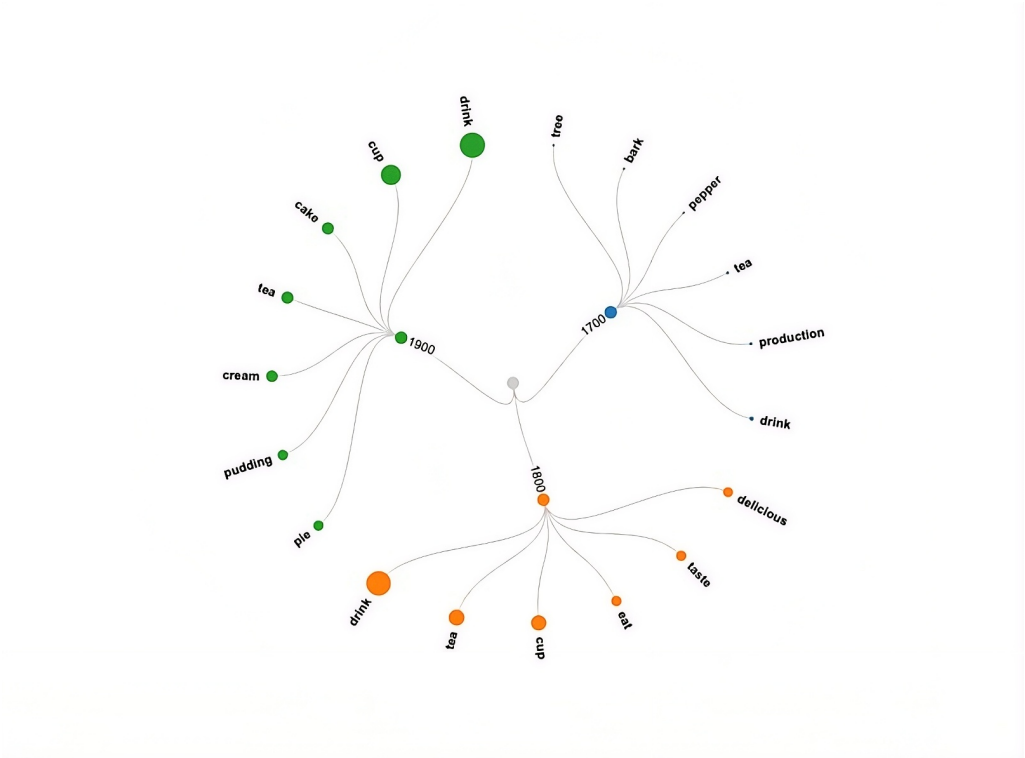

Figure 12
Noun and adjectives co-occurrence for *coffee*

to the context specifically associated with *tea*. In conclusion, we have shown how frame-based analysis can support and highlight patterns that qualitative approaches have traditionally explored on a smaller scale, thus providing quantitative insights into the historical understanding of food-related practices and cultural customs. To further support this kind of exploration, we release a set of tools, data and a web interface (see Sections 8 and 9) to facilitate quantitative approaches even for users without a technical background.

8. Web interface

Given that one of the goals of this system is to facilitate qualitative gustatory research for scholars who may not have a technical background, we adapted the classifier for use via a web interface, available at `https://dh-server.fbk.eu/taste-extractor/`. The interface provides three specialized models for English: a general-purpose model (RoBERTa), a historical model (MacBERTh), and a multilingual model (RoBERTa XLM) that can be also used for zero-shot classification in other languages. To use the interface, either write or paste a text up to 1000 tokens containing references to taste experiences into the 'Insert a text' box. Select then the historical, monolingual or multilingual (zero-shot) model to extract the taste references and download a tsv file in the IOB format, which is in line with previous NER approaches. Being the most accurate overall, the

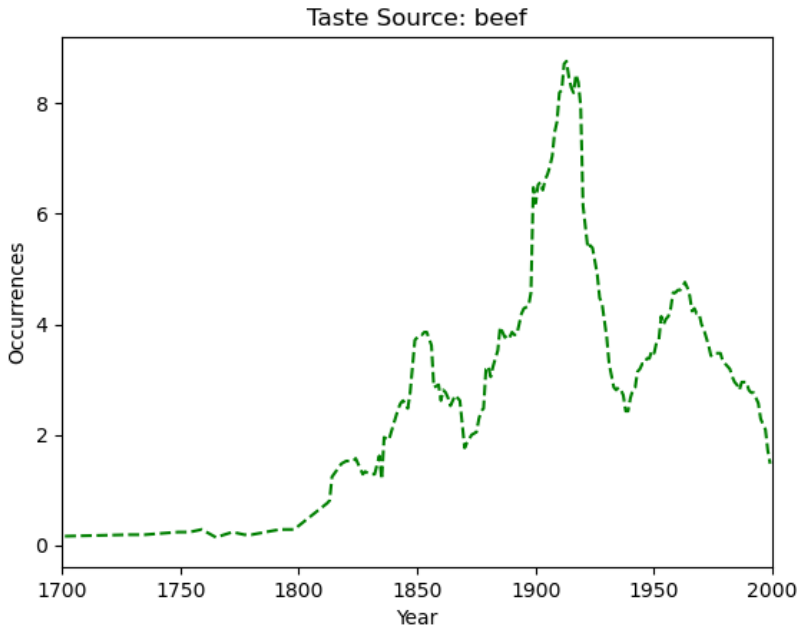

Figure 13
Moving average of the word *beef*.

multilingual model is the recommended choice for the great majority of the cases. A screenshot of the interface is provided in Fig. 14.

9. Code and Data Release

With this paper we release:

- The annotated benchmark, [23]
- The pipeline to extract frame predictions, [24] namely: (1) the book converter for the model format predictions, (2) the models and the code to extract the gustatory frame from texts, and (3) a code to convert the output of the model in a tsv suitable for the analysis of the frame,
- The web interface, [25]
- The notebook for the analysis, including: (1) the computation of moving average for a target Taste Source, (2) the computation of PMI associations for Quality and target Taste Source, (3) the computation of co-occurrences of a target Taste Source, (4) the extraction of Taste Sources co-occurring

23 https://github.com/tpaccosi/Gustatory-Benchmark
24 https://github.com/tpaccosi/Gustatory-Frame-Extraction/
25 https://dh-server.fbk.eu/taste-extractor/

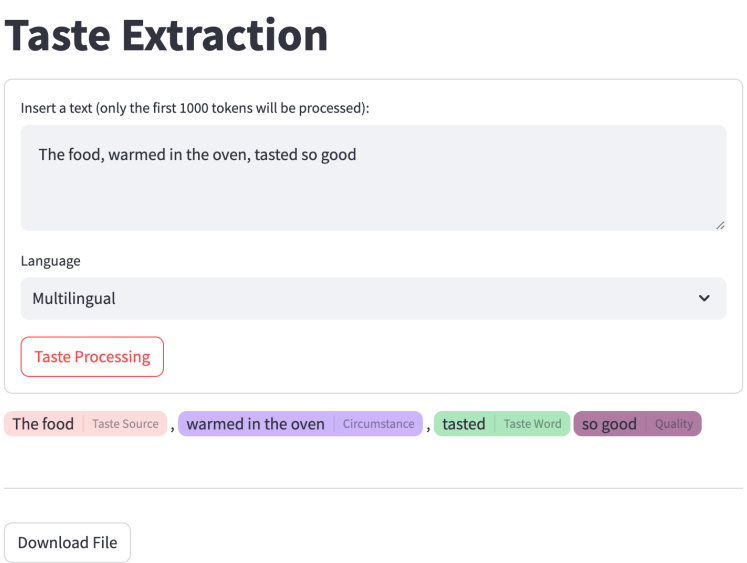

Figure 14
Web interface for gustatory frame extraction

with a target Taste Source, and (5) the taste data automatically extracted from Project Gutenberg. [26]

10. Conclusions and Future Directions

In this paper, we presented a benchmark for gustatory events containing manually annotated taste-related information, built as a counterpart to the one proposed in (Menini et al. 2022). The benchmark is constructed with the same approach adopting a frame-based methodological framework to analyze sensory language. We emphasized the importance of frame-based analysis in capturing and comparing sensory events, by exploring the characterization of positive and negative valence through the analysis of taste and smell words and sources in the benchmark data. The analysis based on frames brings relevant insights into capturing sensory valence from different perspectives, likely supporting the suitability of this approach to deal with humanistic inquiries. We then presented a supervised system to automatically extract taste-related frames, trained on this benchmark, demonstrating that a multitask learning approach is not only effective for this type of extraction but also outperforms single-task models. To further investigate the potential of this frame-based approach to analyze taste from a historical point of view, we conducted a preliminary diachronic analysis of the taste source *tea* in the literary domain. This analysis underscores how a frame-based approach is an effective methodology for tracing and analyzing the evolution of food concepts in texts. Future work should focus on increasing the number of documents in the benchmark and providing less sparse annotations to achieve better temporal balance. The current limited dataset likely contributes to the notable discrepancies in accuracy

26 https://github.com/tpaccosi/Taste-Source-Analysis

across different frame elements, highlighting the need for more annotated instances to support robust generalization. Priority should be given to enriching annotations for frame elements with lower performance and fewer examples in the benchmark, such as Taste_Carrier and Location. Moreover, alternative evaluation metrics and techniques should be explored to better capture and explain performance variations across models. As an additional point of comparison, we also plan to evaluate the performance of general-purpose frame semantic parsers, such as LOME (Xia et al. 2021), on our benchmark.

Limitations

The size of our manually annotated dataset is quite limited, which may affect the generalization of our models. Nonetheless, the results obtained are promising for the application of this model to historical studies of taste. We are also aware that a single case study is not sufficient to fully evaluate the model, but it provides a showcase of its potential and illustrates how the released notebook can be used to adapt the analysis to other historical sources or research questions.

Acknowledgments

Funded by the European Union under grant agreement 101088548 -TRIFECTA. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors would also like to thank Marieke van Erp, the head of the project, for her support.

References

AI, Open. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bagler, Ganesh and Mansi Goel. 2024. Computational gastronomy: capturing culinary creativity by making food computable. *NPJ Systems Biology and Applications*, 10(1):72.

Bagli, Marco. 2021. *Tastes we live by: The linguistic conceptualisation of taste in English*. Walter de Gruyter GmbH & Co KG.

Boscher, Cédric, Christine Largeron, Véronique Eglin, and Elöd Egyed-Zsigmond. 2024. Sense-lm: A synergy between a language model and sensorimotor representations for auditory and olfactory information extraction. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1695–1711, St. Julian's, Malta, March. Association for Computational Linguistics.

Brate, Ryan, Paul Groth, and Marieke van Erp. 2020. Towards olfactory information extraction from text: A case study on detecting smell experiences in novels. In *Proceedings of 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.

Caruana, Rich. 1993. Multitask learning: A knowledge-based source of inductive bias. In *ICML'93: Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48, Amherst, MA, USA, July. Morgan Kaufmann Publishers Inc.

Caruana, Rich. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Cassotti, Pierluigi, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada, July. Association for Computational Linguistics.

Cenikj, Gjorgjina, Gorjan Popovski, Riste Stojanov, Barbara Korousic Seljak, and Tome Eftimov. 2020. Butter: Bidirectional lstm for food named-entity recognition. In *2020 IEEE International*

Conference on Big Data (Big Data), Atlanta, Georgia, December. IEEE Institute of Electrical and Electronic Engineers.
Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, online, July. Association for Computational Linguistics.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
Fillmore, Charles J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.
Giulianelli, Mario, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada, July. Association for Computational Linguistics.
Goel, Mansi and Ganesh Bagler. 2022. Computational gastronomy: A data science approach to food. *Journal of Biosciences*, 47(1):12.
Griffith, Eglcfeld. 1856. *A Universal Formulary Containing the Methods of Preparing and Administering Officinal and other Medicines for Physicians and Pharmaceutists*. Philadelphia: Lea and Blanchard.
Jurafsky, Dan. 2014. *The language of food : a linguist reads the menu*. W.W. Norton & Company, New York.
Krippendorff, Klaus. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania.
Leung Kin, Han Paul. 2007. Tracing the history of tea culture. In *Tea and tourism: Tourists, traditions and transformations*. Clevedon: Channel View.
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
Magnini, Bernardo, Vevake Balaraman, Simone Magnolini, and Marco Guerini. 2018. What's in a food name: Knowledge induction from gazetteers of food main ingredient. In *Proceedings of the Fifth Italian Conference on Computational Linguistics, CLiC-it 2018*, page 241, Turin, Italy, December. CEUR Workshop Proceedings.
Manjavacas Arévalo, Enrique and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, pages 23–36, online, December. Association for Computational Linguistics.
Menini, Stefano. 2024. Semantic frame extraction in multilingual olfactory events. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14622–14627, Turin, Italy, May. Association of Computational Linguistics.
Menini, Stefano, Teresa Paccosi, Serra Sinem Tekiroğlu, and Sara Tonelli. 2023. Scent mining: Extracting olfactory events, smell sources and qualities. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 135–140, Dubrovnik, Croatia, May. Association for Computational Linguistics.
Menini, Stefano, Teresa Paccosi, Sara Tonelli, Marieke Van Erp, Inger Leemans, Pasquale Lisena, Raphael Troncy, William Tullett, Ali Hürriyetoğlu, Ger Dijkstra, et al. 2022. A multilingual benchmark to capture olfactory situations over time. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 1–10, Dublin, Ireland, May. Association for Computational Linguistics.
Paccosi, Teresa, Stefano Menini, Elisa Leonardelli, Ilaria Barzon, and Sara Tonelli. 2023. Scent and sensibility: Perception shifts in the olfactory domain. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 143–152, Singapore, December. Association for Computational Linguistics.

Paccosi, Teresa and Sara Tonelli. 2024a. Benchmarking the semantics of taste: Towards the automatic extraction of gustatory language. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 720–727, Pisa, Italy, December. CEUR Workshop Proceedings.

Paccosi, Teresa and Sara Tonelli. 2024b. A new annotation scheme for the semantics of taste. In *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation@ LREC-COLING 2024*, pages 39–46, Turin, Italy, May. Association for Computational Linguistics.

Popovski, Gorjan, Barbara Koroušić Seljak, and Tome Eftimov. 2019. Foodbase corpus: a new resource of annotated food entities. *Database*, 2019:baz121.

Ruppenhofer, Josef, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute.

Schlechtweg, Dominik, Shafqat Mumtaz Virk, Pauline Sander, Emma Sköldberg, Lukas Theuer Linke, Tuo Zhang, Nina Tahmasebi, Jonas Kuhn, and Sabine Schulte Im Walde. 2024. The durel annotation tool: Human and computational measurement of semantic proximity, sense clusters and semantic change. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 137–149, St. Julians, Malta, March. Association for Computational Linguistics.

Smith, Woodruff D. 2014. From coffeehouse to parlour: the consumption of coffee, tea and sugar in north-western europe in the seventeenth and eighteenth centuries. In *Consuming Habits: Global and Historical Perspectives on How Cultures Define Drugs*. Routledge, pages 142–157.

Stojanov, Riste, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research*, 23(8):28229.

Talea, Anderson. 2025. Beef Tea and Protose Broth: A Study in Convalescent Cookery, 1893–1904. *Global Food History*, 12(1):1–17. Publisher: Routledge.

Tekiroğlu, Serra Sinem, Gözde Özbal, and Carlo Strapparava. 2014. A computational approach to generate a sensorial lexicon. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 114–125, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Van Der Goot, Rob, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, online, April. Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA., December. Curran Associates Inc. 57 Morehouse LaneRed HookNYUnited States.

Winter, Bodo. 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, Bodo. 2019. *Sensory linguistics: Language, perception and metaphor*. John Benjamins Publishing Company.

Wróblewska, Ania, Agnieszka Kaliska, Maciej Pawłowski, Dawid Wiśniewski, Witold Sosnowski, and Agnieszka Ławrynowicz. 2022. Tasteset–recipe dataset and food entities recognition benchmark. *arXiv preprint arXiv:2204.07775*.

Xia, Patrick, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online, April. Association for Computational Linguistics.