Constructing a Recipe Web from Historical Newspapers

Marieke van Erp[1], Melvin Wevers[1], and Hugo Huurdeman[2]

[1] KNAW Humanities Cluster, DHLab, Amsterdam, the Netherlands
{marieke.van.erp,melvin.wevers}@dh.huc.knaw.nl
[2] Universiteit van Amsterdam, Amsterdam, the Netherlands
h.c.huurdeman@uva.nl

Abstract. Historical newspapers provide a lens on customs and habits of the past. For example, recipes published in newspapers highlight what and how we ate and thought about food. The challenge here is that newspaper data is often unstructured and highly varied. Digitised historical newspapers add an additional challenge, namely that of fluctuations in OCR quality. Therefore, it is difficult to locate and extract recipes from them. We present our approach based on distant supervision and automatically extracted lexicons to identify recipes in digitised historical newspapers, to generate recipe tags, and to extract ingredient information. We provide OCR quality indicators and their impact on the extraction process. We enrich the recipes with links to information on the ingredients. Our research shows how natural language processing, machine learning, and semantic web can be combined to construct a rich dataset from heterogeneous newspapers for the historical analysis of food culture.

Keywords: natural language processing, information extraction, food history, digitised newspapers, digital humanities

1 Introduction

There is no dearth of structured recipes available online (cf. Epicurious, Foodnetwork.com).[3] Recipes can also be found in non-structured form in digitized newspapers and magazines. Because of their diachronic nature, these recipes can offer valuable insights into the evolution of food customs, making them of particular interest to historians and ethnologists. However, their lack of structure and varying OCR quality make it more difficult to identify, extract, and use these recipes for analysis. In this paper, we present our work on extracting and enriching recipes from a collection of Dutch historical newspapers (1945-1995).

Scholars in the humanities and social sciences approach what, how, when, where, and why we consume as constitutive signifiers of national and local identities[1]. Diachronic analyses of recipes offer insights into changes in food

[3] http://www.epicurious.com, http://www.foodnetwork.com

culture, shedding light on "analyses of everyday culture, the changing foundations of nations in a globalising world, and of food and drink as subjects of objects of consumption within the dynamic material worlds of late capitalism and late modernity[2]." Van Otterloo argues that the perception of a typical Dutch food culture formed during the 1950s. She also claims that it is difficult to get an overview of all the ideas and perceptions of food and the consumption of food. Computational approaches, however, are able to process large amounts of data and can possibly extract a more comprehensive overview of developments of ideas associated with food.

Newspapers function as transceivers; they are both producer and messenger of public discourse [3,4]. In other words, newspapers both reflect and shape prevailing ideas and tastes in particular periods. Recipes have been part of newspapers at least since the late nineteenth century. In addition, newspapers also contain reports containing views on daily life and customs in a national context. This information regularly appeared in recipes, offering an understanding of food cultures of the past. On the whole, this makes newspapers an invaluable source for studies of food culture.

However, recipes in digitised newspapers are not easily accessible. For instance, a query using the search term 'recept' (recipe) not only retrieves articles containing food recipes, but also recipes for homemade remedies and articles mentioning doctor's prescriptions—the same word in Dutch. Furthermore, not all articles that include recipes include the term 'recipe.' Due to noise introduced in the digitisation process and the diachronic language variation, standard information extraction methods perform poorly on such data. This paper addresses these issues and presents our method and experiments for (1) automatically identifying recipes in newspapers using a classification algorithm, (2) classifying the recipes using a multi-label classifier, (3) extracting ingredients, quantities and units using automatically extracted lexicons, and (4) linking the ingredients to information on their origins. In all steps, we investigate methods for which we can automatically generate training data (via distant supervision) or automatically extracted lexicons from domain-specific and generic resources. This approach also lowers the threshold to transfer the approaches to other domains. Furthermore, we evaluate the quality of the OCR and of our extraction process. All annotations, including the OCR quality indicators, are made available, enabling researchers to gain insights into the quality of the extracted information.

Our contributions are twofold: (1) a distant supervised method for extracting, structuring and enriching recipes from newspapers; (2) a dataset consisting of 27,411 historical recipes extracted from Dutch newspapers (1945-1995), which can be used for further research.

Our software, experiments and data can be found at: `https://github.com/DHLab-nl/historical-recipe-web`. Due to copyright restrictions, the text from the newspaper articles is not included, but can be retrieved via the document IDs.

The remainder of this paper is structured as follows. In Section 2, we discuss the background and related work. In Section 3, we describe the datasets used in this work. Our extraction, structuring and enrichment pipeline and evaluation are described in Section 4. Statistics on our historical recipes dataset are presented in Section 5. We discuss strengths and limitations of our approach in Section 6 and conclude with directions for future work in Section 7.

2 Related Work

The food domain has recently gained some attention in the AI community as a versatile application domain. Various recipe databases are available for research purposes, such as [5] and [6]. These can, for example, be used for the construction of recipe workflows containing specific actions to be carried out [7,8].

Recipe extraction and classification is clearly a multilingual research domain, as [8,9] show by taking Japanese and Italian as their domains, respectively. [10] enrich German recipes with category tags. We apply this type of tag classification in Subsection 4.3, but we amend the feature selection to fit our dataset. Closer to our work is the extraction of ingredients and quantities and units from recipes such as presented in [9] and [11]. However, the main difference with our work is that their corpora are digital-born and thus not affected by fluctuations in text quality from the digitisation process as our corpus is (which also holds for [5,6,7,8]).

We take inspiration from [9] concerning the use of different lexicons for the extraction of ingredients (see Subsection 4.4). The fluctuation in digitisation quality of our corpus affects our options for the application of standard natural language processing tools. There is some work on information extraction from noisy OCR data, such as [12] who investigate the impact of error rates from different OCR engines in a Named Entity Recognition (NER) task using a dictionary, regular expressions, a Maximum Entropy Markov Model, a CRF, and a combination of the approaches. For Dutch ingredient, quantity and unit extraction, there is no training data available as there is for NER. Therefore, we focus on dictionary and regular expression-based methods for that part of our research.

In the Semantic Web domain, the two main dedicated food datasets we found were Open Food Facts[4], an open collaborative database containing information about food in English and French and Foodpedia, a linked dataset containing Russian food products [13]. Although there are dedicated recipe vocabularies such as the BBC Food Ontology,[5] and the Food Ontology[6], the number of datasets using those is limited, not open, or not easily findable.

Some examples of analyses that rich food datasets can provide can be found in [14], which presents an exploratory interface for comparing 487 chocolate chip cookie recipes collected from the Web. Restaurant menus also provide a window into social status as a linguistic analysis of 6,511 restaurant menus by [15] shows. They found that more expensive restaurants use longer and more foreign

[4] `https://world.openfoodfacts.org/data`

[5] `https://www.bbc.co.uk/ontologies/fo`

[6] `http://data.lirmm.fr/ontologies/food`

words. As different newspapers target different audiences, our dataset may also provide such insights, but the core goal of this research paper is to investigate the extent to which distant supervised methods can be used to identify, classify, and structure recipes from a historical newspaper corpus.

3 Data

Using Optical Layout Recognition, pages have been segmented into separate articles, available as images and OCR'ed text. The quality of the digitised text varies throughout the corpus. The age and quality of the original material are important determinants of the ability of the software to recognise the text; hence, older newspapers contain more errors than more recent papers.

The National Library of the Netherlands allows researchers to access data through an API and selected parts of the corpus are available as downloadable data dumps.[7] Access to the source material enables more substantial analyses, which are not possible on resources that are solely accessible through web search interfaces such as the Library of Congress' Chronicling America Corpus.[8]

In addition to our dataset of newspapers, we used a corpus of structured recipes to bootstrap the extraction of ingredients and to train a multi-label classifier to tag the historical recipes. This additional corpus consists of approximately 16,000 recipes from *Allerhande*, the recipe resource from the oldest and one of the largest Dutch supermarket chains.[9] Its recipes have been marked up with schema.org information[10] as well as tags, nutritional information, the source of publication, and ratings.

Data selection

We selected four recently-digitised newspapers because of their higher OCR quality. These newspapers are the liberal *NRC handelsblad* (1970-1994), the social-democratic Amsterdam-based newspaper *Het Parool* (1946-1995), the Catholic *Volkskrant* (1950-1995) and the Protestant newspaper *Trouw* (1950-1995). Table 1 details the descriptive statistics of our dataset.[11]

Apart from their higher OCR quality, the historical period represented by the selected newspapers is of particular interest for research into Dutch food culture. The period after the Second World War exhibited rapid modernisation and industralisation. The recipes might show how these processes affected food

[7] Due to copyright restrictions, a user agreement is required for newspapers published after 1876.

[8] `https://chroniclingamerica.loc.gov/`

[9] `https://www.ah.nl/allerhande/`

[10] `http://schema.org`

[11] Note that the decreased type-token ratio for the *NRC* suggests that the OCR quality in this newspaper is probably the lowest. Of these four newspapers, *NRC* was digitised first, which might explain the lower OCR quality.

Table 1. Statistics of the four selected historical newspapers: number of pages, number of articles, the number of unique tokens (types), the number of tokens in total (tokens) and token to type ratio (TTR)

	Pages	Articles	Types	Tokens	TTR
Parool	14,194	2,380,697	23,651,078	612,036,106	0.039
Volkskrant	13,628	2,248,652	28,616,758	744,275,792	0.038
NRC	7,199	947,198	11,735,250	489,397,816	0.024
Trouw	13,891	2,578,731	24,520,472	656,941,631	0.037

culture and perceptions of cooking within households. The Netherlands also welcomed people from its former colonies Indonesia and Surinam as well as migrant workers from Morocco and Turkey. These migrant communities introduced new recipes and styles of cooking to the Netherlands. We argue that a dataset of historical recipes and their descriptions can be used to better understand how these cuisines were perceived and appropriated in the Netherlands [1,16,17,18].

4 Constructing the Historical Recipe Web

In our workflow, we first generate lists of ingredients, recipe tags, and recipe descriptions from the structured recipe background dataset (Allerhande). We use this dataset to train a recipe tag classifier (described in Subsection 4.3) and to bootstrap an ingredient and quantities and units extractor (described in Subsection 4.4). The first step includes the detection of historical recipes using a seed list and the training of a recipe classifier based on historical recipes (describe in Subsection 4.1). Then, we tag the historical recipes using our tag classifier and we extract the ingredient and quantify information from them. Finally, we enrich the set of structured historical recipes by linking the ingredients to DBpedia, recovering their scientific name, if available, and linking the ingredients to the Global Biodiversity Information Facility to obtain their origin.

4.1 Recipe Identification

From the four newspapers, we selected articles that include the tokens ‘recept’ or ‘recepten’ *and* one of the following tokens: ‘gram, kilogram, pond, keuken, koken, kook, bakken, eetlepel, gerecht, theelepel, snijden’ (recipe, recipes, gram, kilogram, pound, kitchen, cooking, cook, baking, tablespoon, dish, teaspoon, cut). We then manually annotated which of these articles were actually recipes (Table 2). Some recipes are part of a larger article describing an entire menu. In such cases, we treated the article as a single recipe.

Next, we created a training set of the articles annotated as recipes, articles falsely extracted as recipes, and 24,000 articles randomly selected from the four newspapers bar the articles annotated as recipes. This dataset was used to train

Table 2. Results of recipe annotation from seed tokens

	correct	false	total
Volkskrant	1,526	796	2,322
Parool	1,481	971	2,452
Trouw	2,568	926	3,494
NRC	1,913	753	2,666
Total	7,488	3,466	10,954

Table 3. f_1, precision, and recall of the Recipe Classifier

	f_1	precision	recall
articles	0.97	0.96	0.97
recipes	0.95	0.96	0.95

a recipe classifier. After removing the search terms used for the initial query to improve the performance of the classifier, we transformed the text into a TF-IDF feature space based on unigrams and bigrams. On this feature space, we trained three classifiers: a multinomial Naive Bayes, a Support Vector Machine (SVM) with Stochastic Gradient Descent (SGD), and a Linear Support Vector Classification using cross-validated randomized search on hyperparameters. The latter scored the best with an accuracy score of 0.96 using a 5-fold cross validation (see Figure 3 for precision, recall, and f_1 scores).

After applying the trained classifier to the four sets of newspaper articles, the number of recipes found increased drastically, especially for earlier periods, yielding 27,411 articles of which we have a high confidence that they are recipes. Using the classifier resulted in an almost six-fold increase over the initial seed list.

4.2 OCR Quality of the Recipe Dataset

While the Delpher newspaper data was digitised and OCR'ed relatively recently, the OCR quality is not perfect. To get an indication of the OCR quality, we performed a lexicon-based OCR quality check developed at the Dutch Language Institute.[12] This method checks what proportion of tokens present in an article occurs in a range of historical lexicons.[13] Most OCR software will give an indication of the certainty of its decisions by attaching a score to a document or batch of documents. However, these scores often give an indication of the errors

[12] `https://ivdnt.org/the-dutch-language-institute`

[13] `https://www.digitisation.eu/tools-resources/language-resources/historical-and-named-entities-lexica-of-dutch/`

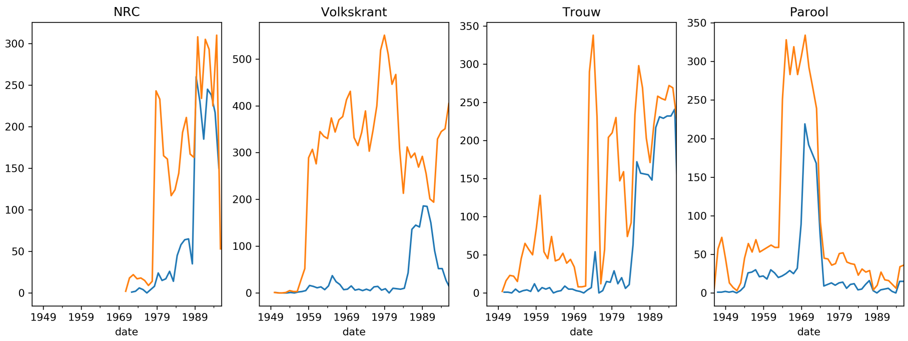


Fig. 1. Retrieved articles using seed list (blue) and using classifier (orange)

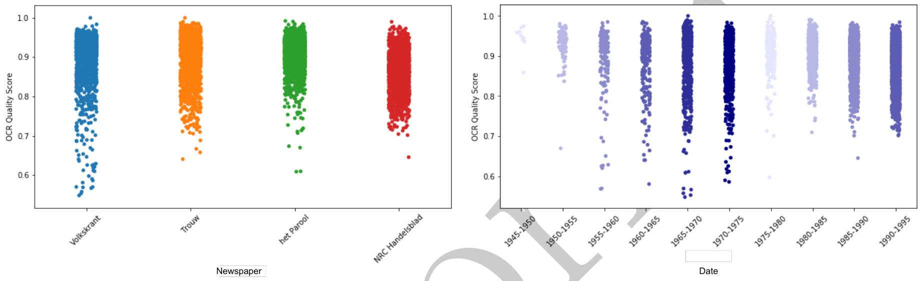


Fig. 2. Lexicon-based OCR quality indicators per newspaper (left) and per five-year period (right)

at the character level, while for our purpose, it is more useful to know how many words (or tokens) are correct in a text, as information extraction techniques do not read as easily over character errors than humans do.

Figure 2 shows the results of this measure on the different newspapers (left) and per 5-year interval (right). Fortunately, the majority of the texts scores about 80%, although there is some difference between the newspapers and the different time periods. The scores are also provided in the historical recipe dataset, such that researchers can choose to exclude articles with a lower OCR score.

4.3 Tag Classifier

To categorise the recipes, we trained a multi-label classifier using the tags associated with recipes in the Allerhande dataset. Recipes in the Allerhande dataset are tagged with one, two, or three tags drawn from a set of 69 tags. These tags either indicate the type of dish (Thai, American, Italian), type of diet (Vegetarian, Healthy, Lactose-free), occasion (Christmas, Easter), or style of cooking (Grilling, Baking, Oven, Fast, Budget).

After initial training of the classifier on all tags, we removed tags with an accuracy score < 0.1, tags occurring in fewer than fifty recipes, and those that

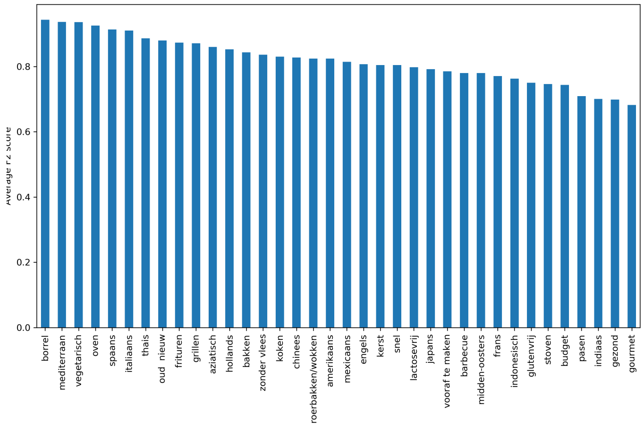


Fig. 3. Accuracy scores per tag of tag classifier on Allerhande dataset

were specific to the Allerhande set such as ‘advertorial’ and ‘wat eten we vandaag’ (what’s for dinner today). Also, we grouped together similar tags, such as ‘healthy’ and ‘slim’, and ‘without meat/fish’ and ‘vegetarian’. These steps resulted in a set of 36 tags.

As input variables, we used the title, description, and cooking instruction fields from the Allerhande set. From this text we removed the names of tags to make the classifier less sensitive to the presence of these words. After converting the text into a TF-IDF feature space with an ngram range of (1, 5), we trained a OneVsRest Classifier balanced Linear SVC. The overall accuracy score of the classifier is 0.75. The Hamming Loss is 0.014, and the average F2 test score: 0.82. Figure 3 shows the scores per tag based on the Allerhande training set.

Subsequently, we applied the tag classifier to the annotated recipes extracted from the historical newspapers. Figure 4 shows the number of tags found in this dataset. In the bar chart, we find that a small set of tags were found relatively often, while others were infrequently found, or not at all. This suggests that some tags are quite specific to the Allerhande data and do not generalize quite well. On the other hand, tags such as ‘vegetarian’, ‘italian’, ‘asian’, and the more specific ‘thai’, ‘grilling’, and ‘deep frying’ were found with high accuracy in historical recipes.

For evaluation, we constructed a dataset of 100 recipes for every tag and 100 recipes that were not tagged. If tags appeared in fewer than 100 recipes, we selected all these recipes, for the other cases we took a sample. The tagged set included 1,197 recipes. We manually annotated recipes with the tags: ‘italian’, ‘vegetarian’, and ‘asian’. These tags occurred relatively often and were easier to tag since they were less ambiguous than for instance, ‘budget’. For these tags, the tagger scored relatively well (Table 4). During manual tagging, we also noticed that recipes tagged as ‘asian’ did not receive the more specific tags ‘japanese’,

Table 4. Evaluation of Tagger on Historical Recipes

	precision	recall	accuracy	f_1
Asian	0.97	0.72	0.95	0.83
Italian	0.83	0.84	0.96	0.84
Vegetarian	0.78	0.45	0.78	0.57

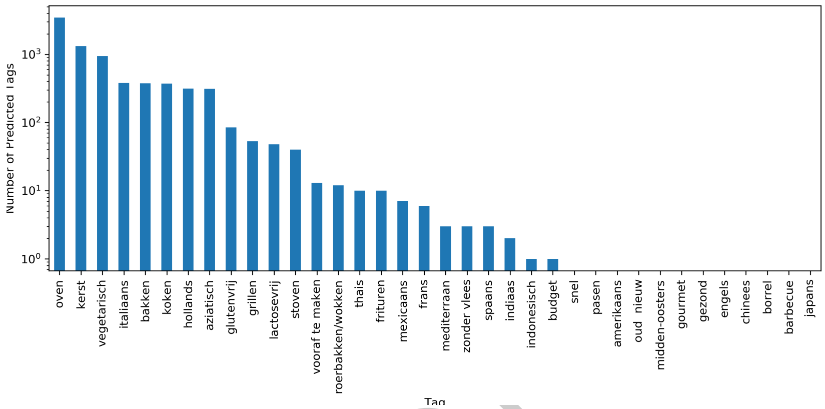
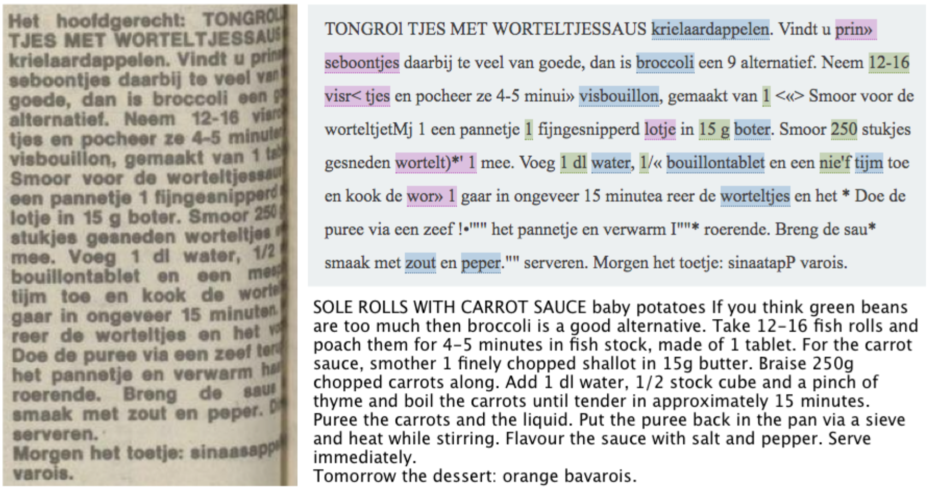


Fig. 4. Frequency of tags found in historical recipes

'indonesian', 'chinese', or 'thai', even though they were described as such. The low recall for 'vegetarian' partly stems from the fact that in the Allerhande desserts, while often vegetarian, are almost never tagged as such. We annotated these recipes as 'vegetarian'. An interesting find was also that a recipe described as 'vegetarian' in a newspaper article was not tagged as 'vegetarian' by our tagger. Here the classifier was actually correct, since the recipe used chicken and trasi, a spice paste made of fermented shrimp. This perhaps suggests a changing concept of vegetarian food.

4.4 Ingredient and Quantity Extraction

Figure 5 illustrates some of the difficulties in extracting information from a digitised newspaper source. As the scan of the newspaper page shows (left), some of the text on the right-hand side is difficult to read because of the fold of the newspaper, resulting in gaps or misrecognised characters in the OCR output (top right). We have annotated the ingredients that do not contain any errors in blue, potential ingredients contained in strings with OCR errors in pink, and quantities in green. Interestingly, not all ingredients are precisely quantified, such as 'a pinch' (literally 'a knife's point' in Dutch). This makes it difficult to, for example, compute the nutritional value of the dish, even if the OCR was perfect and all ingredients and quantities could be recognised correctly.



Het hoofdgerecht: TONGROL
TJES MET WORTELTJESSAUS
krielaardappelen. Vindt u prin
seboontjes daarbij te veel van
goede, dan is broccoli een g
alternatief. Neem 12-16 vis
tjes en pocheer ze 4-5 minute
visbouillon, gemaakt van 1 ta
Smoor voor de worteltjessau
een pannetje 1 fijngesnipperd
lotje in 15 g boter. Smoor 250
stukjes gesneden worteltjes
mee. Voeg 1 dl water, 1/2
bouillontablet en een mes
tijm toe en kook de worte
gaar in ongeveer 15 minuten.
reer de worteltjes en het v
Doe de puree via een zeef ter
het pannetje en verwarm he
roerende. Breng de saus
smaak met zout en peper. D
serveren.
Morgen het toetje: sinaasapp
varois.

TONGROl TJES MET WORTELTJESSAUS krielaardappelen. Vindt u prin» seboontjes daarbij te veel van goede, dan is broccoli een 9 alternatief. Neem 12-16 visr< tjes en pocheer ze 4-5 minui» visbouillon, gemaakt van 1 <«> Smoor voor de worteltjetMj 1 een pannetje 1 fijngesnipperd lotje in 15 g boter. Smoor 250 stukjes gesneden wortelt)*' 1 mee. Voeg 1 dl water, 1/« bouillontablet en een nie'f tijm toe en kook de wor» 1 gaar in ongeveer 15 minutea reer de worteltjes en het * Doe de puree via een zeef !•''''' het pannetje en verwarm I''''* roerende. Breng de sau* smaak met zout en peper.'''' serveren. Morgen het toetje: sinaatapP varois.

SOLE ROLLS WITH CARROT SAUCE baby potatoes If you think green beans are too much then broccoli is a good alternative. Take 12–16 fish rolls and poach them for 4–5 minutes in fish stock, made of 1 tablet. For the carrot sauce, smother 1 finely chopped shallot in 15g butter. Braise 250g chopped carrots along. Add 1 dl water, 1/2 stock cube and a pinch of thyme and boil the carrots until tender in approximately 15 minutes. Puree the carrots and the liquid. Put the puree back in the pan via a sieve and heat while stirring. Flavour the sauce with salt and pepper. Serve immediately.
Tomorrow the dessert: orange bavarois.

Fig. 5. Example of a newspaper recipe scan, its resulting OCR'ed text, marked up with ingredients that our approach should be able to recognise (blue), potential ingredients (pink) and quantities (green) as well as the recipe's English translation. Source: NRC Handelsblad 24 April 1988, page 20, `https://resolver.kb.nl/resolve?urn=KBNRC01:000029338:mpeg21:a0179`

To evaluate the ingredient and quantity extraction, we selected a random stratified sample from the recipe set created using seed list in Subsection 4.1. The sample consists of 100 articles (1.35% of the set) following the same distribution over newspapers and time periods.

Ingredients, quantities and units in the sampled recipes were annotated using the Recogito annotation tool.[14] Furthermore, ingredients that contained OCR-errors were marked separately to gain insight into the proportion of ingredients affected by these errors. Three annotators contributed to the gold standard. Six articles were annotated by all three annotators for which we computed Krippendorff's alpha to measure inter-annotator agreement, yielding a score of 0.85 [19]. Overall, the agreement is high, but we do see disagreement on whether or not parentheses are included and for the OCR category particularly it is unclear when a garbled-up word starts and ends. For example, in one instance Annotator 1 annotated *j ◦ ar,anen'* and Annotator 2: *◦ ar,anen'*.[15]

Ingredient extraction Many of the recipes do not follow a structured format where the ingredients are presented at the start of the article (as web-based recipes or formal cookbooks usually do). Segmenting the articles into 'ingredient'

[14] `http://recogito.pelagios.org/`

[15] The article actually stated '4 bananen'

and 'description' paragraphs is therefore not an option. Experiments with standard NLP tools to identify noun chunks and part-of-speech tags are not robust against the OCR variation in our corpus. Therefore, ingredients are extracted using a dictionary-based tagger. We generate several ingredients lists inspired by [9]. In that work, a domain specific resource was used to bootstrap ingredients from AGROVOC[16] and a combination of three generic resources based on WordNet [20]. As a Dutch version of AGROVOC does not exist, we used the Allerhande corpus to generate a list of unique ingredients consisting of 2,723 food stuffs ranging from 'uien' (onions) to 'Ben Jerry's Cinnamon Buns ijs' (Ben Jerry's Cinnamon Buns ice cream).

We compared the Allerhande list to lists of ingredients from two generic datasets: Dutch DBpedia and Open Dutch Wordnet. From DBpedia, we select resources in the categories 'food' and 'lists of food'.[17]. After excluding some categories (e.g. "List of Belgian Beers", which contained a fair few mentions of breweries), 2,642 potential ingredients remained. Singular nouns were automatically expanded with plural forms using the pattern library.[18] This yielded a total of 4,110 ingredients. From Open Dutch WordNet, we selected lexemes with the superclass 'Food' or 'Plant', yielding 1,602 entries, which were also automatically expanded to their plural forms, added up to 3,204 ingredients.

Table 5 presents the results of four types of ingredients extraction: (1) exact match using the entire list of ingredients; (2) exact match using only ingredients harvested from DBpedia; (3) exact match using ingredients harvested from WordNet; and (4) exact match using the combined lists (AH-DBP-WN). In an effort to tackle spelling variations and OCR errors, we experimented with fuzzy matching, but this only decreased performance by introducing more noise and no gains in recall.

Some ingredients may be mentioned several times in the recipe but we only note each ingredient once, thus performing a type analysis rather than a token analysis. Our gold standard contains 1,538 ingredients without OCR errors and the annotators identified 150 strings denoting ingredients containing OCR errors.

Error analysis The low recall stems from insufficient coverage of the ingredient lists, but simply adding ingredients would not yield 100% recall as there is also variation in parts of ingredients, e.g. 'brandneteltopjes' (tips of nettles) or 'kabeljauwkoppen' (cod heads). Furthermore, recipes occassionally mention ingredients by referring to a brand name, e.g. 'Delfiatablet' (a brand of butter), or by describing a foreign foodstuff, e.g. 'warka-vellen' (Moroccan phyllo).

Errors in precision stem from noise in the lexicons. For example, the Allerhande ingredients list contains 'aardappelsalade' (potato salad) and 'chocoladecake' (chocolate cake), whereas in newspaper article this is the name of the final

[16] `http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus`

[17] `http://nl.dbpedia.org/resource/Categorie:Voedsel;http://nl.dbpedia.org/resource/Categorie:Lijsten_van_voedsel`. The resources typed with `dbo:Food` are mostly beers.

[18] `https://www.clips.uantwerpen.be/pages/pattern-nl`

Table 5. Results of ingredients extraction from recipes. 'Clean ingredients' denotes results on ingredients without OCR errors, 'With OCR errors' denotes results including OCR errors. The number of correct items is the same for both sets as no new mentions from the set of OCR errors was retrieved.

	Clean Ingredients				With OCR errors		
	precision	recall	f_1	correct	precision	recall	f_1
Allerhande	0.70	0.65	0.67	998	0.70	0.59	0.64
DBpedia	0.60	0.33	0.47	513	0.60	0.30	0.45
WordNet	0.62	0.50	0.56	764	0.62	0.45	0.54
AH-DBP-WN	0.56	0.75	0.66	1,154	0.56	0.68	0.62

product. The annotators were instructed to only annotate the base ingredient and not its shape. For example, in 'kokend water' (boiling water) only 'water' is annotated. This decision was made to keep close to unprocessed ingredients and not have to account for variant such as chopped, diced, sliced, grated, etcetera. These variants, however, do occur in the Allerhande ingredients list. In addition to names of dishes, the DBpedia and WordNet lists contain cooking actions such as 'fruiten' (sautée) and other related terms such as 'dier' (animal), 'blikje' (can), and 'ingrediënten' (ingredients). This notwithstanding, our setup is to test the extent to which automatically harvested lexicons can be used for ingredient extraction. Some cleanup would improve the precision, but for the recall an automatically bootstrapped lexicon, or a statistical method will probably yield better results.

Quantity and Unit Extraction Quantities and units are extracted using a regular expression that utilises a list of 91 units generated from the Allerhande dataset containing terms such as 'kilogram' and 'liter', but also 'pakje' (package) and 'pot' (jar). The units were pluralised automatically yielding 182 instances. The matcher checks for an occurrence of one or more digits followed by a unit or a digit followed by an ingredient. This quantity extraction method correctly identified 312 units with a precision of 0.74, a recall of 0.51, and an F_1 of 0.62.

Error analysis Precision errors are often caused by half matches, e.g. recognition of '4 pot' (4 jar) where the full annotation states '4 potten'. Part-of-speech tagging might resolve some of these problems, if the available taggers can be made more robust in dealing with OCR errors. The case for recall is more complex. On the one hand, the units lexicon can be expanded with variations on, for instance, pieces, wine glasses, layers, and tea cups. However, we also found some quite poetic variations on quantities and units expressions, such as 'een paar royale slagen met de pepermolen' (a couple of generous twists on the pepper grinder), 'een niet kinderachtige hoeveelheid' (a not childish amount), and 'een snuf snuf' (a sniff sniff). The use of these variants might be distinctive of particular historical periods.

Table 6. Results of ingredient to DBpedia linking

	precision	recall	f_1	unique	scientific	dbpedia-en
String match	95.56 (280)	10.77 (293)	53.17	293	37	293
Spotlight	85.45 (1,034)	44.47 (1,210)	64.96	438	76	397

4.5 Linking Recipe Elements

The food on our plates is often sourced from all corners of the world. To gain an insight into the different localities from which our ingredients originated, we linked items in our Allerhande ingredients list to the Global Biodiversity Information Facility (GBIF).[19] This resource gives information about different species and their native range. To establish these links, we first collected an ingredient's scientific name from DBpedia, which was then queried in GBIF to obtain its origin. In this step, we also created links between our ingredient list and DBpedia, through which we also obtained links to the English DBpedia. We use two different approaches to generate these links: a simple string match and DBpedia spotlight[21]. The resulting links (Table 6) were judged by one annotator.

Error analysis The precision on the string match is naturally quite high, only in cases where ingredient names are ambiguous this fails. DBpedia Spotlight has more trouble, as it has a higher coverage. It, for example, links 'salsa' to salsa dancing instead of the sauce. Its increase in recall over the string match method is thanks to its access to synonyms such as "Zwaardherik" for 'Rucola' (arugula). There are still quite some ingredients for which no link was found. Some are quite surprising, such as the lack of a link for 'aardbeien' (strawberries), but for ingredients such as 'Amelander verse sladressing' (Amelander fresh salad dressing) or 'kippenbouillontablet' (chicken stock cube) this is not surprising. For many of the processed food items, such as cheese, there is no scientific name and corresponding GBIF entry. There are other interesting sources to relate these to, such as consumer price indices, but we leave this for future work.

5 Dutch Historical Recipe Web

Our extracted and enriched historical recipes dataset of over 27k recipes and over 365k ingredients can for example be used to investigate ingredient combinations in different time periods or popular tags in different newspapers. Table 7 shows the statistics of our recipes dataset.

It should be noted that the newspaper dataset does not include all published newspapers, so any comparative or proportional analyses derived from the newspaper corpus or our dataset will have to take this into account. Recipes may be

[19] https://www.gbif.org/

Table 7. Statistics of Dutch Historical Recipe Web

	Recipes	Tags	Ingredients	Quantities	DBpedia	Scientific	GBIF
Parool	4,440	5,221	46,685	11,620	2,423	277	170
Volkskrant	13,270	16,962	185,872	56,626	7,395	730	349
NRC	3,764	4,943	59,717	17,738	1,850	282	142
Trouw	5,937	7,353	72,859	21,880	3,232	368	168
Total	27,411	34,479	365,133	107,864	14,900	1,657	829

repeated, but differences in OCR performance makes detecting the same recipe not trivial.

6 Discussion

In this paper, we focused on distant supervision approaches to detect and classify recipes from newspapers; to extract ingredients, quantities and units; and to add links to external datasets. The obtained scores show that for the identification and classification tasks, this works quite well, as the recipes from our seed datasets generalise well over to the newspapers dataset.

For the more fine-grained extraction and enrichment, i.e. the ingredients, quantities, units and external links, there are clear limitations to using available lexicons and resources. Although ingredients are not as infinite a set as, for example, named entities, our newspaper dataset shows enough variation to affect the performance of the approach. As the OCR quality affects standard natural language processing tools, such as part-of-speech tagging or noun chunking, it is difficult to bootstrap patterns from the dataset to grow the lexicons. Solutions can be sought in (a) only working with those articles that obtain a high OCR score, (b) cleaning up the OCR, or (c) training NLP systems to deal with noisy text. In our dataset, the OCR lexical coverage scores are provided, so researchers can choose to only use those articles in their analyses. Correcting the OCR is difficult, in particular with images that are already difficult to read for humans, but some tools are becoming available such as PICCL.[20]

7 Conclusions and Future Work

We presented a distant supervised method and experiments to construct a recipe dataset from historical newspapers. To the best of our knowledge, we are the first to combine natural language processing, machine learning, and semantic web for information extraction from noisy OCR data. Our evaluations show that articles denoting recipes can be identified with an F_1 score of 0.96, tags can be assigned

[20] https://github.com/LanguageMachines/PICCL

with F_2 scores between 0.57 and 0.84, ingredients can be identified with an F_1 of 0.67, quantities and units with an F_1 of 0.62 and link with an F_1 of 0.64. These results leave room for improvement, but the approach does not require manually labeled training data. The resulting 27,411 recipes can be used by (humanities) researchers interested in food culture to more easily access relevant sources. We will continue to expand this dataset with additional newspapers and time periods and explore diachronic lexicons and machine learning methods to improve the classification and extraction.

The lexicon-based method that was used for the ingredients and quantities and units extraction is limited by the scope of available lexicons and cleanliness. The Allerhande lexicon, which was derived from schema.org ingredient elements, shows that such markup allows flexibility on the content provider's side, but makes it difficult to repurpose, for example, to use as an ingredients lexicon. Furthermore, the coverage of the Dutch DBpedia in the food domain was also lower and less well-structured than expected.

We have also assessed the impact of OCR errors in the newspapers corpus by providing an indication of the article's lexical coverage and by annotating OCR problems in the ingredients lists in our evaluation dataset. The use of PICCL and other methods will be investigated to improve the quality of the sources.

As our method relies on distant supervision and automatically extracted lexicons, it can easily be ported to other domains to construct similar datasets from (historical) newspapers or magazines such as sport reports or music reviews.

The dataset, software and experiments described in this paper can be found at: `https://github.com/DHLab-nl/historical-recipe-web`

Acknowledgements

The authors thank the National Library of the Netherlands for making available the newspaper collection for research purposes as well as for organising the HackaLOD hackathon with Rijksmuseum and Netwerk Digitaal Erfgoed where this project got started. We thank Jesse de Does for the OCR quality measure, Marten Postma and Emiel van Miltenburg for querying Open Dutch WordNet, and Richard Zijdeman for fruitful discussions on the dataset concept. No Hawaiian pizzas were consumed during the writing of this paper.

References

1. van Otterloo, A.H.: Eten en eetlust in Nederland, 1840-1990: een historisch-sociologische studie. B. Bakker, Amsterdam (1990)
2. Wilson, T.M., ed.: Food, drink and identity in Europe. European studies. Rodopi, Amsterdam (2006)
3. Schudson, M.: The Power of News. Harvard University Press, Cambridge (1982)
4. Marchand, R.: Advertising the American Dream: Making Way for Modernity, 1920-1940. University of California Press, Berkeley (1985)

5. Harashima, J., Ariga, M., Murata, K., Ioki, M.: A large-scale recipe and meal data collection as infrastructure for food research. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (may 2016)
6. Tasse, D., Smith, N.A.: Sour cream: Toward semantic processing of recipes. Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA (2008)
7. Maeta, H., Sasada, T., Mori, S.: A framework for recipe text interpretation. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM (2014) 553–558
8. Mori, S., Maeta, H., Yamakata, Y., Sasada, T.: Flow graph corpus from recipe texts. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (may 2014)
9. Mazzei, A.: On the lexical coverage of some resources on italian cooking recipes. In: Proc. of CLiC-it 2014, First Italian Conference on Computational Linguistics. (2014) 254–259
10. Kicherer, H., Dittrich, M., Grebe, L., Scheible, C., Klinger, R.: What you use, not what you do: Automatic classification of recipes. In: International Conference on Applications of Natural Language to Information Systems, Springer (2017) 197–209
11. Greene, E.: Extracting structured data from recipes using conditional random fields. The New York Times Open Blog (2015)
12. Packer, T.L., Lutes, J.F., Stewart, A.P., Embley, D.W., Ringger, E.K., Seppi, K.D., Jensen, L.S.: Extracting person names from diverse and noisy ocr text. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data, ACM (2010) 19–26
13. Kolchin, M., Chistyakov, A., Lapaev, M., Khaydarova, R.: Foodpedia: Russian food products as a linked data dataset. In: International Semantic Web Conference, Springer (2015) 87–90
14. Chang, M., Hare, V.M., Kim, J., Agrawala, M.: Recipescape: Mining and analyzing diverse processes in cooking recipes. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM (2017) 1524–1531
15. Jurafsky, D., Chahuneau, V., Routledge, B., Smith, N.: Linguistic markers of status in food culture: Bordieu's distinction in a menu corpus. Journal of Cultural Analytics (2016)
16. Schuyt, K., Taverne, E.: Dutch Culture in a European Perspective: 1950, Prosperity and Welfare. Palgrave Macmillan, Basingstoke (2004)
17. Hoving, I., Dibbits, H., Schrover, M., eds.: Cultuur en migratie in Nederland. Veranderingen in het Alledaagse, 1950-2000. Sdu Uitgevers, The Hague (2005)
18. Schot, J., Rip, A., Lintsen, H., eds.: Technology and the Making of the Netherlands: The Age of Contested Modernization, 1890-1970. MIT Press, Cambridge (2010)
19. Krippendorff, K.: Computing krippendorff's alpha-reliability. (2011)
20. Postma, M., van Miltenburg, E., Segers, R., Schoen, A., Vossen, P.: Open Dutch WordNet. In: Proceedings of the Eight Global Wordnet Conference, Bucharest, Romania (2016)
21. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics). (2013)