The discourse of the French method: making old knowledge on market gardening accessible to machines and humans.

David Colliaux[1,*], Remi van Trijp[1]

[1]Sony Computer Science Laboratories - Paris, 6 Rue Amyot, 75005 Paris, France

Abstract

A vast amount of our cultural heritage is at risk of getting lost because it resides in old books that are difficult to access. It is therefore important to make this information available to human readers but also to machine analysis, so that new representations and insights based on this knowledge can be constructed. In our case study, we use a host of digital tools to extract and analyze a corpus of 19th century French texts about the practices of market gardening in Paris, and to apply a variety of possible visualizations in an integrated interface. Our work includes a Named Entity and Linking procedure for creating maps of the locations mentioned in these texts as well as the social networks of people cited in the books. We also consider how the analysis of verbs can approximate and represent the know-how of market gardening: we analyze the statistics of those verbs compared to their usage in a general corpus for French, and map the verbs using word embeddings. Finally, we also consider a semantic frame analysis to extract causal relations from texts to evaluate how well these relations support the biological knowledge embedded in those texts (such as how too much exposure to the sun may affect the quality of the garden's produce). Altogether, we show how the visualizations based on Natural Language Processing and Textual Statistics could support a convivial navigation through the corpus.

Keywords

digital humanities, grounded language, corpus linguistics

1. Introduction

Digital libraries gather large corpora of texts which are beyond human possibilities of reading. One of the tasks of digital humanities [21] is thus to organize and analyze those texts so that they are easy to navigate. For instance, through distant reading [16], we may construct curves, graphs and maps that make this large quantity of information graspable for the human mind. Moreover, it is necessary that the information is accessible not only to humans but also to machines, so that further processing may be applied to those texts.

A large collection of works dedicated their efforts in this direction, applied to literary texts [16] and the press [6], showing the potential of text mining and natural language processing for such corpora. However, less attention has been paid to manuals, even though such texts

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark

*Corresponding author.

david.colliaux@sony.com (D. Colliaux); remi.vantrijp@sony.com (R. van Trijp)

https://csl.sony.fr/member/david-colliaux-phd/ (D. Colliaux); https://csl.sony.fr/member/remi-van-trijp-phd/ (R. van Trijp)

0000-0003-1898-4864 (D. Colliaux); 0000-0002-0475-8367 (R. van Trijp)

are essential as they encapsulate the knowledge of a particular era about a certain topic. In our case, we focused on 19th century manuals about market gardening. Those manuals are both a record of the practices of the time and the beginning of the crystallization of this knowledge into a science, namely agronomy.

19th century texts are particularly interesting because shortly after that period, from the second part of the 20th centuray onwards, agriculture went through radical changes with the green revolution and the introduction of chemicals to control the growth and the environment of plants. These changes, which were driven by the agronomical institutions, were so sweeping that we can reasonably ask whether some part of the old knowledge was lost. To answer this question, it is necessary to mine the older texts; and their analysis will also help visualize some interesting aspects of the history of agriculture.

We present here how we built the corpus, the preprocessing of the data and some analysis we did on the texts. First, we performed Named Entity Recognition and Linking to gather information on the places and people cited in those books. Then, we analyzed the verbs appearing in the corpus through semantic embeddings. And finally, we collected sentences expressing causal relations as those are most susceptible of containing agronomical knowledge. For each of these analyses, we provide visualization which can help navigate the corpus in an interactive manner.

2. The good Old Manuals corpus

The gardening manuals of the 19th century are a memory of the development of very efficient methods for growing vegetables in an urban environment (as many of these books are focused on the practices in the Paris area). These methods of cultivating very densely mixtures of crops on small plots of land have inspired a movement in California and more recently in Europe commonly referred to as the Biointensive French Method [14], or French Method [4] for short. The French method is related to more recent practices like agroecology [23] or permaculture [7], although the French Method insists on how to force the culture of vegetables out of season to be able to sell products at higher price early in the season or late in the season. One book in particular, *Manuel pratique de la culture maraîchère de Paris* by Moreau and Daverne, was particularly influential according to the actors of this revival [12], but there is a rich collection of literature on the topics in the 19th century, among which we picked references to include in our corpus. We describe below how the manuals were selected to compose the Good Old Manuals corpus (GOM).

2.1. Selection of the books

The first selection of books was collated by looking at the recommended readings accessible on an online platform about agroecological practices. The GOM1 corpus is thus composed of seven books listed in the table below. Additionally, we included 14 more books in the full GOM corpus after discussions with specialists of market gardening. All books are related to market gardening and were published between 1802 and 1912. For the following textual analysis, we only consider the GOM1 section of the corpus. The list of books included in the full GOM

Table 1
List of the books included in the GOM1 corpus.

Author	Date	Title
Combles, Charles-Jean De	1802	L'école du jardin potager
Noisette, Louis	1825	Manuel complet du jardinier maraîcher
Courtois-Gérard, Claude Joseph	1843	Manuel pratique du jardinage
Moreau, J.G. et Daverne, Jean-Jacques	1845	Manuel pratique de la culture maraîchere de Paris
Deby, Julien et Rodigas, François/Emile	1853	Manuel de culture maraîchère
Gressent, Vincent	1863	Le potager moderne
Desmoulins, Philippe	1871	Guide pratique du jardinier français

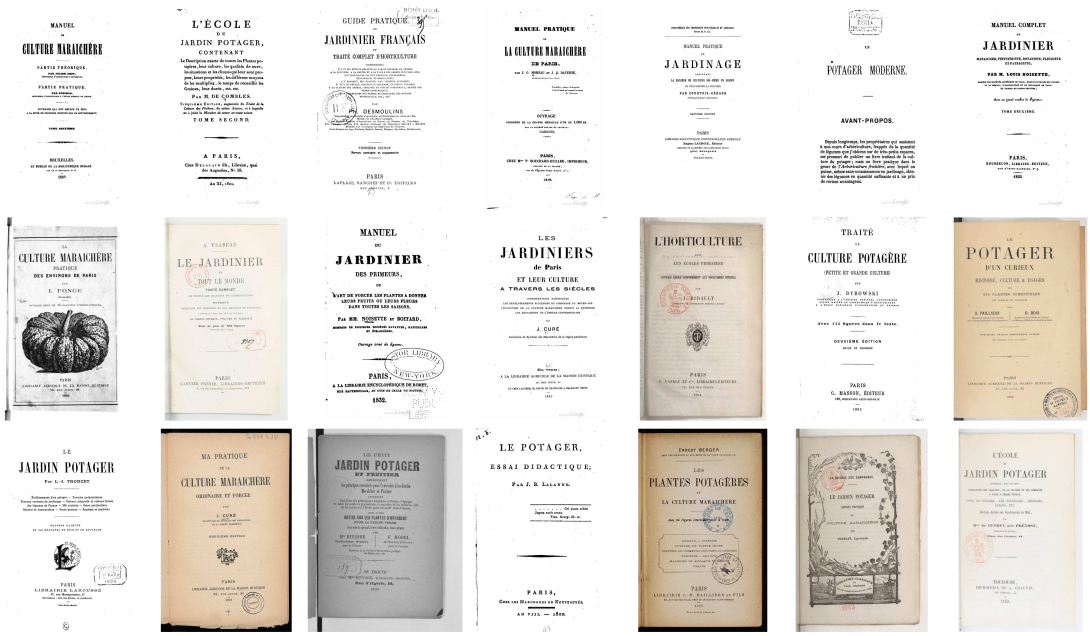

Figure 1: Covers of the books included in the GOM corpus.

corpus is available on the companion website [1].

2.2. Text extraction and preprocessing

The first step in our analysis is to extract the layout of each page, identifying regions of the page occupied by text paragraphs, title, figures or tables using an image segmentation algorithm based on Faster RCNN trained on a large collection of publications [24]. In this process, we could extract 1269 figures and 120 tables. The regions of the images classified as text were then fed to the Tesseract library [20] for optical character recognition (OCR).

As expected, the resulting text still includes many mistakes, so a first preprocessing was done to substitute characters unlikely to appear in the text by their most likely replacement (for ex-

[1] https://sonycslparis.github.io/gom-webapp/

ample ä->à). Next, to correct spelling mistakes from the OCR, we filtered out-of-vocabulary words (using the reference lexicon MORPHALOU3; [19]), for example "avans" instead of "avons". A Bayesian model [17] combining the estimation of the most likely mistakes (using the confusion matrix of the characters [2]) and the closest neighbors using the edit distance with a weight different for words at 1 edit distance and 2 edit distance. For a string s, we select the candidate valid word w maximizing P(w).P(s|w). Where P(w) is the frequency of occurence in a base corpus (FRANTEXT [2] in our case) and P(s|w) is the probability of subsitutions leading from s to w as given by the confusion matrix. From this process, we managed to reduce the number of out-of-vocabulary words from 80000 to 8000.

3. Named entity recognition and linking

It is important to identify the places and people cited in the GOM corpus so that the texts can be properly situated in their appropriate geography and history. For this, we used the out-of-vocabulary words, and selected the ones written starting with a capital letter. We then matched this list to a dictionary of geographical locations including their localization as GPS coordinates. In the remaining words, we checked manually, through web search, in the most commonly cited if those correspond to personalities.

Additionally, for places, there is a common ambiguity in our corpus on whether the name of a location is used to refer to the location or to a variety of plant originating from this location. To disambiguate this, we manually annotated all the mentions of names of locations as referring to the location or to a variety of plant originating from this location.

Based on this recognition of places and people, we were able to visualize both aspects. First, in a graph on Fig.2, we represented the authors and the most cited people (more than 2 times). We drew an edge between an author and a cited person if this person was cited by the author. We see that some authors cite generously, while some others only mention a few people. For example, in the Moreau & Daverne, only Héricart de Thury and Mr Gontier are cited. The book they wrote was a response to a call emitted by the Royal Society of Horticulture, whose director was Héricart de Thury; and Mr Gontier was a market gardener in the region of Nantes and who was among the first to experiment with an innovative technology of the time, the thermosiphon. For places, on Fig.3, we placed circles on a map of France with the radius denoting the frequency of occurrence of the name of place in the GOM1 corpus. We notice that there are many mentions of places in the Paris region, which is expected since a lot of the practices we are interested in are originating from the Paris region.

4. Mapping the key verbs in the GOM corpus

It is interesting to focus on the verbs mentioned in the GOM corpus as they reflect the actions that are important to a market gardener on their farm. We are particularly interested in the verbs that are specific to market gardening, which can be considered as a keyword identification problem. For this, we first lemmatize and POS tag the texts using spacy, a widely used tool for

[2]We used the confusion matrix available at https://github.com/shaneweisz/OCR-Character-Confusion/blob/master/confusion_matrix/confusion_matrix_base.pkl

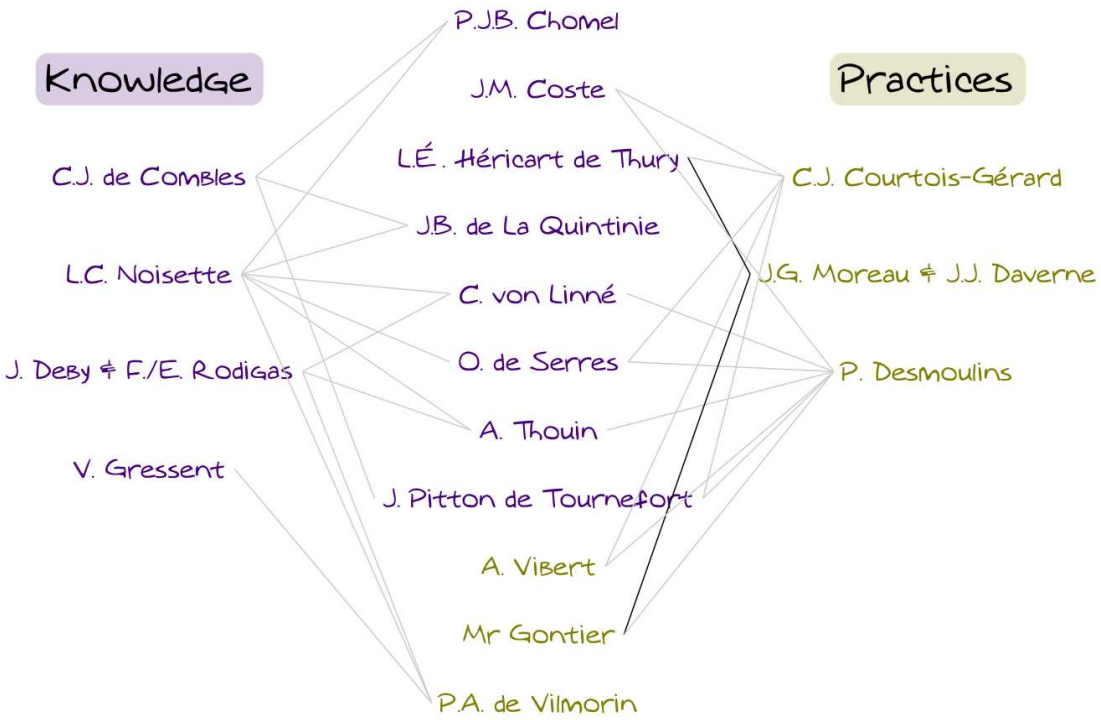

Figure 2: Citations in the GOM corpus. Authors are listed in the left and right columns; while cited people are listed in the central columns. Names in purple refer to people mostly on the knowledge side (professors of agronomy or botany for example) and names in yellow refer to people involved on the practical side (market gardeners, seed sellers,...).

various NLP tasks [3]. Then, similarly to the keyness commonly used in corpus linguistics [18], we measure for each verb the logarithm of the ratio between f_G the frequency of occurrences in the GOM1 corpus and f_F the frequency of occurrences of the verb in a reference corpus, FRANTEXT [2], which gathers 31 M words from periodicals the 19th and 20th century : $k = log(\frac{f_G}{f_F})$

The word cloud in Fig.4 shows the verbs with a size proportional to this index in yellow and the verbs not appearing in FRANTEXT in red with a size proportional to the log of the frequency of occurrences in GOM1.

In the previous representation the location of words has no interpretation and we also want to represent the words in a space where two words located close together would have similar meaning (in the distributional sense). That representation can be useful, for example, to show groups of words clustered together having a similar meaning. We represented each verb using its embedding in a word2vec model trained on a large French corpus [1] and we visualize the map of verbs after reducing the dimension of the embedding to 2 dimensions using UMAP [15] in Fig.5. We can for example identify a cluster of verbs describing actions of the farmer in the field (sarcler-palisser-semer) or verbs related to biological processes of the crops (pommer-

[3]https://spacy.io

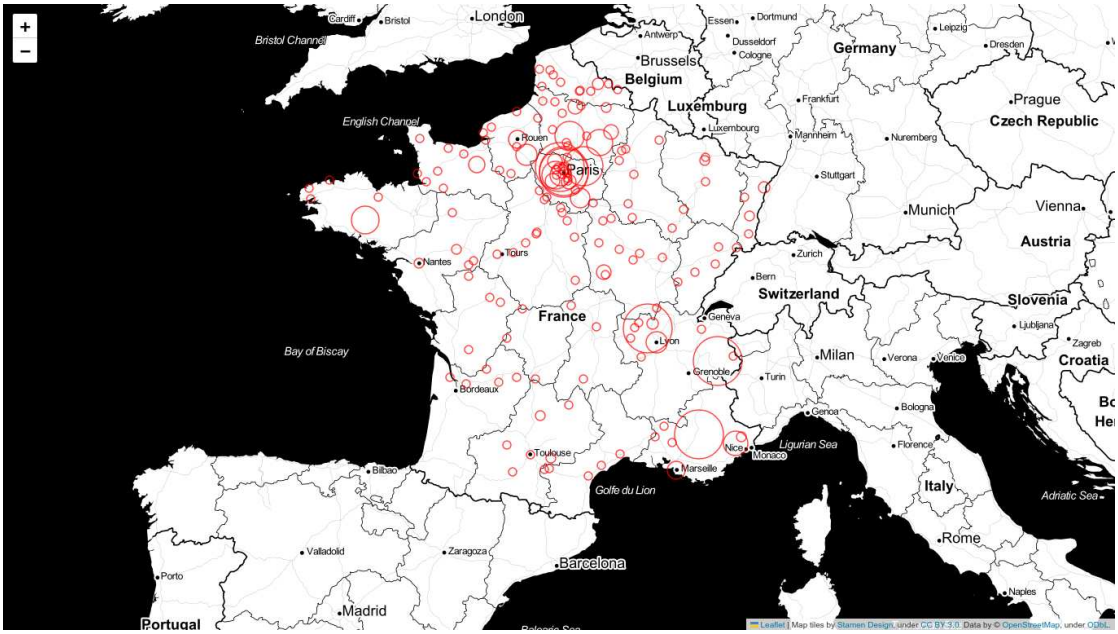

Figure 3: Map of the locations mentioned in the GOM1 corpus. The circles size reflect the number of occurrences in the corpus.

tacheter-fleurir) being grouped together. Such a map is useful to navigate the content of the manuals and the embeddings may be useful to classify parts of the text.

The GOM corpus gathers an rich mixture of practical advice and practical knowledge. It is interesting to study whether the discourse in those books reflects this dichotomy between practices and knowledge. A key feature of the transition of discourse from practice to knowledge is nominalization, a linguistic process where nouns are derived from verbs [11]. Thus in the particular example of the verb arroser ("to water"), we plot the usage statistics in each of the 7 books of the GOM1 corpus. We see, in Fig.6 top panel, that some authors favor much more the use of the verb than the noun, denoting a more practical and less abstract discourse. Also, it is interesting to note that in the case of the verb arroser, there were actually two forms for the corresponding noun: arrosage and arrosement (both meaning "the watering (of crops)"). By plotting the frequencies of occurrence of these two terms in large corpora (Gallica and Google books), it shows that the 19th century is precisely the time during which those 2 terms coexisted, arrosement being used more frequently before; and arrosage becoming dominant after the 19th century. Some references (ATILF) mention a small difference in the meaning of those 2 terms, arrosement being more related to a passive manner for plants to receive water and arrosage referring to a more active process from a human to provide the water.

5. Extracting causality frames

We were also interested in capturing the parts of the discourse reflecting causal relations because in the sentences expressing causality, we may find elements of biological knowledge. For

Figure 4: This word cloud of verbs illustrates which actions were important for market gardener (indicated though size). Red verbs do not appear in the FRANTEXT reference corpus and are therefore specific to market gardening.

example, let us consider the following sentence from Moreau et Daverne:

”Autre observation : la pratique nous a appris que, pendant l'été, si nous arrosons nos romaines durant le grand soleil avec l'eau froide de nos puits, quand elles sont près de se coiffer ou déjà coiffées, cela détermine dans leur intérieur des taches de pourriture; nous disons alors que la romaine est mouchetée : dans cet état, elle n'est plus bonne pour la vente.” “Another observation: practice has taught us that, during the summer, if we water our romaine plants in the hot sun with cold water from our wells, when they are about to be capped or have already been capped, this causes spots of rot inside them; we then say that the romaine is speckled: in this state, it is no longer fit for sale.”

Here, the authors draw a causal relation between on the one hand the watering of the crops with cold water when it's hot at a specific growth stage of the crops; and on the other hand the rotting of their leaves. Even though knowledge was too scarce at the time to fully explain this phenomenon, namely that these conditions are favoring the growth of fungi, it is clearly some kind of knowledge about biology that is encapsulated in the text.

To detect such causal relationships in a systematic matter, we are currently performing a Frame-Semantic analysis [8] of the corpus. A Semantic Frame is a structured piece of knowledge that can be considered as a template of a scene with several open slots (called Frame Elements) that need to be filled in. One example is the Causality Frame, which comes with ‘core’ Frame Elements such as Cause and Effect, and ‘non-core’ elements that further qualify

Figure 5: A vector representation of verbs allows us to identify clusters of related activities. One cluster contains actions that focus on work in the field (such as 'sarcler', 'semer', and 'palisser') in the region on the right; while another cluster at the bottom left groups together biological processes of crops (such as 'pommer', 'fleurir' and 'tacheter')..

the relation. The linguistic sister theory of Frame Semantics is called Construction Grammar [9], which explores how semantic frames get expressed in language through associations of form and meaning called constructions. There are typically two types of constructions involved. The first kind are frame-evoking constructions (usually lexical items or multiword expressions), which activate a semantic frame. In French, numerous words and multiword expressions evoke the Causality frame, such as à cause de "because of", parce que "because", occasionner "to bring about", suite à "due to", and so on. The second type are grammatical constructions (typically argument structure constructions; [10]), which identify which phrases of a sentence should be mapped onto which Frame Elements.

Our Semantic Frame Extractor has been implemented in Fluid Construction Grammar (FCG; [22]), an open-source computational grammar formalism for engineering Construction Grammars, following the methodology described by [3], who developed a Causality Frame Extractor for English. Our approach integrates several knowledge sources: • Input sentences are preprocessed using both a dependency parser and a constituency parser (such as the Berkeley Neural Parser; [13]). These different structures are integrated in a single syntactic representation of a sentence using feature structures. During the training phrase, annotations of semantic frames are mapped onto the syntactic analysis to extract recurrent patterns of form-meaning

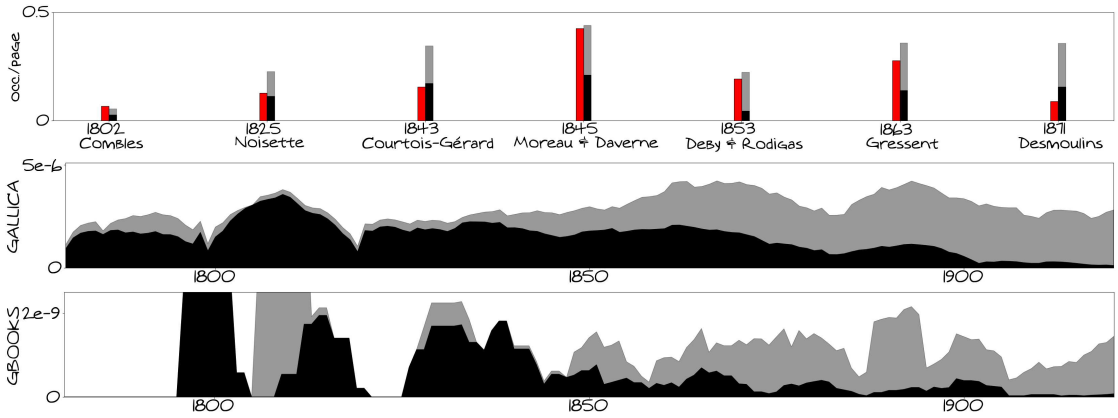

Figure 6: (Top) Comparison of the usage, in the GOM corpus, of the verb "arroser" (red) compared to its nominalizations "arrosage" (in black) and "arrosement" (in gray). Comparaison of the frequency of occurence of "arrosage" (in black) and "arrosement" (in gray) in Gallica (Middle) and Google books (Bottom).

associations (constructions). Patterns that are not frequent enough are pruned because they typically result from annotation errors. The semantic annotations were taken from the French FrameNet, developed within the ASFALDA project [5]. The French FrameNet project has explicitly focused on Causality as one of its main domains, and includes 11 distinct Causality frames and 217 distinct frame-evoking elements. Fig.7 illustrates the kind of information that can be extracted using this method. On the left is an input sentence, and on the right is a Causality frame that was detected. As can be seen, the verb form détermine (here: "causes") is the frame-evoking element (FEE). It has designated its subject (cela "that") as the Cause, and its direct object (tâches de pourriture "spots of rot") as the Effect.

In its current form, a Causal Frame extractor is already useful because it can search through a text for instances of causal language, and then present the results to the human reader. We are currently evaluating how well a frame extractor trained on contemporary French data can be applied to the Good Old Manual corpus. For this, we are annotating a test set of causal expressions that can be found in the corpus. Moreover, as can be seen in Figure 7, the Frame Extractor currently identifies Frame Elements through syntactic relations, so the syntactic subject cela was assigned the role of Cause rather than the semantic subject (printed in italics), which is what really matters for extracting knowledge. Future work will therefore have to include anaphor resolution and tracking entities across longer spans in the discourse.

6. Conclusion

Old texts are often treasure troves of past knowledge that has become almost inaccessible or even forgotten as societies evolve. Especially "good old" manuals, which have so far been neglected, offer a great potential source of information about the knowledge and practices of a given time and place. In this paper, we have illustrated how a suite of techniques from Digital Humanities, natural language processing, statistical analysis and data visualization, can be

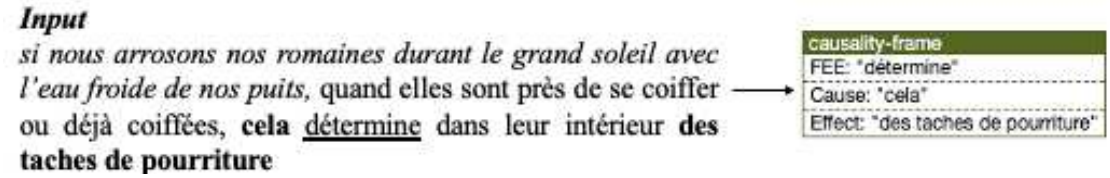

Figure 7: This Figure shows an input sentence on the left, with its Frame Elements indicated in boldface, and its frame-evoking element underlined. On the right is a Causality frame that was extracted from this sentence, as it is visualized in Fluid Construction Grammar's web interface.

exploited to make such texts not only accessible, but also more meaningful to human readers.

More specifically, we have introduced the Good Old Manual corpus of 19th century texts about French market gardening, particularly in the Paris region. These techniques have recently gained a renewed interest because they offer insights into increased efficiency for farming on small plots of lands, known as the French Method. We have demonstrated how the most prominent actors at the time can be situated in a social and geographic network through named entity linking; how activities that are relevant and meaningful to specific topics such as market gardening can be visualized through word clouds and word embedding spaces, and how more fine-grained knowledge could potentially be mined through semantic parsing.

References

[1] H. Abdine, C. Xypolopoulos, M. K. Eddine, and M. Vazirgiannis. "Evaluation of word embeddings from large-scale French web content". In: *arXiv preprint arXiv:2105.01990* (2021).

[2] P. Bernard, J. Lecomte, J. Dendien, and J.-M. Pierrel. "Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella." In: *Lrec*. Citeseer. 2002.

[3] K. Beuls, P. Van Eecke, and V. S. Cangalovic. "A computational construction grammar approach to semantic frame extraction". In: *Linguistics Vanguard* 7.1 (2021), p. 20180015.

[4] C. de Carné-Carnavalet. *Le maraîchage sur petite surface: La French Method: une agriculture urbaine ou périurbaine*. Editions de Terran, 2020.

[5] M. Djemaa, M. Candito, P. Muller, and L. Vieu. "Corpus annotation within the French FrameNet: a domain-by-domain methodology". In: *Tenth international conference on language resources and evaluation (LREC 2016)*. 2016.

[6] M. Ehrmann, M. Düring, C. Neudecker, and A. Doucet. "Computational Approaches to Digitised Historical Newspapers (Dagstuhl Seminar 22292)". In: (2023).

[7] R. S. Ferguson and S. T. Lovell. "Permaculture for agroecology: design, movement, practice, and worldview. A review". In: *Agronomy for sustainable development* 34 (2014), pp. 251–274.

[8] C. J. Fillmore and C. Baker. "A frames approach to semantic analysis". In: (2009).

[9] M. Fried and J.-O. Östman. “Construction Grammar: A thumbnail sketch”. In: *Construction Grammar in a cross-language perspective* 1 (2004), pp. 1–86.

[10] A. E. Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

[11] M. A. K. Halliday and J. R. Martin. *Writing science: Literacy and discursive power*. Routledge, 2003.

[12] P. Hervé-Gruyer and C. Hervé-Gruyer. *Miraculous abundance: One quarter acre, two French farmers, and enough food to feed the world*. Chelsea Green Publishing, 2016.

[13] N. Kitaev and D. Klein. “Constituency parsing with a self-attentive encoder”. In: *arXiv preprint arXiv:1805.01052* (2018).

[14] O. Martin. “French Intensive Gardening: A Retrospective”. In: (2008).

[15] L. McInnes, J. Healy, and J. Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[16] F. Moretti. *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

[17] P. Norvig. “How to write a spelling corrector”. In: *De: http://norvig. com/spell-correct. html* (2007).

[18] P. Rayson. “From key words to key semantic domains”. In: *International journal of corpus linguistics* 13.4 (2008), pp. 519–549.

[19] L. Romary, S. Salmon-Alt, and G. Francopoulo. “Standards going concrete: from LMF to Morphalou”. In: *The 20th International Conference on Computational Linguistics-COLING 2004*. 2004.

[20] R. Smith. “An overview of the Tesseract OCR engine”. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. Ieee. 2007, pp. 629–633.

[21] M. Terras, J. Nyhan, and E. Vanhoutte. *Defining digital humanities: a reader*. Routledge, 2016.

[22] R. van Trijp, K. Beuls, and P. Van Eecke. “The FCG Editor: An innovative environment for engineering computational construction grammars”. In: *Plos One* 17.6 (2022), e0269708.

[23] A. Wezel, S. Bellon, T. Doré, C. Francis, D. Vallod, and C. David. “Agroecology as a science, a movement and a practice. A review”. In: *Agronomy for sustainable development* 29 (2009), pp. 503–515.

[24] X. Zhong, J. Tang, and A. J. Yepes. “Publaynet: largest dataset ever for document layout analysis”. In: *2019 International conference on document analysis and recognition (ICDAR)*. Ieee. 2019, pp. 1015–1022.