



Modern Methods for Medieval Marseille: Prototyping Natural Language Processing Workflows for Entity Extraction from Highly Specialized Corpora

Citation

Humphrey, Emma Louise. 2022. Modern Methods for Medieval Marseille: Prototyping Natural Language Processing Workflows for Entity Extraction from Highly Specialized Corpora. Bachelor's thesis, Harvard College.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37372705>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

**Modern Methods for Medieval Marseille:
Prototyping Natural Language Processing Workflows
for Entity Extraction from Highly Specialized Corpora**

Emma Humphrey

Presented to the Department of Applied Mathematics
in partial fulfillment of the requirements
for a Bachelor of Arts degree with Honors

Harvard College
Cambridge, Massachusetts

March 26, 2022

Abstract

This thesis constitutes an investigation into the ways in which current Natural Language Processing (NLP) tools and methods can be applied to historical corpora in a variety of languages from the Medieval period. In doing so, it is most specifically concerned with a set of household inventories from early 15th century Marseille, written in Latin, and detailing the tangible objects associated with a vast set of individual households. For this corpus I have developed a process which extracts objects and their attributes from dependency parsed renderings of these inventories, retraining a model for dependency parsing based on their particularities and developing a methodology for extraction in order to create computationally relevant, searchable, versions of these material culture records, demonstrating along the way the divergences from typical process necessary to bridge contemporary NLP and historical language.

Acknowledgments

I am beyond grateful at the conclusion of this project for all of the support I received throughout the thesis-writing process. Primarily, I am thankful for my thesis advisor, Gabriel Pizzorno, who was immensely generous to me with his time, his knowledge, and his patience. He possesses not only a rare breed of expertise but also an uncanny ability to remain both sharp and silly while mentoring, sure of the necessary next steps and able to guide me there while also laughing at and recognizing the inherent ridiculousness present in much of research. I would very literally never have been able to do this without him. I am further grateful to my roommates, who make me happy and keep me sane and with whom I cannot wait to waste the rest of this semester. Finally, to my family, who have supported me up until this point and always, who believe me capable of anything, and who had wonderful title suggestions which I did not use <3.

1 Introduction

This thesis details the development of a prototype end-to-end process for the translation, extraction, and classification of objects described in household inventories from the late Middle Ages. This effort was undertaken as part of the Documentary Archaeology of the Late Medieval Europe (DALME) project out of the Harvard Department of History, which is working to make these records available as computer-actionable data for researchers around the world¹. I have been working directly with Professor Gabriel Pizzorno on much of this project, collaborating with and learning from him at first as we retrained Latin Dependency Parsing models, and then working on my own to develop a methodology for extracting objects and attributes from our early results. Together we have worked to apply cutting-edge Natural Language Processing (NLP) methods to these corpora, manipulating and developing tools for this specific data set as necessary, in order to streamline the process and make a greater volume of this source material available for analysis than previously possible.

This thesis is perhaps unconventional in that it details a research process which is ongoing, an exploration of the possibilities and limitations of applying existing tools and methods to these types of datasets and working to overcome the limitations found wherever possible. I do not have results so much as I have a better awareness of algorithmic constraints, of linguistics, of all the ways in which the work that we have completed will need to be manipulated and then manipulated again in order to work for other dialects and contexts of interest. Much of the process has been, if not quite trial and error, then research and roadblocks. The frame is a real research problem: an attempt to fill a lacuna in current interdisciplinary capability, a bridge between two disciplines which construct themselves, often unconsciously, such that they actively repel one another.

¹ See: <https://dalme.org/>

This has felt, often in immediate succession, both helpless and exhilarating. I spent much of the first half of this project desperately trying to learn enough Latin grammar such that I could recognize which elements of existing NLP models would be relevant, and then the next day enough about an NLP model to tell whether it could work for our corpus. Both sides of this work are highly specific, and reconciling them often felt like my biggest challenge. But, there is a compelling power in coaxing the computer towards comprehension of language so specific and flawed and contrary to its nature. In looking at something ancient with all the tools of modernity and coming up short, reminded at every corner of the necessity for humanity in untangling the nebulous net of our history. I have proven, with lots of coaxing and compromise, that this untangling is possible for one very narrow window into the medieval world. In the process, I have further proven to myself that it is futile, and funny, to try and understand one another across the centuries using anything other than empathy. And that it is rewarding regardless, to try.

2 Problem and Historical Context

Europe was, during the Medieval period, stereotypically steeped in grime and darkness. If you believe what you read in Steven Pinker's *Better Angels*, it was a stretch of time characterized by ruthless, sexual torture². If you're primarily influenced by mythologies or by your local Renaissance Fair, you might consider the medieval a time of knights and their jousting battles, of quests for the benefit of pale, helpless damsels who affixed long ribbons to the tips of their pointy hats. You may likely have been biased against the people of the Middle Ages in your early history classes, been taught to think of them as dirty and diseased, to picture rats crawling out of the river Thames and through the body-littered streets of London during the bubonic plague. While we would, perhaps, be more empathetic about their relative fault in encouraging a pandemic today, these stereotypes persist, and they combine to create a pervasive mental image of the Middle Ages which offers an incomplete picture of daily life for most during this period.

However, these details are not entirely without grounding in some reality, or at least in the way reality has been tampered with, and turned a profit out of, throughout history. Pinker, when doing his research, really visited medieval museums of torture in order to make his case. Many of the gruesome devices on display there are counterfeit eighteenth century playthings of the rich, misinterpreted and misattributed³. Regardless, these objects remain in museums. They sell tickets to tourists across Europe.

Art museums, when accounting for the Medieval period, often dedicate a dark corner room full of suits of armor and stained-glass depictions of tortured saints. This creates for us a Middle Ages of either king or clergy, brave knight or doomed worshipper, and it erases centuries of daily toil, prevents us from considering the

² Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking.

³ Bishop, C. (2014). The 'pear of anguish': Truth, torture and dark medievalism. *International Journal of Cultural Studies*, 17(6), 591–602.
<https://doi.org/10.1177/1367877914528531>.

lives of the masses as potentially distinct, interesting, and defined by circumstances outside of those we have been taught to immediately conjure.

Scholars today are particularly interested in pulling back this conventionalized curtain from the Middle Ages. The DALME project aims to “increase our understanding of Europe’s material horizons during the later Middle Ages,” and is working to catalog, translate, and interpret a vast cache of primary source material from across Europe during this time period⁴. Some of this material includes lists of tangible objects included in legal records, from testaments to merchant inventories. These documents offer a rich window into the daily lives of a wide swathe of medieval society. The DALME project is working towards the creation of a database of these objects that not only makes them accessible and searchable, but also comparable, both across regions, languages, and periods, and to other sources of material culture information, such as museum collections and archaeological excavations. Key to achieving this goal is the establishment of a semantic link between the primary textual description of the object and its corresponding value in a common ontology, for example the Getty Art and Architecture Thesaurus (AAT), which is used for describing objects and material attributes⁵. Through these semantic links, something described as a copper pitcher would be connected to the AAT values for `copper (metal)` [300011020], `copper (color)` [300311190], and `pitcher (vessel)` [300194765]. The objects and their attributes can then be manipulated based on these conceptual values, bypassing the language-specific terms used in the original documents.

This is, however, a difficult task. Whichever clichés we wish to defy about the “dark ages,” they were a time of inconsistent record keeping and generally low literacy, such that the materials we do have are riddled with inconsistencies in both their grammar and their content. What’s more, we’re disproportionately reliant

⁴ Smail, D., Pizzorno, G., & Morreale, L. (2020b). Overview. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/overview/>.

⁵ See: <https://www.getty.edu/research/tools/vocabularies/aat/about.html>

on these sources to paint a complete picture of material culture. Archaeological findings are always limited from any time period, and especially so for objects of daily life in the poorer classes. Things like textiles and metals tend not to survive in the archaeological record. Our understanding is bounded by the things which can physically withstand the test of time, or which were carefully passed down and have landed in museum collections.

However, within our acknowledgement of these inconsistencies, there was record keeping, and a sense of literacy in most communities, and proof of these truths has prevailed. For the early Middle Ages, the archaeological record remains the most consistent resource for studying material culture. But ca 1250, writing on paper became much more popular throughout Europe. Record keeping was suddenly accessible to all classes of people, and this is demonstrated in the explosion of documents found dating to this era and afterwards⁶. The DALME project is interested in a subset of this sudden documentation, household inventories, which demonstrate in vivid, diverse detail “the deep interdependence of humans and things.”⁷ Inventories were an opportunity to lay official claims, to work through complicated arrangements, to catalogue ownership. Communities found a myriad of needs for this category of record keeping, and while the implementation and exact function varied from place to place throughout medieval Europe, the surviving records teach us lots about the people who made them and the conditions in which they lived.

Imagine for a moment that you are a recent widow in 15th century Marseille. Lady Massileta, perhaps, widow of Guillemus de Cavallione in mid-November, 1405⁸. Your husband was a man of little means, and now he is dead, and as a

⁶ Smail, D., Pizzorno, G., & Morreale, L. (2020a). Methodology: A Documentary Archaeology. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/methodology/>.

⁷ Ibid

⁸ Smail, D. L. (2022). Inventory of Guillelmus de Cavallione. In D. Smail, G. Pizzorno, & L. Morreale (Eds.), *The Documentary Archaeology of Late Medieval Europe*.

result you are responsible for putting his affairs in order. This involves a household inventory of your late husband's possessions, a legal list which will determine whether it is worthwhile for you to accept your inheritance, or if his debts are greater than his assets. "Husband's possessions" is a distinct and specific label. He was, and is now especially, separate from you, even if he was understood to control the household's wealth in life. The inventory will, in a broad sense, cover everything related to his estate. But there are nuances worth noting as you and your late husband's brother Jacob, who is the executor of the document on your behalf, make your considerations. For example, you are understood to own objects adding up to the equivalent of your dowry. If you came to your husband with a silver candlestick which the two of you, in your marriage, had sold – you, as the widow to the estate, are owed other goods in the house equal to the value of your original candlestick. There are other subtleties as well: the shoes your children wear on their feet every day are theirs, not considered a part of the household wealth, the clothes which you wear on your back are the same. Objects which had been promised to others, either as a means of paying a debt or because of some original provenance, are marked as such and not included in the inventory. A shrewd wife, as perhaps you are, may ensure that many objects are said to be claimed this way in order to hide the true assets of the estate and further protect your children, as Thomas Keun has argued was common practice as these inheritances began to be settled with greater institutional oversight⁹. Land, animals, and other large-scale resources were easy to quantify, and easy to see. Household wealth, the materials which made up a life, were considerably more difficult to account for as completely, and the records of your household would likely reflect this uneasy balance between the requirements of record keeping and the vicissitudes of daily life.

⁹ <https://purl.dalme.org/55a1cca4-8d71-4fb0-abb8-4ae0b350acaa/>.
Keuhn, T. (2008). Estate inventories as legal instruments of credit in renaissance italy. *Heirs, kin, and creditors in renaissance florence*. Cambridge University Press.

These inconsistencies are interesting, but challenging for scholars, who are often interested in developing a sense of general norms – what kinds of lamps did people own? How many pots did they have? Of what materials? This variability, however, while palpable in individual records, largely evens out over the context of hundreds and hundreds of household inventories from any given region. The problem, then, is to translate, parse, and account for every object across these hundreds of inventories, which leads then to a nice match with another problem in contemporary history research: the desire to apply contemporary data science approaches to large data sets, to reach conclusions bolstered by statistical analysis. This inconvenient marriage of modern data processing methods and historical data sets, of making the masses of records specific and quantifiable, is the focus of this project.

One of the most important things that a contemporary data science approach requires, however, is a large set of clean and organized data, something which the historical record rarely ever hands us. If we go back, quickly, to your role as a Medieval Marseille widow, the person you would have needed to hire to make this record of your remaining estate would have been a notary, a new profession of relative prestige across Europe as paper records became more common. Notaries were educated and often formally associated with the church, popping up quickly in towns throughout Italy and Southern France around the middle of the twelfth century, and comfortably straddling their role as ecclesiastical officers, and thus individuals with the protection of the church, and the largely secular, delicate nature of their essential, evidence-preserving work¹⁰.

Notaries were literate men, as demanded by their profession, but not necessarily literate in the way that we would define such a qualification today. Many would speak and be most comfortable in the dialect of their region: Provençal, in the

¹⁰ Murray, J. M. (1993). The profession of notary public in medieval flanders. *The Legal History Review*, 61, 3–31.

case of Marseille, or Tuscan in Lucca, Italy. Records, however, were often written in Latin, and while these officials knew Latin, they were not taught to the same standard as, say, a monk or a bishop. As a result, the records in their hand are riddled with both small errors and dialectical leanings. Many of these mistakes are the kind of thing which a trained eye would mentally fix and immediately understand, small errors like mismatched use of the six Latin cases on the end of nouns and adjectives or misplaced adpositions (prepositions, postpositions, etc) within a sentence. Even if perfect correctness in the grammar could be assumed, the DALME records contain inventories in a vast array of dialects, often several within the same regional area, each requiring a high level of expertise for the mere process of transcription.

In order to process these inventories on the scale necessary for the normalization of object-based inconsistencies, as described above, some sort of automation for translating, parsing, and cataloguing the inventories, despite their variability, is necessary. NLP is the obvious, and well-researched solution here. But popular NLP models are built for languages for which there is a rich existing corpus of translated and tagged sample text, something the medieval dialects which interest us, lack. NLP models are additionally easily tripped up by small errors such as irregularly applied Roman/Arabic numerals, or the omission of a full stop at the end of each object in a list¹¹. These small mistakes, which litter our corpora of household inventories, necessitate exhaustive pre-processing and manual oversight.

The primary goal, then, is to find a way to bend existing computational linguistic tools to this very specific dataset, harnessing the power of contemporary NLP advancements towards cleaning, parsing and organizing these inventories into a format which is easily understandable and more readily usable. Then, tak-

¹¹ Smail, D., Pizzorno, G., & Morreale, L. (2021). Workflow. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/workflow/>.

ing these records and extracting the information of interest to us - the objects and their attributes – contributing, item by item, to a larger sense of what medieval people held, and kept, and used, and doing so on a massive scale in an automated and replicable manner.

3 Background and Early Attempts

Since DALME’s collection of documentary sources spans so many languages, Professor Pizzorno and I discussed and began our process with a set of well-defined and documented tools for contemporary NLP, in the hopes that whichever methodology we used for retrofitting those models in one language might follow a similar, and replicable process for other languages – eliminating the need to start completely from scratch for each collection of inventories.

To this end, we chose Stanza pipelines from the Stanford NLP group¹², a set of what will be referenced here as “pre-trained models,” which are based on major Universal Dependency treebanks (UD Treebanks, moving forward, cross-language linguistic typology and dependency grammar models) consisting of everything from the text of an original King James Bible to Italian tweets. The pre-trained models are trained on a very wide range of sources. This, in theory, is a positive for NLP models – it is often purported that “more data is better data,” when it comes to training computers to understand language. However, there are often “serious inconsistencies in the way the dependency relations are used in the respective treebanks,” originating from differences in the ways people communicate across platforms in our contemporary era, or “idiosyncrasies in the linguistic treatment of a certain construction pursued in one treebanking team”¹³. This holds potential for further harm in our project in particular, where something trained on a treebank which is tracking contemporary language usage, such as social media posts, could bias the dependency parsings further away from the language that our sources are already not mapping particularly well. As a result, we needed to evaluate the performance of the different models against our data.

We began this evaluation with a set of one hundred inventories from the city

¹² See: https://stanfordnlp.github.io/stanza/available_models.html

¹³ Dönicke, T., Yu, X., & Kuhn, J. (2020). ‘identifying and handling cross-treebank inconsistencies in ud: A pilot study. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 67–75.

of Marseille in southern France, as these had already been manually annotated with universal part of speech (UPOS) and lemma information. There were four models pre-trained on Latin treebanks on POS-tagging and lemmatization tasks¹⁴ and after testing each of them, the resulting scores, in the 0.59-0.67 range for POS-tagging and 0.52-0.60 for lemmatization, suggested that the models would perform very poorly on more complex tasks, such as dependency parsing, with scores estimated well below 0.5¹⁵.

So, we had pre-trained models but not very successful ones, and were in need of tangible ways of improving their performance. In NLP neural network approaches, words, alongside semantic information, are represented via word embedding vectors. Stanza provides functionality¹⁶ to download and inspect these vectors, and to run their pipelines using new vector spaces. This course of action was immediately of interest, particularly because of research regarding the limitations of word embedding vectors. Word embeddings are susceptible to meaning conflation deficiency, or “the inability to discriminate among different meanings of a word”¹⁷. Word embeddings, by design, represent each word as a point in the corresponding vector space representation of a given language. However, this is inherently limiting. Many words in all languages have different meanings in different contexts. Further, and more specifically for the DALME corpora, even if we ignore this potential source of error in a properly trained, perfectly relevant neural pipeline – our sources contain specific uses of words which are particular to a certain dialect or region. If the existing vector space representation of Latin which we inherited

¹⁴ Each of the models tested can be found below:
https://universaldependencies.org/treebanks/la_ittb/index.html,
https://universaldependencies.org/treebanks/la_llct/index.html,
https://universaldependencies.org/treebanks/la_perseus/index.html,
https://universaldependencies.org/treebanks/la_proiel/index.html

¹⁵ Pizzorno, G. (forthcoming). Bootstrapping machine learning models for low-resourced, highly specific language corpora: A case study from late-medieval marseille. *Digital Humanities Quarterly*.

¹⁶ https://stanfordnlp.github.io/stanza/word_vectors.html

¹⁷ Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *ArXiv:1805.04032 [Cs]*.

from Stanza didn’t account for these word embeddings, as we knew that it often did not, it was reasonable to think that perhaps a good starting point would be to supplement the semantic matrix: to bend it more closely towards our dialects of interest.

There has been much work done towards correcting for undesirable word embeddings, particularly because they are often based on context, meaning that biased source material (as all source material is bound to be), creates biased associations. This often manifests in NLP models as gender bias. The model sees and codifies associations such as **king:queen::man:woman**, where such a breakdown is valid, but it also sees them in **doctor:nurse::man:woman**, where we would likely prefer that it did not. Such inapplicable associations certainly exist between our ideal, mythical tagged corpora for our dialects of interest and the Latin pipeline we had in our reality. Models with implicit matrix factorization can become unbiased by using subspace projection on de-biased vectors after the model has been established. Ethayarajh and colleagues, for example, managed to de-bias their vectors by subtracting “from each biased word vector its projection on a ‘gender bias subspace.’”¹⁸. The biases within the DALME corpora, however, are not as simply constrained, with potential permutations of problematic subspaces stretching out widely, and beyond simple definition or subtraction. As a result, while we continued to consider word embeddings and their implications, it seemed unlikely that the errors and lack of accuracy occurring for the household inventories would be fixed in this way.

While removing specific biases appeared to be an inaccessible goal, there has also been research into manipulating the dimensionality of word embedding vectors to impact the extent to which they fit the corpus on which they were trained.

¹⁸ Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705. <https://doi.org/10.18653/v1/P19-1166>.

Yin and Shen, in 2018¹⁹, pointed out the often-random choice of dimensionality within most NLP research and application, with many teams using 300 since that was the number used in the skipgram word2vec model in 2013²⁰. Yin and Shen developed a new metric, pairwise inner product (PIP) loss, the normalization of the difference between the PIP matrix of two word embedding matrices on the same vocabulary. Minimizing PIP loss then, between various embedding matrices, can establish the best dimensionality choice for any given data set. This approach seemed promising: here was a way to manipulate word embeddings by comparison which didn’t require direct, linguistic-based manipulation of the embeddings themselves. However, this method, as applied, did not show significant improvements in performance. Dimensionality reduction resulted in score increases in the 1-2% range for the core NLP tasks in the pipeline²¹.

Our next attempt was tangentially inspired by the work of a research group attempting to develop the first treebank for Occitan, a Romance language spoken in the south of France but which has “no official status in the country” today, and no existing UD treebank²². This team’s initial treebank was based exclusively on texts in the Lengadocian dialect of Occitan, and when they wanted to extend the corpus to handle three additional dialects, they found that “parsing models trained on Occitan dialects achieve better results than a delexicalized model trained on other Romance languages despite the latter training corpus being much larger (20K vs 900K tokens)”²³. A very small, annotated corpus of text in the

¹⁹ Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *CoRR*, *abs/1812.04224*. <http://arxiv.org/abs/1812.04224>.

²⁰ Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. <http://arxiv.org/abs/1301.3781>.

²¹ Pizzorno, G. (forthcoming). Bootstrapping machine learning models for low-resourced, highly specific language corpora: A case study from late-medieval marseille. *Digital Humanities Quarterly*.

²² Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., & Sibille, J. (2020). A four-dialect treebank for Occitan: Building process and parsing experiments. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 140–149. <https://aclanthology.org/2020.vardial-1.13>.

²³ Ibid

closer dialect performed better on the new dialects than a significantly larger corpus of generalized Romance languages. This latter method is very similar to what we had been trying to no avail, attempting to manipulate the robust and existing Latin pipeline to fit our specific purpose. We had been convinced by the traditional NLP dogma that more data is better data, many contemporary NLP approaches are moving towards Data Augmentation, demanding larger training sets for better accuracy²⁴. But would less, more specific data, be more effective than the generalized massive Stanza sets from which we had been starting?

To answer this question, Professor Pizzorno manually annotated a set of 200 statistically representative sentences from the Marseille corpus, and then used this very small dataset to train a model. The resulting scores showed a significant increase in the model’s accuracy. Given the initial success of this approach, especially when compared to the very minor improvements achieved earlier, we proceeded to retrain successive models. There was significant improvement in accuracy with each consecutive batch. The graph below shows shrinking orange bars representing the time needed to create the new seed of test data, while the blue bars, depicting the accuracy of the model, continued to improve with each consecutive retraining. New seeds were each parsed with the latest model, such that the improved accuracy continued to build upon itself.

²⁴ Bayer, M., Kaufhold, M., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2021). Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *CoRR*, *abs/2103.14453*. <https://arxiv.org/abs/2103.14453>.

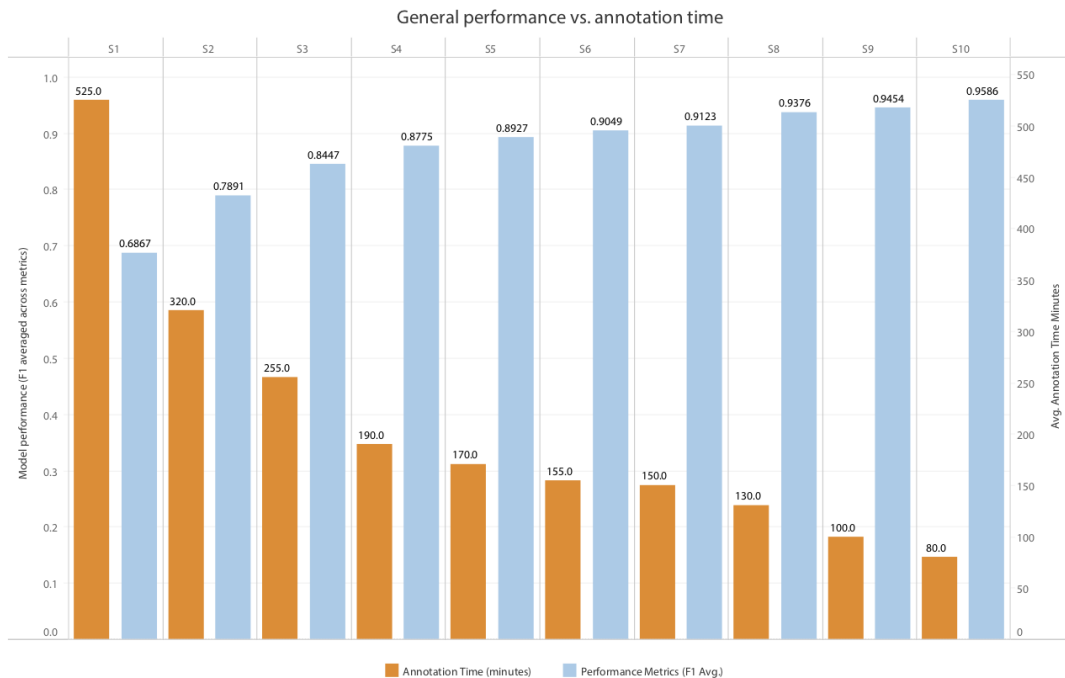


Figure 1: Performance of the Latin Pipeline vs. amount of time to manually annotate the sentences for the associated seed.

This was a very exciting process to watch. But, the DALME project contains inventories in several more languages and dialects than just this specific corner of dialectal Latin on which we had been focused, and while we were glad to have developed a method for this one subset, we knew that this several-round manual annotation process could not be maintained for all of DALME’s languages from both a time and an expertise standpoint.

As a result, I spent much of the latter half of the fall attempting to generalize these results and test the extent to which manual annotation would be necessary in our other languages and dialects of interest. I began with the Italian corpus which, similarly to the Latin, is mostly comprised of descriptive lists of objects left behind as a record of various peoples’ estates. I ran the most basic, overarching pre-trained Stanza pipeline for Italian on a random sample from the cleaned dataset, and I found that unlike for the Latin model, the contemporary Italian

did pretty well predicting the Medieval Italian, and produced parsings which were comprehensible. However, I lacked a “gold standard” corpus to compare the pre-trained Italian model’s performance against. These records have been transcribed, but they have not been translated, and in many of our dialects, there are relatively few people in the world who even could do the work of manually annotating these records to create the kind of reference material we would need in order to calculate any mathematical measure of “success.” The only way to test the effectiveness of the contemporary Italian was to look at the sentences one by one and make a judgement call. Based on these judgements I feel reasonably hopeful that the Latin is one of the more difficult cases within the DALME corpus, and that there will be less manual effort necessary for future subsets of the corpus. But, as always when working with real data, there are bound to be surprise challenges within each of the languages, something which is inherently a problem for this diverse corpus. For the rest of this thesis, results are based on the Latin dependency parsings taken from the retrained model after the several-seed process shown above, with an eye towards the fact that this process would need to be repeated, with new seeds and new manual annotations, for any other language.

4 Parsing the Parsings: An Adventure with Open Information Extraction

We had trained an effective translator and dependency parser. However, the goal was always to extract objects and attributes from these records, a “downstream” NLP problem requiring further manipulation of the dependency parsed sentences²⁵.

The Stanza pipeline was returning a parse-tree, which translated to CONLL-U format included detailed information about each sentence: the characteristics detected for each individual word and the relationships between those words. A typical entry looked something like this:

```
sent_id = 9CEA8040-7762-407A-84E1-AE98CE387B79
text = Item unum morterium lapideum.
1 Item item ADV d----- _ 3 advmod 3:advmod Gloss=next
2 unum unus NUM m-s---nn- Case=Nom|Gender=Neut|Number=Sing 3 nummod 3:nummod Gloss=one
3 morterium morterium NOUN n-s---nn- Case=Nom|Gender=Neut|Number=Sing 0 root 0:root Gloss=mortar
4 lapideum lapideus ADJ a-s---nnp Case=Nom|Degree=Pos|Gender=Neut|Number=Sing 3 amod 3:amod Gloss=of
stone|SpaceAfter=No
5 . . PUNCT u----- _ 3 punct 3:punct Gloss=full_stop
```

This sentence details the existence of a stone mortar (you can see the English translation of each word under *Gloss*, the final entry in each line). Looking at the entry to the left of *Gloss*, you can see that the word “one” here, at index 2, has the relationship *nummod* with the third indexed word in the sentence, the root, “mortar.” This is the power of dependency parsing, it tells us, quite literally, that word 2 is a numerical reference to how many of word three there are, and so on and so forth²⁶.

We can see these relationships expanded a bit in the slightly more complicated

²⁵ This is a collaborative and ongoing project under the larger DALME umbrella. Code progress can be seen at <https://github.com/DALME/dalme-testing/tree/emma-testing>, but DALME has not yet converged all updates and results.

²⁶ For more on these relationships, see: <https://universaldependencies.org/u/dep/>

example below, which describes a slate-blue tunic:

```
sent_id = 639CEC73-5AD2-4D0A-833C-8EDA628838E1
text = Item i alia tunica de livido panno.
1 Item item ADV d----- _ 4 advmod 4:advmod Gloss=next
2 i unus NUM m----- NumForm=Digit 4 nummod 4:nummod Gloss=one
3 alia alius DET p-s---fa- Case=Acc|Gender=Fem|Number=Sing|PronType=Ind 4 det 4:det Gloss=other
4 tunica tunica NOUN n-s---fn- Case=Nom|Gender=Fem|Number=Sing 0 root 0:root Gloss=tunic
5 de de ADP r----- AdpType=Prep 7 case 7:case Gloss=of
6 livido lividus ADJ a-s---nbp Case=Abl|Degree=Pos|Gender=Neut|Number=Sing 7 amod 7:amod Gloss=slate-blue
7 panno pannus NOUN n-s---mb- Case=Abl|Gender=Masc|Number=Sing 4 acl 4:acl Gloss=cloth|SpaceAfter=No
8 . . PUNCT u----- _ 4 punct 4:punct Gloss=full_stop
```

Or, from a different part of the house, and a different kind of sentence, a set of one hundred pine boards of varying width, from 5 to 8 palms across.

```
sent_id = 7018EF44-FBB5-438B-B220-D614CB5C83A2
text = Item centum Tabule pyni de quinque palmis usque octo palmis.
1 Item item ADV d----- _ 3 advmod 3:advmod Gloss=next
2 centum centum NUM m----- NumType=Card 3 nummod 3:nummod Gloss=one hundred
3 Tabule Tabula NOUN n-p---mn- Case=Nom|Gender=Masc|Number=Sing 0 root 0:root Gloss=board
4 pyni pinus NOUN n----- _ 3 nmod 3:nmod Gloss=pine
5 de de ADP r----- AdpType=Prep 7 case 7:case Gloss=of
6 quinque quinque NUM m----- NumType=Card 7 nummod 7:nummod Gloss=five
7 palmis palmus NOUN n-p---md- Case=Dat|Function=Unit|Gender=Masc|Number=Plur 3 acl 3:acl Gloss=palm
8 usque usque CCONJ c----- _ 10 cc 10:cc Gloss=up to
9 octo octo NUM m----- NumType=Card 10 nummod 10:nummod Gloss=eight
10 palmis palmus NOUN n-p---md- Case=Dat|Function=Unit|Gender=Masc|Number=Plur 7 conj 7:conj Gloss=palm|SpaceAfter=No
11 . . PUNCT u----- _ 3 punct 3:punct Gloss=full_stop
```

These entries seem ideal; all the information that I need to extract objects and attributes is glaringly present within them. However, the glare is irreversibly tinted by our human understanding of the information. Of course *we* know what to extract from “one mortar of stone,” but that ability is not helpful at scale. The dependency parsing outputs contained a vast array of sentence types and grammatical constructions, and I needed to develop an exact, mathematical way to extract the information of interest to DALME, and only that information, regardless of form and using computational methods.

In my first pass attempting to solve this problem I took each CoNLL-U entry

and extracted all the numbers, nouns and adjectives based on the part of speech tags for each word. On a very basic level, this simplest brute force approach performs well in terms of retrieving the primary words of interest. However, it is not rigorous, nor is it based in the dependency relations we had just worked so hard to correctly discover. Further, it did not show the relationships between the words, it just returned them in one big list. In more complicated examples like this last one above, several of the important words, like ‘palm’ are marked under relationships like *acl*, which mark the start of a new clause. In this case, the dependency parser recognized that ‘palm’ was a new type of subject with a further description, a unit of measurement with a number attached. I needed to extract clause-wise, preserving the ways in which extra descriptions were attributed to the root objects, such that palm and the number of palms would both always be related to each other *and* shown to modify the original object: 100 pine boards. Extracting based on these relations would obviously require a more sophisticated approach, and so I started researching information extraction based on dependency parsings, knowing that these relationships between words would be key to my ability to extract the correct ones and associate them with one another appropriately.

This led me to the branch of research known as Open Information Extraction (OIE), in which people have worked to solve this same downstream NLP problem. OIE works by pulling out a triple consisting of a subject, object, and relation of related terms or phrases from a given sentence. For example, in the sentence from our corpus detailed above: “Next one other tunic of slate blue cloth,” a successful OIE algorithm would pull the subject (‘one tunic’) relation (‘of’) and object, (‘slate blue cloth.’) OIE was developed with the web and a diverse array of linguistic variety in mind, making it appealing for my purposes, where I was looking for algorithmic insights into how to handle such a strange dataset. Best of all, this means of understanding texts, extracting the important bits is ”inher-

ently unsupervised”²⁷. Having just finished a component of this project which seemed like it would always require more supervision, more manual training, it felt prescient to try, in this part, for a more autonomous method.

OIE is a well-documented field²⁸, with swaths of existing research papers and open-source algorithms available to the public. However, while I was initially hopeful that one of these existing methodologies would match our dataset and goals, I soon found that OIE is overwhelmingly stratified by language. Most approaches, although they may vary in their exact methodologies, apply some sort of rule-based grammar comparison to each sentence – shifting sentences into boxes based on word relations and word types, and then extracting subclauses and related terms based on specific grammatical rules applied to each of these boxes, or sentence types. These rules vary widely by language. In English, for example, a common set of rules would be based on the verb-phrase within each sentence and could include patterns like *subject, verb phrase, direct object*, or *subject, verb phrase, verb prepositional phrase*, or *subject, verb phrase, attribute*. Once sentences are sorted into categories based on which of these grammatical subsets they possess, the algorithm would then pull the subject, object, relation triples from the given construction, following a set of comparative conditionals through to the final extracted output triple²⁹.

These concepts are general, and applicable to this corpus in a broad sense. But they are also based on full, grammatically correct sentences. You may have noticed from the above examples that the Marseille corpus, while stratified based on sentence-mimicking “full-stop” demarcations, is mostly just a list. Our “sentences” very rarely contain verb phrases, the typical hinge upon which a sentence

²⁷ Etzioni, O., Banko, M., & Cafarella, M. J. (2006). Machine reading. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, 1517–1519.

²⁸ See: <https://github.com/gkiril/oie-resources>

²⁹ Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2012). Dependency-based open information extraction. *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 10–18. <https://aclanthology.org/W12-0702>.

turns and is thus, by most OIE models, extracted from. Even if someone had written an OIE algorithm based on the inventories’ very specific kind of Provençal-tinged Latin (and to my knowledge, nobody has publicly written one for *any* Latin), it would not be likely to work for our corpus. The DALME project is interested in the material culture, in the objects, and since these sentences are largely a list of tangible things and since we did the manual annotating which led to the eventual grammatical results, the parsings reflect this subject bias. As a result of all these factors, manipulating one of these existing rule-based approaches for this dataset would not have been feasible.

The conclusion I drew at this point was that I would need to develop my own grammatical gauntlet, a set of rules, following the general OIE model, but particular to this corpus. This strictly rule-based approach, however, presented more problems the closer I looked. Secondary clauses were very often dropped in this system. As one paper pointed out, TEXTRUNNER, one of the earliest OIE leaders would only extract (*Faust, made, a deal*) from the sentence *Faust made a deal with the devil*, highlighting the complicated nature of tracking all the relations within a sentence³⁰. These “incoherent extractions” made up for upwards of 30% of outputs on these highly-specialized and researched models³¹, and it seemed unlikely that I would be able, grammatically, to reach even this level of granularity on my own without a much larger margin of error. The Marseille corpus’ sentences are full of these one-step-further descriptions, these ripe opportunities for incoherent extraction, with the attributes building on one another in several cases: “In the closet, *a cloak, of slate blue cloth, with a fur hood.*” I would drown defining rules for all the potential turns a description could take.

Instead, I was sure there must be a generalizable method for OIE, where

³⁰ Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1535–1545. <https://aclanthology.org/D11-1142>.

³¹ Ibid

machine learning could identify common dependency relation patterns across languages and extract the relevant components without a granular set of specific rules. A paper out of the University of Montpellier³² claimed to have done just that, taken advantage of language-independent similarities between universal dependency parsing outputs to build a model which could perform Open Information Extraction on dependency parsed sentences of any language. The research by Atmani and Lafourcade utilizes a neural OIE model, which feeds a feature vector extracted from each sentence’s dependency parse through a Bi-LSTM layer. They are then able to, after a series of matrix computations, apply the Viterbi algorithm to get the most likely series of predicate phrases³³.

Their method was based on relations, not rules, and they had released a model, already trained, which not only could be manipulated for our specific set of strange sentences but would then also be applicable to any later set of strange sentences in a different language we happened to encounter – it could work for all future permutations of this process for the DALME researchers. It also performed remarkably well, according to the paper. Better than many of the other models I had been studying before, including ones which, while still utilizing a set of language-specific rules, also based their extractions on CONLL-U files, bolstering the correctness of their extractions with dependency parsing³⁴.

While this seemed like a magical solution for several very nice days, after a week or so of attempting to apply their methodology, it became clear to me that this would not be as simple as it seemed in the abstract. Their repository is poorly

³² Atmani, M., & Lafourcade, M. (2021). Universal Dependencies for Multilingual Open Information Extraction. D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, & B. Heinisch (Eds.), *3rd conference on language, data and knowledge (ldk 2021)* (24:1–24:15). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/OASICS.LDK.2021.24>

Keywords: Natural Language Processing, Information Extraction, Machine Learning.
³³ Ibid

³⁴ Corro, L., & Gemulla, R. (2013). Clausie: Clause-based open information extraction. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 355–366. <https://doi.org/10.1145/2488388.2488420>.

documented, which made it very difficult to integrate into our own processes. Further, even without clause extraction based on a language-specific set of rules, their model was still trained, understandably, towards full sentences with verb phrases. My object-as-root list entries would never match the expected inter-sentence relations on which this model and ones like it had been trained. And I was finally realizing that I didn't want them to. If the results I needed from each sentence were so specific, then it was clear that the approach to extracting them would need to be as well. Plus, the algorithm could be simpler this way. It changed the problem from retrofitting existing linguistic tools which leaned, naturally, in the wrong direction, to developing a method from scratch for accessing objects and attributes based on the constructions I had been given.

In my first attempts at creating this new methodology, I tried to follow the examples set within the Multilingual OIE paper. I represented each of our sentences as a matrix, with each word in the sentence as a column, and each potential relationship between words as a row with ones in the spaces of overlap where the given relationship held between two words, such that all of a given word's relationships would be represented within its individual column vector. This created large, inelegant sparse matrices, but I thought that I could make a set of comparison vectors for the patterns of interest to me, ones based on number, root and descriptors or secondary clauses (*nummod*, *root advmod*; *nummod*, *root*; *nummod*, *root*, *acl*), rather than based on the verb phrases I had seen so often implemented. Once I had the matrix for each sentence and a set of comparison vectors, I could compare across, zeroing in on matches and proceeding from there.

However, almost as soon as I had grown confident in this data structure, I found multiple problems with the approach. Namely, it was essentially just as linear a process as combing through the sentences and comparing at the level of the word. Since I was interested in exact matches, not span, there was no elegant way to isolate the presence of a vector within each matrix, I would have needed

to iterate over each vector individually. Further, again and always the issue, my sentences of comparison are imperfect and highly varied. While it's true that many of them follow a pattern which could be isolated, just as many add extra descriptors, contributing to the context of the object in question but making the task of identifying perfectly identical patterns a largely useless one.

Knowing that a better data structure was necessary, I returned to the drawing board and to the first OIE research papers I had found, the ones which I had initially ruled out because their implementations were too language-specific, back when I was looking for an algorithm I could apply immediately. I returned to the paper from Paul Gamallo and Marcus Garcia where they describe a general strategy for dependency-based open information extraction³⁵. Their algorithm is based on Directed Acyclic Graph (DAG) representations of each sentence, where the graphs are used to identify essential clauses and then passed through a relatively small set of clause rules. Their set of grammatical rules “only considers verb-based clause triples and only extracts one triple per clause,” in line with most of its OIE contemporaries³⁶. As discussed above, this verb-focused system doesn't match the needs of my dataset, and so in terms of rule implementation their methodology was not applicable. But, for mathematical representation of the sentences themselves, the DAGs made wonderful sense. The relationships between words would become obvious in this way, with the root of the sentence serving as the literal root of the graph, and all subclauses of interest sitting in child node groups, where they could be analyzed in the context of their relationship to the primary subject or acknowledged as their own separate object and attribute group.

Looking again at the CONLL-U outputs from which I was working, the hi-

³⁵ Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2012). Dependency-based open information extraction. *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 10–18. <https://aclanthology.org/W12-0702>.

³⁶ Ibid

erarchical advantages of a DAG system are immediately clear. In the example with the slate-blue tunic, for instance, we can see that the dependency parsing is already a graph, with the relationship component of each word quite literally pointing back to the root it's referencing.

```
sent_id = 639CEC73-5AD2-4D0A-833C-8EDA628838E1
text = Item i alia tunica de livido panno.
1 Item item ADV d----- _ 4 advmod 4:advmod Gloss=next
2 i unus NUM m----- NumForm=Digit 4 nummod 4:nummod Gloss=one
3 alia alius DET p-s---fa- Case=Acc|Gender=Fem|Number=Sing|PronType=Ind 4 det 4:det Gloss=other
4 tunica tunica NOUN n-s---fn- Case=Nom|Gender=Fem|Number=Sing 0 root 0:root Gloss=tunic
5 de de ADP r----- AdpType=Prep 7 case 7:case Gloss=of
6 livido lividus ADJ a-s---nbp Case=Abl|Degree=Pos|Gender=Neut|Number=Sing 7 amod 7:amod Gloss=slate-blue
7 panno pannus NOUN n-s---mb- Case=Abl|Gender=Masc|Number=Sing 4 acl 4:acl Gloss=cloth|SpaceAfter=No
8 . . PUNCT u----- _ 4 punct 4:punct Gloss=full_stop
```

Our dependency parser recognizes that tunic is the root, the primary object of the sentence, and then it also acknowledges that the sentence contains a subclause providing more information about a specific aspect of the tunic: that it's made of slate blue cloth. The *acl* relationship signifies that there's an additional clause present, and that 'of' and 'slate blue' point to the object in the clause, the cloth, rather than to the tunic itself, while the whole sub clause points to the root, the tunic, to demonstrate that it's a step further description of that same original object.

Throughout the OIE and more general NLP research I've done over the course of this project I had encountered a wide variety of methods for representing sentences and their meanings. In a September 2020 issue of Computational Linguistics, several scholars worked to investigate the different ways in which linguistic theories and NLP methods work to represent sentence meaning for computation³⁷. They conclude with suggestions for ways of unifying these techniques,

³⁷ Žabokrtský, Z., Zeman, D., & Ševčíková, M. (2020). Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3), 605–665. https://doi.org/10.1162/coli_a.00385.

largely based in the Universal Dependencies project for surface-syntactic annotations. Along the way to this conclusion, however, they investigate the varied types of graph-based representations used to represent sentences for different languages and different projects. There are Paninian structures: rooted trees with syntactic phrases in each node, Sequoia frameworks with directed graphs on top of const trees, representing content words in each node using two separate structs. There are rooted trees with extra at-will edges when necessary, directed graphs with cycles and without them, nodes for tokens, for concepts, for phrases³⁸. I was less concerned with their arguments for converging these options than I was in the variety preceding this normalization, and in picking from that variety the form which best suited the needs of this project.

The CoNLL-X files from which I would construct my sentence representations each had a very clear root, the core subject identified by our dependency parser. We had trained for that root to be an object worth extracting and so, a rooted graph made the most sense. I knew that I needed the nodes to be represented by tokens, and more specifically by the relationship of each token to its root node. My grammatical comparisons would be based on these relationships, the adjective moderators and the counts which qualify the objects. By storing these relationships in each individual node they would be most easily accessible to me for comparison. Finally, I determined that it would not make sense to have cycles within the graph, even though some words, like the “of” in our current example, really should reference backwards, following those cycles would make my later plan to recurse over each layer of the graph more difficult and so I resolved to handle such cases in post-processing. Each of these decisions led to the development of DAGs which look like the below, for the same slate-blue tunic example we investigated above, with each layer of the graph representing a new clause and a further set of either attributes or an additional object and attribute. You will

³⁸ Ibid

note that I’ve removed extraneous aspects of each sentence, like the punctuation and the largely procedural ”next,” *advmod* at the start of each one.

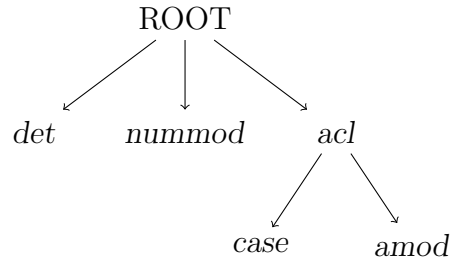


Figure 2: DAG representation with relationships as nodes, as implemented in the algorithm

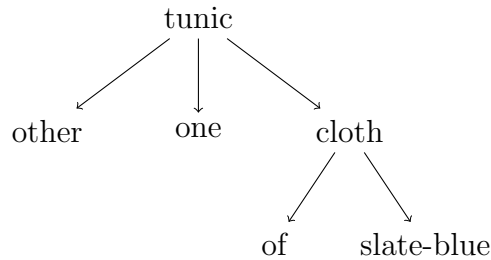


Figure 3: DAG representation with individual translated tokens as nodes, for clarity

With these graphical representations carefully constructed and in hand, the final step was to do the actual extraction of the objects and attributes for each sentence. As mentioned above, the goal of parsing these sentences has always been to connect each item and its description to the AAT, a hierarchical vocabulary where something like “water pitcher” is associated with **vessels for serving and consuming food**, **culinary containers**, and eventually categories as general as **containers (receptacles)** and **Furnishings and Equipment**.

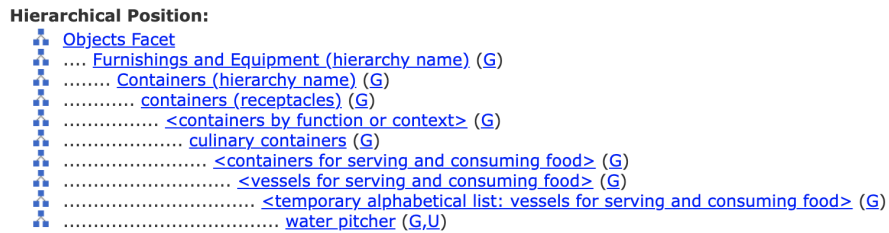


Figure 4: Example of Getty Art and Architecture Thesaurus hierarchical structure for “water pitcher”

If we knew that the water pitcher was, for example, made ‘of pewter,’ that would be a subcategory unto itself, enriching the picture of the pitcher which we’re able to glean from the Getty connection. Sorting in this way will help scholars who are interested in either pewter objects or culinary containers to easily conjure everything we know about this specific water pitcher, allowing for multiple avenues of research access.

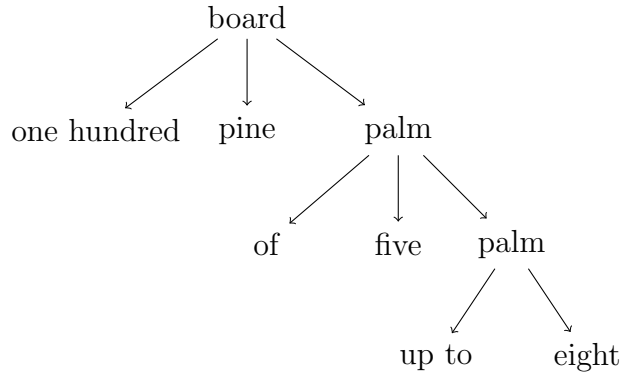
Thinking of the objects this way further allowed me to develop the methodology for parsing and extracting the information from each sentence’s graph, using recursion to pattern match within each individual sub-clause and attributing the extra details to their associated object accordingly. Recursion and Directed Acyclic graphs are famous friends, and since I knew that I wanted to represent each object from the bottom up, checking as I went to which original root a description applied and connecting the two accordingly, recursing over each level of the graph and tracing back to the root was the most thorough means of doing so while also ensuring that I didn’t drop any clauses or produce any incoherent extractions along the way.

Looking at the DAGs above, it is easy to visualize how this process functions. The recursive algorithm is first passed the index of the root node, the DAG, and the dictionary representation of the parsed original sentence. In cases where there is a secondary clause present, represented in most of our sentences by an ‘acl’ or a ‘conj’ relation, it recurses, passing itself the index of the secondary root – the attribute root, and then passing onto the next layer of description. Once it reaches a level with no further additions it compares to my set of grammatical rules based on the relations present in that level and prints with the postprocessing mentioned above such that ‘of’ comes before the descriptors in an intuitive fashion for us to read, and in a hierarchical manner reflective of the AAT’s functionality. The result is something like the below, for our tunic sentence, with the DAG representation first, with word indexes in each level, and then what was extracted from the

sentence.

```
4:  [1, 2, 3, 7, 8], 7:  [5, 6]
tunic :  ['one', 'other'] ['of', 'slate-blue', 'cloth']
```

Or for our hundred pieces of wood of varied widths, the dag and then the extraction:



```
3:  [1, 2, 4, 7, 11], 7:  [5, 6, 10], 10:  [8, 9]
board :  ['one hundred', 'pine'] palm :  ['of', 'five'] palm :  ['up
to', 'eight']
```

Each layer is still printed as its own separate list of attributes, with the root, the central object, at the front and the subsequent roots, the points within the sentence to which the later clauses refer, also interspersed with their relations on display. This mirrors the functionality of the AAT hierarchy to which these extracted clauses are then connected. We get, with this structure, that there are one hundred pine boards and that they range, from five palms up to eight palms in width. The two separate palm clauses are separated, in this case, because the secondary clauses are not often directly related in this way where they're two sides of the same scale, but they are easily merged in post-processing to create:

```
3:  [1, 2, 4, 7, 11], 7:  [5, 6, 10], 10:  [8, 9]
board :  ['one hundred', 'pine'] palm :  ['of', 'five', 'up to', 'eight']
```

in which all related words are grouped in the same subcategory. Using these outputs, we can stratify and apply methodologies to each sentence and the objects which they represent, meaning that we have met the goal of this project, to extract the objects and their attributes from each sentence and present them in such a way that they are easily both registered with the AAT and searchable for the researchers of the future. From this format semantic linkage is a simple matter of querying for the AAT's corresponding entry and filling in our additional clauses of details from there, effectively beginning DALME's searchable inventory database system with the Marseille corpus.

5 Conclusion

This project was, in many ways, an exploration. A first pass to see how far we could go towards illuminating this very specific corner of the medieval world using modern methods. As a result, there is of course much farther to go, and lots to consider surrounding the limitations of these types of interdisciplinary attempts.

In terms of next steps, the DALME collections contain far more languages and dialects than we were able to directly investigate during this thesis. The process of retraining existing models, stanza or otherwise, to produce dependency parsings of these inventories will necessarily be unique for each language brought under the purview of DALME’s push towards automation. As discussed, we are hopeful that much less manual annotation will be necessary for these future languages, but under the system presented in this thesis, the process will still require knowledgeable supervision and human decision making, not to mention the development of a gold standard for these inventories, necessitating the involvement of language and dialect specialists.

Once the dependency parses are acquired, however, my extraction process should remain relevant. While the languages vary, the structure of these inventories remains largely consistent. Across Europe, household items were being logged and recorded in a very similar manner. As a result, the sentences recorded – their object focus, their lack of verb phrases – are widely applicable, and so my recursive algorithm inclined towards this structure should persist.

There are, of course, improvements to be made to this part of the process as well, largely surrounding the singular focus of the recursive algorithm on working for the sentence fragments which we find in droves within these inventories. As discussed, these sentences are almost always rooted in a noun or noun phrase, with later clauses either adding an additional object to the list or further describing the root. Almost is an operative term here, though, and there are some sentences which do not follow this structure, or any one currently accommodated

by the conditionals in my recursive call. These capabilities could be expanded in the future, potentially involving some pre-processing to split sentences into two initial groups: traditional or partial (inventory) sentences, and then sending them through two separate recursive processes accordingly.

When considering the future work necessary for this project I can't help but reflect on the work which was necessary just to get started, to reach a point where it felt as if I was accomplishing anything substantial or helping Professor Pizzorno at all. I spent most of the fall semester wading hopelessly through Universal Dependency results. With no formal linguistics, never mind Latin linguistics training, it was unclear to me for a long time how I would ever understand enough of the actual content to even think about improving its accessibility through quantitative means. I understood Natural Language Processing methods and theories in the abstract, but I also was certainly not intimately familiar enough with them to know where and how I could break the rules, which buttons were ripe for pushing, able to be tweaked towards my desired outputs. As I grew more comfortable and knowledgeable in both fields, I felt them as increasingly disparate. Specific needs of the inventory corpora had no comparable vocabulary in NLP literature. Existing OIE algorithms were trained to be as broad as possible, trending them away from being something I could even think of incorporating into my processes.

It was frustrating, but perfectly logical, that these disciplines should be so innately opposed. Computer Science tends towards the global, it wants to serve the highest volume of content in the quickest way possible. History is often concerned with minutiae, specializing within specializations until we end up focused on household inventories written in medieval Provençal. There is power, certainly, in the faster, big-picture algorithms dominating computational research today. But there is power, too, in harnessing this capability for something very localized and particular, in forcing the giant Machine Learning model to bend and consider the dialectical leanings of a tired, 15th century notary.

Is it a useful application of time, of computational resources, the dialectical leanings of a 15th century notary, the objects he inked into the historical record? I think so. It's fun; the cross-century, cross-cultural awareness of the things which made up these people's lives. I have coats with hoods in my closet, boards of varying widths in my garage. So did they. Is not the humanist argument for constant improvements in tech the ability to bring us all closer, to increase awareness of the similarities across our great big world?

But if fun, and recorded material culture, and the time-space continuum of *stuff* is not sufficiently convincing to you, then the ramifications of this problem to contemporary analogs should be. There are just as many modern languages and dialects equally unserved by Natural Language Processing methods as Provençal, languages in which today's citizens are writing social media posts, letters, literature, and for which there are no accurate existing models. The capabilities of computation lean so heavily on the languages of those with the resources and education to create them. As we develop, from our places of privilege, we further solidify those languages and their specificities as a general norm, biasing the world of computational linguistics towards ourselves.

The tedious correcting and re-correcting necessary to train the model and develop the methodology presented in this thesis would be necessary ten-fold in order to account for the multitudes of applying Natural Language Processing tools to an entire under represented language or dialect. Not only did I have a very specific language, but also a very limited data set and an exact goal: get the objects and their attributes. Machine Learning and NLP techniques are used to monitor and assess language on the web constantly, from the monitoring hate speech on social media websites to processing emergency requests. There is a pressing need, in this day where people increasingly turn to the internet for both assistance and malintentioned disinformation, to accurately handle this constant influx of text and to fairly account for the people behind it. There is no way to

do this without the ability to better bridge computation and linguistics, without the development of methods which are applicable based on broad truths about language and then have the capacity to be highly specialized for all of the niches which make our speech and writing interesting, and for all of the languages in which we communicate.

This project is an example of the kind of work currently necessary in order to attempt to bridge these sorts of topics towards mutually increased understanding. A record of the research and roadblocks and requisite, constant future improvement demanded by the progress of our contemporary society and the wheedling, fascinating aggravations worth uncovering in our past. It is also, I think, an example of the research and roadblocks necessary to make this same understanding possible in our near future, to better comprehend the lives of our own contemporaries, not through what they own - although still fascinating! - but through what they have to say.

References

- Atmani, M., & Lafourcade, M. (2021). Universal Dependencies for Multilingual Open Information Extraction. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, & B. Heinisch (Eds.), *3rd conference on language, data and knowledge (ldk 2021)* (24:1–24:15). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/OASIcs.LDK.2021.24>
- Keywords: Natural Language Processing, Information Extraction, Machine Learning
- Bayer, M., Kaufhold, M., Buchhold, B., Keller, M., Dallmeyer, J., & Reuter, C. (2021). Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *CoRR*, *abs/2103.14453*. <https://arxiv.org/abs/2103.14453>
- Bishop, C. (2014). The ‘pear of anguish’: Truth, torture and dark medievalism. *International Journal of Cultural Studies*, *17*(6), 591–602. <https://doi.org/10.1177/1367877914528531>
- Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *ArXiv:1805.04032 [Cs]*.
- Corro, L., & Gemulla, R. (2013). Clausie: Clause-based open information extraction. *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 355–366. <https://doi.org/10.1145/2488388.2488420>
- Dönicke, T., Yu, X., & Kuhn, J. (2020). ‘identifying and handling cross-treebank inconsistencies in ud: A pilot study. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 67–75.
- Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705. <https://doi.org/10.18653/v1/P19-1166>

- Etzioni, O., Banko, M., & Cafarella, M. J. (2006). Machine reading. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, 1517–1519.
- Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1535–1545. <https://aclanthology.org/D11-1142>
- Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2012). Dependency-based open information extraction. *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 10–18. <https://aclanthology.org/W12-0702>
- Keuhn, T. (2008). Estate inventories as legal instruments of credit in renaissance italy. *Heirs, kin, and creditors in renaissance florence*. Cambridge University Press.
- Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. <http://arxiv.org/abs/1301.3781>
- Miletic, A., Bras, M., Vergez-Couret, M., Esher, L., Poujade, C., & Sibille, J. (2020). A four-dialect treebank for Occitan: Building process and parsing experiments. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 140–149. <https://aclanthology.org/2020.vardial-1.13>
- Murray, J. M. (1993). The profession of notary public in medieval flanders. *The Legal History Review*, 61, 3–31.
- Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking.
- Pizzorno, G. (forthcoming). Bootstrapping machine learning models for low-resourced, highly specific language corpora: A case study from late-medieval marseille. *Digital Humanities Quarterly*.

- Smail, D., Pizzorno, G., & Morreale, L. (2020a). Methodology: A Documentary Archaeology. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/methodology/>
- Smail, D., Pizzorno, G., & Morreale, L. (2020b). Overview. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/overview/>
- Smail, D., Pizzorno, G., & Morreale, L. (2021). Workflow. *The Documentary Archaeology of Late Medieval Europe*. <http://dalme.org/project/workflow/>
- Smail, D. L. (2022). Inventory of Guillelmus de Cavallione. In D. Smail, G. Pizzorno, & L. Morreale (Eds.), *The Documentary Archaeology of Late Medieval Europe*. <https://purl.dalme.org/55a1cca4-8d71-4fb0-abb8-4ae0b350acaa/>
- Yin, Z., & Shen, Y. (2018). On the dimensionality of word embedding. *CoRR*, *abs/1812.04224*. <http://arxiv.org/abs/1812.04224>
- Žabokrtský, Z., Zeman, D., & Ševčíková, M. (2020). Sentence meaning representations across languages: What can we learn from existing frameworks? *Computational Linguistics*, 46(3), 605–665. https://doi.org/10.1162/coli_a_00385