The Bodin Corpus: A multilingual parallel text case-study

Peter Nadel

May 29, 2022

Abstract

This project investigates the efficacy and limitations of providing linguistic and morpho-syntactic annotations to multilingual texts. I use the the French, Latin and English versions of Jean Bodin's *Six Bookes of the Commonweale* as a case study. Employing a range of natural language processing tools, each sentence is matched with a sentence from the other languages, and then annotations are added to each word in that sentence. Word-level alignment is attempted, but the hurdles for this task on the Bodin corpus and at large are explained. Each word is also paired with a section in a reference grammar, which elucidates the annotations attached to it. While common in some well-resourced languages, these morpho-syntactic annotations are just beginning to be explored for under-resourced languages like Latin, and so this project also seeks to explore differences between these two groups how they manifest themselves in practice.

1 Introduction

As classicists, understanding the interaction between a source language and a translation is paramount in examining the motivations and rationales behind a translator and how that source text is interpreted for generations to come. To see behind this veil of translation, classicists devote themselves to years of studying not just ancient languages but also common methods of how to translate from these languages into their mother tongues. With this fluency achieved, classicists seek to engage with source texts as they were written, slowly relying less and less on a translation and more and more on their own understanding of the ancient language itself. While this model has been prolific in generating novel interpretations of classical texts, it neglects the way that most people, those who are not classicists, interact with material written in a dead language. Indeed, in order to access the works of Aeschylus or of Juvenal, readers must resort to canonical English translations, most of which, as any classicist would attest, make use of antiquated or anachronistic language to render difficult or else untranslatable sections of the original. However, this process of translation is not the subject of this project. Instead this project seeks to synthesize state of the art computational tools and standard close-reading techniques to show readers where and how source texts deviate from their translations. To that end, a platform, through word-level morpho-syntactic and alignment annotations, to easily visualize the similarities and differences between source and translation texts would offer general readers with an opportunity to interact with a language they do not have any experience with, non-classicist researchers with a greater ability to rigorously analyze a text in translation, and classicists with new method for interrogating fundamental questions concerning translation.

The French, English and Latin editions of Jean Bodin's *Six Bookes of the Commonweale*, herein all together referred to as the Bodin Corpus, provide a unique case study to humanists and data scientists exploring these questions. As I will explore, the peculiar publication history of this work, as well as its foundational role in the development of Enlightenment political science, make it a useful case-study in investigating how complicated, multilingual texts can be made more accessible for students and researchers, regardless of their linguistic background.

I build on the efforts of many projects with similar goals, but where these seek to create a platform generally for downstream use, I will approach the Bodin corpus as a case-study of how to construct the digital edition of a multilingual text. Nevertheless, comparison between my objectives and these similar projects is productive. First, the Ugarit Translation Alignment Editor (http://ugarit.ialigner.com/), developed by Tariq Yousef at the University of Leipzig in 2016, is an incredibly flexible manual word-level alignment platform, intended for use in research and classrooms. Palladino, Foradi and Yousef [4] present a pilot study of Ugarit and its use in Classics courses at Tufts and Furman Universities. For

both graduate and undergraduate students, classroom assignments took the form of manually aligning original text in either Latin or Ancient Greek to its canonical English translation, sometimes multiple English translations. In this way, as I mentioned above, manual alignment sought to formally integrate the work of past translators (work that classics students necessarily take advantage of anyways) into course work. As the study tells us:

> The recognition of the limited capabilities of translation to efficiently convey concepts in Classical languages, while at first caused disappointment and criticism, convinced the students of the need to look more in depth into the original language to understand its inherent characteristics and the way it expressed ideas.

This observation is key and stands at the foundation of my work on the Bodin corpus. As we will see, the translations between the different books are not as simple as modern translations of canonical Greek and Latin texts, themselves not simple. Instead, they depend on each other for formulating meaning and are rarely direct translations. Thus, researchers and students of the Bodin corpus must confront the conceptual porousness of translation as a task before they can begin to understand how each language version contributes to historical meaning.

This introduction, thus, behooves us to further explore the history of the Bodin corpus and how the unique facts of its production make multilingual text annotation difficult.

2 The History of Bodin's *Six Bookes*

Because it will have significant bearing on the results of this project, I will sketch a brief history of Bodin's Six Bookes of the Commonweale in order to show how my digital edition converses and interacts with analogue ones.

2.1 French to Latin (1576-1586)

As Kenneth Douglas McRae explains in his introduction to the Harvard Political Classics 1962 edition of the 1606 English translation of Bodin's treatise: "Serious study of the République is complicated by the existence of two versions of the text, both written by Bodin, published ten years apart, and differing substantially one from the other" [3]. He continues, "Rather than a translation, the Latin version represents a complete redrafting and a rewriting of the French text" (Ibid, 28). In order to grasp what he means by this contention, consider the very first sentence of each version:

> French: "REPUBLIQUE est un droit gouvernement de plusieurs mesnages, et de ce qui leur est commun, avec puissance souveraine." [2]
>
> Latin: "Respublica est familiarum rerúmque inter ipsas communium summa potestate ac ratione moderata multitudo [1]

Although these two sentences mean nearly the same thing, the Latin adds the phrase, "ac ratione moderata," "and with a managed system of reckoning," to its definition of a Republic. It is important to remember that Bodin's France was, throughout much of his life, embroiled in domestic and foreign wars of religion, with Catholic and Huguenot parties rising to and falling from power with the result that in this environment Bodin was either forced to check the true opinions that he held about the government and later represent them in the Latin version, or else as a result of ten years of disappointment in the French court, reform his beliefs in the writing of the Latin version. Whatever the cause of this shift may be, it is clear that by the time of the Latin version's publication, Bodin no longer believed that "puissance souveraine" alone could sustain a Republic, but that it required a rational hand to steer it in the right direction [3]. As McRae points out, "Much had happened to Bodin between the writing of the two versions. He had married and fathered a family... He had participated in [Prince Francois d'] Alençon's ill-starred venture in Flanders... and [had been] disillusioned with King Henry III" (Ibid, 29). Too, no longer was he serving that king, and instead found himself in an "unsought retirement" (Ibid). His days as a rising advocate in the political discussion circles of the duke d'Alençon were well behind him, along with the delicate fabric of court administration. Despite being freer in his ability to comment on the functioning of the government, his power and ability to affect change was equally reduced. Thus, where the French original would only reference monarchy in

an attempt to sway powerful readers of whom Bodin had the ear, the Latin mentions other forms of governments like aristocracies and democracies, showing that Bodin tailored his reworked treatise for his broadening audience and that no longer did his conclusions simply form a French discourse on the role of government, but rather were a part of a wider European conversation. As McRae posits, "All of these changes... were undertaken in order to produce a general study of politics which would be applicable to all countries" (Ibid, 32).

2.2 Knolles's Translation (1606)

Although Bodin died in 1596, this conversation did not cease, and instead his reworked Latin edition achieved popularity on the continent and in England. However, questions of translation again reared their heads. While much of Europe was content with the Latin edition, translations appeared in Spanish, Italian, German and English. Unlike the first two which are direct translations of the French original, the German and the 1606 Richard Knolles's translation attempt to blend the two versions into a single consistent work, with McRae again commenting, "[Knolles's] translation stands as a work of independent judgment. No translator today could properly adopt this procedure..." (Ibid, 38). To better understand this observation, compare the first sentence of the English translation to the French and Latin versions:

> "A Commonweale is a lawfull government of many families, and of that which unto them in common belongeth, with a puissant soveraigntie." [3]

The syntax of this translation copies that of the French exactly, including the hypotactic style of the subordinated clause, "et de ce qui" and shunning the paratactic coordination found in the Latin noun phrase, "rerúmque inter ipsas communium." But Knolles is able to subtly weave features peculiar to the Latin into his English semantics. The genitive implication of ownership, not present in the French, appears in the English with the verb "belongeth." Too, he shifts his translation of the French "droit," usually when an adjective translated as *right* or *upstanding* to one reflecting the noun's meaning, "lawfull," to compensate for the missing ablative phrase, "ac ratione moderata." [3] With these differences in mind, it is difficult to grasp Bodin's *Commonweale* in its entirety. With work spanning languages and contexts, scholars are attracted to two main areas: (1) how Bodin's political thought developed in his life time, achieved by comparing the French and Latin versions; and (2) how younger contemporaries of Bodin, like Richard Knolles, interpreted the differences they saw between the two versions and eventually how their interpretations gained currency in the development of seventeenth century political thought in England and elsewhere. Interested in investigating how research in these fields can be facilitated by state-of-the-art natural language processing systems, we can now turn to my methods and approach in confronting the several obstacles at the foundation of studying the *Commonweale*.

3 Methods

3.1 Alignment Techniques

For the purposes of natural language processing, bilingual text alignment can be formally defined as, given a sequence of source words in language X of length m, $x =< x_1, x_2 \dots x_m >$ and a corresponding sequence of target words in language Y of length n, $y =< y_1, y_2 \dots y_n >$, an algorithm that finds pairs of source and target words such that $A = < x_i, y_i >: x_i \in x, y_i \in y$, where x_i and y_i are semantically or syntactically similar within in the context of their unaligned sentences [6]. AwesomeAlign uses contextualized word embeddings from a language model to generate word representations in the form of continuous vectors. This model can be trained from any source, but for this project and as it is suggested by AwesomeAlign's creators Zi-Yi Dou, Graham Neubig, I will be using and modifying the multilingual Bidirectional Encoder Representations from Transformers (BERT) model found on Hugging Face [5]. This model is trained on the top 104 languages by size on Wikipedia, using masked language model (MLM) and next sentence prediction (NSP) as objectives in model training (Ibid). MLM, given a particular sentence in the training set, randomly masks fifteen percent of the words and passes it into the prediction model. Unlike traditional recursive neural networks which usually see one word after another and thus learn word sequence, this method allows for bidirectional learning and

is less biased towards particular syntax found in the training set. Similarly, NSP joins two masked sentences in the pretraining and text preparation stage and according to Hugging Face's documentation "Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not." Both of these methods allow for a incredibly flexible yet lightweight language model that can, as we will see, be fine-tuned to particular use-cases [5].

With this BERT model, we can begin to leverage AwesomeAlign. First the BERT model extracts a contextualized word embedding for each word in the word sequences x and y, such that: $h_x =< h_x1, h_x2 \ldots h_xm >$ and $h_y =< h_y1, h_y2 \ldots h_yn >$. Taking these contextualized word embeddings, AwesomeAlign's aligner uses two methods for determining unidirectional word pairs, probability thresholding and optimal transport. After this step, bidirectional alignments are extracted from the intersection of the two unidirectional matrices. In probability thresholding, a similarity matrix is computed from the dot product of h_x and h_y. Then, as in many machine learning tasks, the values in this matrix are converted from contextualized word embeddings (i.e. vector representations) into values on a probability simplex (i.e. values between -1 and 1), using either the softmax or α-entmax normalization functions (more about the pros and cons of each of these methods can be found in the original AwesomeAlign paper [6]). If a word pair exceeds a certain threshold, those words are said to be aligned. For optimal transport, the cosine distances between each of the values in h_x and h_y is computed and scaled between 0 and 1 using min-max normalization. As in probability thresholding, if this value passes a certain threshold, the two words are said to be aligned. AwesomeAlign then returns a list of paired word indices sorted by confidence of alignment [6].

As described above, the Bodin corpus offers significant challenges and obstacles to this simple framework. First, as we have seen, AwesomeAlign operates under one significant assumption: that the word sequences x and y are complete translations of each other. The complicated history of Bodin's treatise means that many sentence pairs do not satisfy this assumption. Even the notion of a sentence pair, a concept which underlies AwesomeAlign's implementation, is thrown into question. Consider the following examples. First, these three sentences, in a list of all sentences in each language, appear at the same index position (i = 8):

> French: "C'est pourquoy ils ne doivent jouyr du droit de guerre commun à tous peuples, ny se prevaloir des loix que les vainqueurs donnent aux vaincus." [2];
>
> English: "And therefore in all wise and well ordered Commonweales, whether question be of the publike faith for the more safetie to bee given; of leagues offensive or defensive to bee made; of warre to bee denounced, or undertaken, either for the defending of the frontiers of the kingdom, or for the composing of the controversies and differences of Princes amongst themselves; robbers and pirats are still excluded from all the benefit of the law of Armes." [3];
>
> Latin: "Primum Rempublicam diximus ratione moderatam esse oportere: quia Reipublicæ nomen sanctum est, vt ab ea latronum ac piratarum cætus arceantur, quibuscum nulla contractuum fides, nulla iuris societas esse debet, sed summa distractio." [1]

To anyone knowledgeable in French, English and Latin, these sentences do not say the same thing, yet AwesomeAlign has no way to know that, that is to know whether or not two sentences are in fact translations of each other. While it may be quite simple for a human to do so, this task remains difficult, sometimes impossible, for a computer. Thus, I needed to employ some method by which sentences and not just words could be aligned. Otherwise, most of the text could not be parsed by AwesomeAlign. There are three methods currently available for sentence level alignments, all improvements on the last. The first was described by William Gale and Kenneth Church in their 1993 paper "A Program for Aligning Sentences in a Bilingual Corpora." The Gale-Church algorithm that this paper popularized works by comparing lengths of sentences and matching similarly sized sentences by means of a statistical model [7]. This method is rough and approximate, as it is often used to align many thousands of sentences, in which, due to the Law of Large Numbers and the Central Limit Theorem, each misaligned sentence does not affect the average number of correctly aligned sentences so much so that the bilingual corpus is rendered useless. Indeed the Gale-Church algorithm is still used to align large bilingual corpora for machine learning tasks such as BERT. However, the second method of sentence alignment, Rico Sennrich's Bleualign, has, since its publication in 2010, become

much more popular for this task [12]. Bleualign begins with a standard implementation of the Gale-Church algorithm, but on top of it calculates a BLEU score derived from a machine translated version of the original text. In fact, this BLEU score uses the same calculation as Gale-Church, but because it is drawn from a machine translation of the source text, it is more accurate than Gale-Church in approximating similar sentence length. The last method and the one which I decided on using, is called Vecalign and, like Bleualign, starts with Gale-Church as a basis. Similar to AwesomeAlign's implementation, Vecalign uses document embeddings similar to BERT to generate a statistical model of the bilingual corpus, aligning all those sentence pairs which pass a certain threshold. Although Vecalign is at the cutting edge of sentence alignment, the task for which it is most used is aligning large bilingual corpora for machine learning tasks and not aligning sentences for annotation and presentation [14].

Even though Vecalign can give us a concordance of aligned sentences, AwesomeAlign still has difficulty aligning at the word-level two sentences with similar, but not the same content. Take again the first line of the first chapter in each language (i = 0):

> French: "REPUBLIQUE est un droit gouvernement de plusieurs mesnages, et de ce qui leur est commun, avec puissance souveraine." [2];
>
> English "A Commonweale is a lawfull government of many families, and of that which unto them in common belongeth, with a puissant soveraigntie." [3];
>
> Latin: "Respublica est familiarum rerúmque inter ipsas communium summa potestate ac ratione moderata multitudo." [1]

While each sentence above conveys a similar point, there are marked differences that AwesomeAlign fails to pick up on. Between the French and English, the words themselves are quite a similar and AwesomeAlign picks up on these similarities, aligning every word in the sentence correctly, with the exception of the French word "est" (its second usage in the sentence) and the English word "belongeth." Although both are the main verb of their subordinate clauses and in this context have a similar semantic meaning, AwesomeAlign's similarity score does not meet the threshold for alignment. Lowering this threshold may help in aligning this pair, but would likely let it more false positives than true positives.

This phenomenon is even better seen in the French to Latin alignment. The Latin sentence, despite expressing similar ideas as the French, is clearly not a translation of the French (for reasons why refer to the History of The Commonweale section). And indeed AwesomeAlign is only able to match the words which would jump out as aligned to a human reader of the sentences, like Republique and Respublica, commun and communium. Indeed in these cases, AwesomeAlign, when paired with a sentence aligner like Vecalign, can give us insight into how different two sentences are, as opposed to how similar they are. In this simple example, we can see how Bodin's role as translator is radically different than that role might imply. And while that insight may be obvious from reading alone, AwesomeAlign provides us with metrics to see just how different each Latin word is from its French counterpart.

3.2 Stanza annotations and web development with Django

Along with the alignment annotations described above, I also set out to provide morpho-syntactic annotations on each word, as well as display all of these annotations in a simple interface built off of a database-driven back-end.

Stanza from the Stanford NLP Group provides efficient tools for linguistic analysis of a wide variety of languages [11]. As opposed to Explosion AI's spaCy [8], Stanza provides a built-in Latin language model trained on the Universal Dependency Latin IT (Index Thomisticus) treebanks. This database, taken from the Index Thomisticus, a corpus of all of the works of Thomas Aquinas as well as texts from 61 other authors which was lemmatized and morphologically tagged by father Roberto Busa SJ and others started in the 1940s, gives this Stanza model a vast source from which to parse Latin texts [10]. Alongside this model, I used the English Web Treebank (EWT) model and the French GSD model, which are the two largest treebanks for each language. When a document, in this case a single sentence from any of the Bodin versions, is passed into its language model, the model returns each word as a Stanza Word object, a dictionary-like object which has built into it the word's text, its lemma, its part of speech (in two different schemas), its morphological features, the index position of the word which it modifies, known as the 'head' and the Universal Dependency relation tag associated with that modification, know as 'deprel.' Because each language shares the same vocabulary of tags, direct comparison between words in different languages is facilitated, especially for people who are only

Measure	Score
Precision	.476
Recall	.562
F1	.516

Table 1: Vecalign Precision, Recall, F1 scores.

practiced in only one or two of the languages (Ibid). As another benefit of using Stanza over spaCy, this Word object can be simply appended to the JSON representation of each the word and thus can be easily be parsed when entered into the SQLite database running on the website. Each word then is recorded in this database under its language class and brings with it a series of associated information which then can be displayed on the front-end fo the website. I felt that it was critical to parallel comparison to not just be able to quickly access morpho-syntactic data but have it in a standardized form for which explanations and examples can be found in the documentations of each treebank. Too, with this tag inventory, those who cannot speak a language can begin to appreciate the structure and not just the content of the material. For this reason, too, I included grammar references that can give the reader more insight into how and why a particular form was used, going beyond what Universal Dependency tags can provide and giving the reader a more profound way to interact and engage with a text in a language they may not know or are learning.

4 Evaluation

4.1 Alignment Techniques

In both sentence-level and word-level alignment tasks, there are few comparative case studies that seek to present and annotate multilingual corpuses, let alone those that attempt to do so for texts which use antiquated and historical language, as we see in the Bodin corpus. Indeed, as I mentioned above and will develop below, these tools are mostly used to generate, enhance and evaluate downstream machine translation tasks. The difference between this aim and my own for the Bodin corpus will be explained in the context of evaluating the efficacy of these alignment tools.

4.1.1 Vecalign

Sentence-level alignment, using Vecalign, proved to be scalable for the Bodin corpus, but still fell short of expectations. As stated above and in Brian Thompson and Philipp Koehn's 2019 paper introducing Vecalign, "One of the primary applications of sentence alignment is creating bitext for training MT systems" [14]. And indeed it is in this capacity that Thompson and Koehn evaluate Vecalign. Attempting to re-align "noisy, web-crawled data in two low-resource language pairs: Sinhala–English and Nepali–English," Thompson and Koehn use Vecalign to recreate downstream machine translation systems (Ibid). Realize that this task is quite a different one that what I used Vecalign for in the case of the Bodin corpus. In fact, the first step of this process is filtering out unsatisfactory sentence pairs, whether this arises out of wrong language sentences or a low Vecalign score between sentences with high token overlap. Although dropping sentence pairs that do not align may be justifiable, indeed the expectation, in downstream machine translation tasks, for the proposes of my work on the Bodin corpus, this best practice would have resulted in nearly half of the first book being dropped (See Table 1).

Thus, through a complicated manipulation of the console output of Vecalign, I needed to devise a system by which unaligned lines in either language could be accounted for and reintroduced into the stream of aligned sentences. For, as I discussed in this history section, many lines were not supposed to be aligned to any line the compliment language, a feature which no current alignment algorithm can account for and which makes evaluation in the standard sense, designed for machine translation tasks, quite difficult. This process outputs two collections of sentence indices, one aligned and one unaligned, which are used directly in the text output through Django queries on the SQLite database running the backend of the website. This fundamental distinction between aligned and unaligned sentences does not exist for downstream machine translation tasks, or else it does, but only for intermediary evaluation against a gold standard (i.e. human generated) of sentence pairs, yet for corpus presentation and

Model and package	Precision	Recall	F1 Score
AwesomeAlign with mBERT	.759	.061	.112
AwesomeAlign with Historical mBERT	.710	.060	.110
AwesomeAlign with Montaigne Corpus	.810	.063	.117
SimAlgin	.755	.065	.121

Table 2: Word-level scores.

annotation, a task, which, excepting this work, tends to be over looked in favor of more immediately impressive machine translation results, knowing which lines are unaligned and which are aligned is the first hurdle. Vecalign provides a useful inventory of utilities, yet it is clear given the manipulation necessary to access the indices of unaligned lines that this task of presentation and annotation was not considered in its design. All of that said, Vecalign produces better results when compared to a human aligned gold standard than both BLEUalign and the standard Gale-Church implementation found in BLEUalign [14]. And indeed the evaluation values in Table 1 (measured by F1-score) are somewhat misleading. Although we see just over fifty percent of lines are correctly aligned by Vecalign, this value includes lines not meant to be aligned as false negatives for the purpose of F1-score calculations. Importantly, I did not seek to remove these from the evaluation as the goal of the project was to see how well off-the-shelf tools could function on the Bodin corpus, with as little intervention on my part as possible.

4.1.2 AwesomeAlign

AwesomeAlign provided very poor results for automatic word-level alignment between the two language pairs (see Table 2).

Although these F-1 score results are rather low, especially compared to the numbers reported in the original AwesomeAlign paper, we see that AwesomeAlign has comparative results to the other popular word-level aligner Simalign. AwesomeAlign also allows the user to substitute the standard multilingual BERT model, described above, for other models. For the purposes of evaluation, I included two others: the multilingual Historical BERT model [13], and a multilingual BERT model [5], fine-tuned on a human aligned corpus of Montaigne's Essays. This first model was produced in late 2021 by the Digital Library team at the Bavarian State Library and incorporates corpuses of historical English, French, German, Finnish and Swedish, the contents of which can be explored on this model's page on the HuggingFace model zoo. Obviously, it would be an oversimplification to categorize the linguistic traditions of these languages to historical and not, and the use-case for his model is not completely clear, especially because it does not enhance the word-level alignment quality of AwesomeAlign by any significant degree. The second model takes a human aligned corpus of sentence pairs from Montaigne's *Essays*, and uses it to fine-tune the standard multilingual BERT model. Montaigne penned his *Essays* over the course of the 1570s to the 1590s and thus the spelling and syntax of his French resembles that of Bodin's. Too it was aligned to the 1603 English edition translated by John Florio, which is similar in spelling and syntax to Robert Knolles's version of the Bodin corpus. Thus, this model attempts to account for the variation and the lack of standardization in orthography and style that the neither modern nor the historical multilingual BERT models are able to do. Even still we see that this model only preforms slightly better than the modern model, although its accuracy is significantly higher than the other models. This observation leads to a general conclusion concerning word-level alignment of the Bodin corpus, namely that accuracy scores for the the task are quite high, hovering between seventy to eighty percent while precision sits between five and six percent.

These values tell us that while AwesomeAlign is able to correctly identify aligned words, it is only able to do so for a very small minority of all words in the text. In other words, although the rate of false positives is low, the rate of false negative is much higher, almost by a factor of ten. This issue arises from the one-to-one approach that I alluded to in the methods section. Because AwesomeAlign for the purposes of calculating its cosine similarity score, understands each word by itself and does not take into account that words position in the syntactic tree, auxiliary and helping verbs, that is words with functional and not semantic meaning, are either dropped or misaligned. Consider an example apart from the confusing linguistic tradition of the Bodin corpus (see Figure 1): aligning the the first article of the Universal Declaration of Human Rights, a document written to be translated into as many languages as possible, we see near perfect alignment, even with the functional marker of the passive

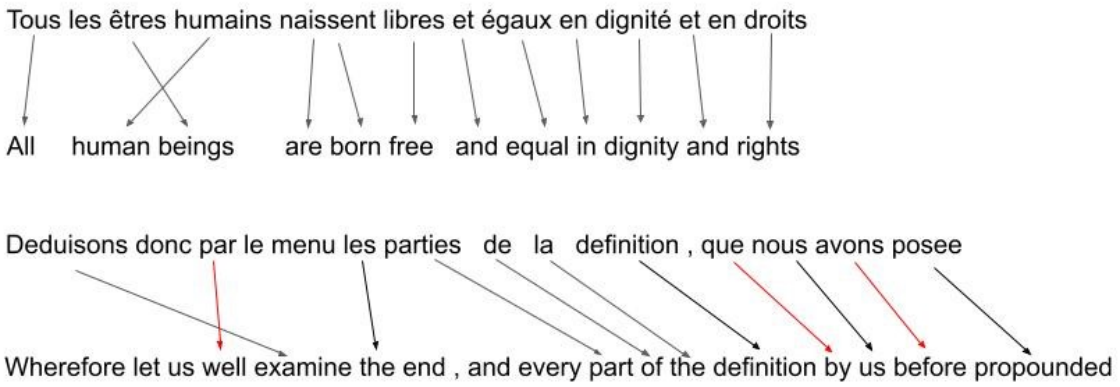

Figure 1: Alignment of the first article of the UN Declaration of Human Rights [9] compared to the alignment of sentence seven of the French and English versions of the Bodin corpus [2] [3]

voice in English ("are") being correctly aligned to the corresponding French word ("naissent") [9]. Compare this to the results taken from sentence seven of Bodin: for the moment putting aside semantic differences in Knolles's translation, key words like "Deduisons" and "posee" are correctly aligned with their English counterparts ("examine" and "proposed," respectively), yet English functional markers, in the first case, of the first person, plural, hortatory usage of the subjunctive, "let us" and, in the second, of the first person, plural, imperfect "have" are dropped and would, for the purposes of evaluation, be considered false negatives. Even when the French uses a similar construction, as in the case of the imperfect with "avons", we see that AwesomeAlign does not just drop this word, but misaligns it to the word just preceding the past participle "posee," in the English "before." Interestingly, if there were no adverb "before," which is not a necessary part of the sentence and not found in the French, alignment of the functional word "avons" to the word preceding the correctly aligned participles ("posee" and "proposed") would give back the correct the alignment. However, Knolles did not use the phrase "have proposed," opting to drop the unneeded auxiliary verb. Perhaps this shows an under developed learning of syntactic structures in the multilingual BERT model, especially when encountering texts that are closely related but not direct translations, as in the case of this sentence and much of Bodin. Whatever the cause of these inaccuracies, the irresolution of functional words, especially in verbal forms, combined with the difficulty of aligning non-direct translations, leads to the low recall score for AwesomeAlign on the Bodin corpus.

4.2 Morpho-syntactic annotations

Unlike the alignment techniques described above, the morpho-syntactic annotations that were added as attributes to each word proved successful both because the task itself was much more suited to the technology available and because that technology was much more flexible in usage. That said, the parsing of each chapter, through the Stanza language model I mentioned in the Methods section, was resource intensive with processing times of about an hour to an hour and half per chapter. The quality of the morpho-syntactic parsing was consistently accurate, especially in determining the case and number of nouns and the tense, mood, number, voice and person of verbs. Significant errors only occurred where optical character recognition (OCR) was unable to correctly parse Bodin's words. This problem was most evident in the Latin version of the text, the published edition of which consists of images of the original typesetting from the 1586 edition. Although these images are sometimes difficult for a human to read, errors only come at the ends of lines where a dash is used to join a word that is continued on to the next line. OCR issues also accounted for alignment inaccuracies.

While the quality of the parsing was successful, the morpho-syntactic tags themselves, while perhaps useful to experts, would raise questions for language-learners and general users alike. Especially for the Latin language model, the output of this parsing is difficult to understand without a guide or correspondence between the reported categories and grammatical terms commonly used in language classes. As touched on before, Stanza language models generate a lemma, two types of part of speech tagging, the universal part of speech (UPOS) and the treebank-specific part of speech (XPOS), mor-

phological features known as feats, and a name for the dependency relation that that word participates in with its head word in the tree, called deprel. [10] This deprel tagset especially, but also the XPOS categories, do not use the same vernacular that a student would find in a classroom. Thus, I did some work to build correspondence between these tags and natural language attributes that might be useful to someone not trained in the specific tags set used by the Stanza language models. This tagset is, often but not exclusively, borrowed from the Universal Dependencies project that I mentioned before and if it seeks to gain currency in language learning and multilingual environments like the Bodin corpus, much more work will need to be done in order to facilitate student and researcher engagement. Whether this shared set of tags is generated following Universal Dependencies standards, along the lines of traditional language classes, or a mixture of both does not change the fact that these tags must be shared across the different languages. Although these tags may introduce some ambiguity, what is most important here is that the vocabulary of tags is the same across each language, or else the transferability of grammatical knowledge between languages is impossible. I will come back to structural and systematic areas of future research but for now, it suffices to present some illustrative examples which will highlight the abilities and inabilities of this schema.

Deprel tags represent a branch in the treebank or, in more formal terms, the edge between two nodes, where each node is a word in the sentence. The deprel tag associated with each word is the edge or connection between that word and its head, the word which it modifies directly. The highest verbal form in each sentence, that is the verbal form which governs the independent clause without which the sentence would be incomplete, may only have one type of deprel tag: root. The head of this word is, in Stanza, always called [root], because it is the root of sentence and conceptual head of the whole sentence. As a result, verbs in independent clauses always receive the deprel: root. This is often times sufficient as it tells the reader that this word constitutes the so-called main verb of the sentence, but at other time it is not, or else must be enhanced in order to be useful. Take for example the seventh sentence of the Latin noted above in Figure 3. In this sentence, the highest verbal form is "exigamus," the first person, plural, imperfect, active, subjunctive form of the verb exigo, meaning to put forth or posit. As we see in the French and English, the subjunctive mood of this verb matters a lot to the meaning that Bodin intended for this sentence. Especially in Latin, independent usages of the subjunctive, like this, signify an important juncture in the text, generally where the author is appealing directly to the audience. The deprel tag for this word is correctly identified as 'root,' yet this category misses the value that Bodin puts on this word by choosing a syntax that places a verb in a subordinate mood as the highest verbal form of his independent clause. Indeed, in other treebanking systems, like the one created by Professor J. Matthew Harrington for his courses at Tufts, this independent usage of the subjunctive is included as an attribute in the dependency relation because the relationship between the imaginary head of the sentence and its main verb is distinctly different when that verb is subjunctive rather than indicative. This point may seem minor, but if the teaching of languages will revolve around these tags, as it does in Professor Harrington's classrooms, they must be as explicit in their meaning as possible.

Another place of confusion in these deprel tags is in the representation of copulative or periphrastic constructions. As we saw in the evaluation of the alignment techniques, language models have some trouble confronting these difficult syntactic structures. These structures are some of the most difficult to master in any language, so it no surprise that they are difficult to annotate. The fundamental problem is that although only a single verbal forms is being expressed, two word are employed. Each word, by the strict logic of the treebank, must have a head word, meaning it must modify some other word. Generally in the case of periphrastics, the auxiliary verbal form is tagged as AUX and the participle forming the periphrastic is tagged as the highest verbal form with the deprel tag of the clause it is in. For example, in sentence five, we see several protases of a conditional construction with the final one being: "si quæ optimus sagittarius facturus esset, ea ipsa fecerit." In this case, "esset" is tagged as AUX and "facturus" as "advcl," reflecting that the protasis is a type of adverbial clause. However, facturus is a future participle not a finite verb, instead esset is the finite verb which governs the clause. This complex structure introduces an ambiguity into the tree which language learners may find unintuitive and altogether different than the way this construction is usually discussed and taught in Latin classes, where esset facturus is just another form of the verb facio and not decomposable. All tagset have their limitations, but this feature seems difficult to express in any tagset, with even Professor Harrington's tags being forced into a similar, albeit opposite convention (giving esset precedence over facturus, which would likely be tagged as an adjective in predicate position relative to the subject).

It is for these places of confusion that I added reference grammars into the structure of the website. The reference grammars give an overview of the syntactic structure being referenced by the word selected and allow the user to navigate through potential explanations of what they see in the text without leaving the Bodin corpus. I will return to this feature when discussing future areas of research.

5 Conclusion and Future Research

As a case-study, the Bodin corpus shows us that multilingual parallel text platforms are not just feasible but scalable to large datasets. Too, I was also able to investigate the usability and value of morpho-syntactic annotations and providing in-line reference grammars to elucidate the usage of certain grammatical structures. Finally, the Bodin corpus, not just one language or language pair, can at last be found in a single repository.

That said, this project, as was the goal, openned many more questions than it answered. Obviously, sentence- and word-level alignment is a fundamental first step if multilingual text platforms of this nature will gain currency in research or teaching. As I alluded to above, there are two obstacles which co-constitute each other: first, AwesomeAlign's core algorithm uses cosine similarity and other synthetic scoring methods to determine similarity between two BERT encoded words, but does this method actually do a good job of discovering similar words?; and second, the multilingual BERT model may not have the resolution to identify similar syntactic structures across languages, especially if translations are less strict than AwesomeAlign expects. At present, language models alone like multilingual BERT and libraries that use it like AwesomeAlign do not have the ability to create word-level alignments for presentation and further annotation, but coupled with syntactic tree analysis AwesomeAlign and multilingual BERT may be able to achieve better results on non-direct translations, like we see in the Bodin corpus. Navigating the tree of each sentence was a process that I tried to avoid, except when accessing the morpho-syntactic annotations because it tends to bog down processing times, making testing and debugging a significant hurdle. Training a supervised learning model on how translators tend to translate certain syntactic structures may provide a way forward in simplifying AwesomeAlign's task. This preprocessing step would first require a human generated concordance between source deprel tags, their clauses and target tags and clauses, with a labeling system that could facilitate AwesomeAlign's implementation. Importantly, though, this method would not be generalizable and would depend on the corpus in question.

As for morpho-syntactic annotations, future research must be devoted to integration with existing knowledge. Already, I have begun to this effort, but the reference grammars that appear at the bottom on the page are just PDF file and thus not interactive or dynamic. The first step would be to fold the millions of treebank sentences in English, French and Latin into a dictionary platform which could provide statistic on word usage based on the frequency of that word in the treebank repositories. I have begun word on this type of platform for the Latin dictionary (Lewis and Short) and that work can be found in the form of a Streamlit app found here. Here, one can enter the lemmatized form of a verb and see the statistics of what syntactic dependencies that word occurs with and in paired with the actual dictionary entry. The goal is to give readers an intuition of what is common usage and what is not, something difficult to identify in one's mother tongue let alone a second or third language. Indeed, the morpho-syntactic annotations serve a similar purpose, to draw attention to syntactic and semantic values that are inconsistent across three languages. As I mentioned before, it is for this reason that the vocabulary between each language must be consistent. The same would be true of these enhanced dictionaries.

Generalizability is the ultimate objective of this research. One can imagine system which could take in a source document and as many target texts as a user desires and incorporate them together into a single document analysis screen. The technology to complete this goal, for the most part, already exist. However, all of these software packages, libraries and scripts are not designed to work with each other. Thus, the future of this work will not be to draw everything we know about natural language processing into one platform; instead, it will be to have the discretion of choosing flexible solutions to facilitate as many uses as possible.

References

[1] Jean Bodin. *Io. Bodini Andegavensis De republica libri sex*. Scholar Select. 1586. ISBN: 9781374080164.

[2] Jean Bodin. *Les Six Livre de la Republique*. URL: http://cts.perseids.org/read/pdlpsci/bodin/livrep/perseus-fre1/1.1.

[3] Jean Bodin. *The Six Bookes of a Commonweale*. Harvard University Press. 1962. ISBN: 9780674733145.

[4] Maryam Foradi Chiara Palladino and Tariq Yousef. "Translation Alignment for Historical Language Learning: a Case Study". In: *Digital Humanities Quarterly* 15.3 (2021).

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[6] Zi-Yi Dou and Graham Neubig. "Word Alignment by Fine-tuning Embeddings on Parallel Corpora". In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. 2021.

[7] William A. Gale and Kenneth W. Church. "A Program for Aligning Sentences in Bilingual Corpora". In: *Computational Linguistics* 19.1 (1993), pp. 75–102.

[8] Matthew Honnibal and Ines Montani. "spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear. 2020.

[9] United Nations. *Universal Declaration of Human Rights*. 1948. URL: https://www.un.org/fr/about-us/universal-declaration-of-human-rights%20and%20https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[10] Joakim Nivre et al. "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497.

[11] Peng Qi et al. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

[12] Rico Sennrich. "Iterative, MT-based sentence alignment of parallel texts." In: *Nordic Conference of Computational Linguistics* (2011).

[13] Bayerische Staatsbibliothek. *bert-base-historic-multilingual-cased*. URL: https://huggingface.co/dbmdz/bert-base-historic-multilingual-cased.

[14] Brian Thompson and Philipp Koehn. "Exploiting Sentence Order in Document Alignment". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5997–6007. DOI: 10.18653/v1/2020.emnlp-main.483. URL: https://aclanthology.org/2020.emnlp-main.483.