Towards the automatic detection of syntactic differences

Kroon, M.S.

CHAPTER 4

Detecting syntactic differences automatically using word alignment

Author contributions: MK, SB, JO and SvdP conceptualized the research; MK designed the algorithms, wrote the tools, analyzed the data, and wrote the paper; SB, JO and SvdP supervised and critically reviewed the research.

Abstract

The key question of this Chapter is whether extensive linguistic knowledge about a language can be leveraged in order to detect grammatical properties of a less well-described language and differences between the two languages. To this end, word alignment is used to map source language words to target language words with the aim of detecting syntactic features of the target language and differences between source and target language by semi-automatically analysing this mapping. Three tools are developed to detect syntactic properties and differences. The tools are evaluated on the language pair English-Hungarian. It is concluded that the tools can be used effectively to form many correct hypotheses on differences between the languages in several syntactic domains, though some room for improvement remains.

4.1 Introduction

In the previous Chapter the possibility of using the Minimum Description Length principle in the automatic detection of syntactic differences was invest-

igated. The key question of this Chapter is whether extensive linguistic knowledge about a language can be leveraged in order to detect morpho-syntactic features[1] of another, less well-described language and differences between the two languages. It is assumed in this research that knowledge about only the source language is available, while no knowledge about the language under investigation (the target language) is available and the utterances in a corpus are not enriched with grammatical information, reflecting an extreme case of investigating an under-resourced and under-researched language. By aligning the utterances in a parallel corpus on a word level, the knowledge about the source language can be analysed automatically and mapped onto the target language in order to arrive at conclusions about morpho-syntactic properties of the target language.

For the purpose of the detection of morpho-syntactic properties of the target language, as well as differences between it and the source language, a three-step process is proposed: Preprocessing, Attribute extraction and Discovering features; cf. Figure 4.1. For the last step, three distinct tools were developed: the Data Grouper for Attribute Exploration, the Generalization Tree Inducer, and the Affix-Attribute Associator.[2] Section 4.2 consists of an extensive description of the overall proposed method, as well as detailed descriptions of the workings of the developed tools.

The remainder of this Chapter consists of a description of the setup for the evaluation of the process and tools (Section 4.3), a detailed results section of said evaluation (Section 4.4), the discussion of the proposed method and its results (Section 4.5), and a concluding section (Section 4.6).

4.2 Method

The proposed approach assumes zero knowledge about the target language, while assuming the availability of linguistic knowledge about a different language, henceforth the source language, as well as the availability of natural language processing tools, such as parsers and taggers, for the source language. In order to be able to conclude anything about the morpho-syntactic nature of the target language or to be able to extract any morpho-syntactic differences between the source language and the target language, there must be a mapping between the two languages. In this approach this mapping is achieved by leveraging parallel data and using bitext word alignment. These alignments are combined with the linguistic annotation of the words in the source language, leading to suggestions for morpho-syntactic features of the target language for a linguist to investigate. In this section we describe the process of going from raw parallel text corpora to the extraction of morpho-syntactic features of the target language and differences between it and the source language.

[1] Recall that by morpho-syntactic features I mean all morphological and all syntactic properties of a language. This reading is used throughout the dissertation.

[2] The code is made available on `https://github.com/mskroon/DeSDA`

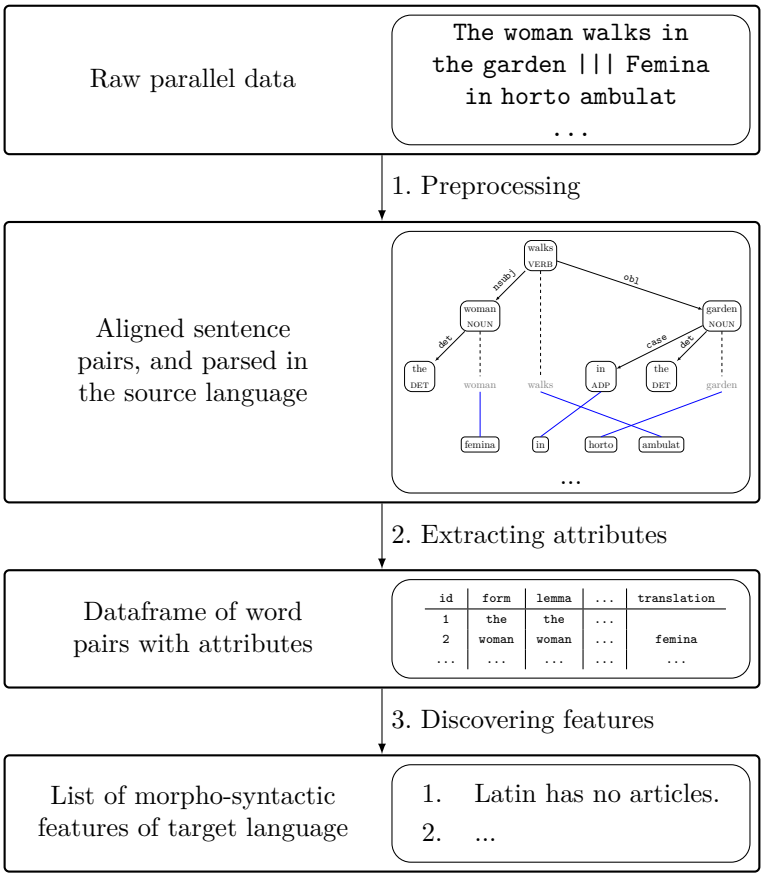

Figure 4.1: A list of morpho-syntactic features of the target language and differences between it and the source language is extracted from raw parallel data. The method consists of three steps: Preprocessing, Extracting attributes and Discovering features.

This process is divided into three steps, as illustrated in Figure 4.1. In the first step, Preprocessing, raw data is aligned on word level and, for the source language, parsed and tagged. In the second step word-internal and contextual morpho-syntactic attributes are extracted from the dependency parses to create a dataframe of words and attributes. In the third and last step three newly developed tools process the dataframe in order to detect morpho-syntactic features of the target language and differences between it and the source language. All of these steps will be described in detail below.

4.2.1 Preprocessing

First, the data of both the source and the target language need to be tokenized. In the current setup this is done with a language-independent tokenizer, that separates tokens based on whitespace, and splits punctuation symbols from tokens to treat them as separate tokens, and lower-cases words. However, a language-specific tokenizer may be more appropriate, depending on the language and the research goals.

Next, the parallel data are aligned on a word level. In principle any alignment algorithm or tool can be used – for the purpose of this research it was opted to deploy `eflomal` (Östling and Tiedemann 2016),[3] short for Efficient Low-Memory Aligner, a well-established statistical aligner that outperforms other popular statistical aligners such as `fast_align` (Dyer, Chahuneau and Smith 2013) and Giza++ (Och and Ney 2003) in both speed and alignment quality.

The task of word alignment can be defined as identifying which words in a translationally equivalent, parallel sentence pair correspond to each other. This is a notably hard problem, because it often involves word order differences, word omissions or insertions, and single words corresponding to multiple words, or a phrase. Due to this and a general danger for high computational complexity, there has been extensive research on the task (cf. Och and Ney 2003 and Tiedemann 2011, who give good overviews and descriptions of existing alignment models), in which three distinct families of approaches can be identified: heuristic, statistical and neural.

Heuristic models are the simplest, as they obtain word alignments through the 'similarity' between words of the two languages. One could think of applying the Dice coefficient (Dice 1945), which quantifies the similarity between two samples based on the intersection of the sample sets; in the task of word alignment, this straightforwardly constitutes the number of sentence pairs in which a word of the source language and a word of the target language occur together, relative to the total number of sentence pairs in which the words occur, whether alone or together. The higher this coefficient, the more often two words occur together, relatively, which indicates they may be each other's translations. While heuristic models are easy to implement and interpret, the problem with heuristic models is that the choice of similarity measure is arbitrary (Och and Ney 2003).

Statistical models, in comparison, have measures that are more soundly defined in probability theory, and often outperform simple heuristic models. They are distinguished by the fact that the alignments are the result of statistical estimation of a generative translation model that generates the target language sentence from the source language sentence using a set of latent alignment variables (Östling and Tiedemann 2016). The word alignments for the sentence pair (i.e. the latent variables) are then inferred from the generat-

[3] The source code and documentation of `eflomal` can be found at `https://github.com/robertostling/eflomal`

ive model, typically using a form of the expectation maximization algorithm (Dempster, Laird and Rubin 1977). The best alignments for the sentence pair are then those that return the highest probability of the source language sentence generating the target language sentence. The inference, however, can be done in multiple ways, and many extensions or adaptations to a model using the expectation maximization algorithm have been proposed, among which is Östling and Tiedemann (2016: `eflomal`), who use a Bayesian model with Markov Chain Monte Carlo inference.

Recent years have seen the rise of neural approaches in word alignment, specifically those using word embeddings to retrieve word alignments. An example of a recent neural aligner is SimAlign (Jalili Sabet et al. 2020), which uses the cosine similarity between the word vectors to obtain the word alignments, in a way reminiscent of existing heuristic approaches. Neural approaches such as SimAlign seem to outperform statistical approaches and have the advantage that the embeddings can be trained on non-parallel data. However, word embeddings are famous for requiring vast amounts of data (usually in the order of millions of sentences) to achieve good quality embeddings. Apart from that neural approaches are much more computationally demanding than statistical approaches.

The advantages and disadvantages of different approaches, then, leave neural approaches to be most effective for language pairs for which parallel data are not sufficiently abundant, while very large non-parallel corpora exist for both languages separately. Statistical aligners give good results – and are faster – for language pairs that do have sufficiently abundant parallel data. Therefore, seeing as the quantities of data large enough to train good quality word embeddings may not be available for most languages, especially those that may be of specific interest to comparative syntactic research, `eflomal` was used in this research, also considering that the existence of sufficient parallel data in order to extract syntactic differences was a prerequisite in the setup of this research.

After alignment, the data of the source language are parsed in Universal Dependencies (UD) (Nivre et al. 2016), with UDPipe (Straka and Straková 2017). Dependency parses are used, as opposed to constituency parses, because dependency parses directly and explicitly contain syntactic relations between words, which were considered to be essential for the purposes of this research. Having access to the syntactic relations between words allows the linguist to detect differences in the order of arguments, or the position of functional elements relative to their heads. In parsing, the UD programme's annotation conventions were followed, since it is one of the most widely used dependency-grammar programmes, but in practice any dependency programme could have been used.

UDPipe is a well-established dependency parser for UD, for which many pre-trained models are available. Easy to implement with binaries in many programming languages readily available, UDPipe achieves (near) state-of-the-art parses, however sentence parses are rarely completely perfect. Depending on the model used, the labelled attachment scores (a standard measure in dependency parsing that corresponds to the percentage of words that were

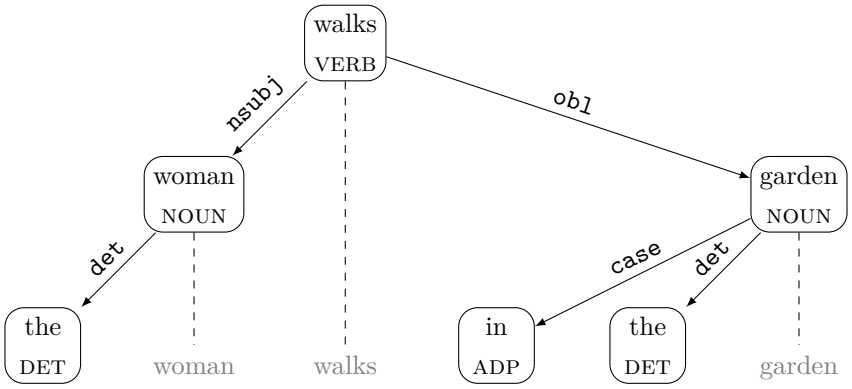

Figure 4.2: An example of an English sentence parsed in UD.

attached to the correct syntactic head with the correct syntactic relation or label) for English range between 82 to 86 per cent.

Additional to inducing a dependency tree with syntactic relations between words, UDPipe also lemmatizes, tags words for part-of-speech (POS) and provides morphological tags, which allows for the generalization over word categories and morpho-syntactic sub-categories. The accuracy of these POS tags and morphological tags range between 93 and 96 per cent for English models.

4.2.2 Extracting attributes

After tokenization, tagging and parsing, words of the source language have several attributes attached to them. These annotated words are passed on to the tools in the next step as rows in a dataframe; each row then contains a token with its attributes.[4] In this subsection all attributes in the dataframe will be described, some of which are deduced from context in the dependency parses or from the alignments.

As mentioned above, UDPipe parses sentences in UD, lemmatizes and tags words for POS and morphological tags. The result is formatted in CoNLL-U by default. The following relevant fields in the CoNLL-U output are taken up in the dataframe as columns, i.e. attributes:

`id:` contains the index of the token in the source sentence.

`form:` contains the form of the token in which it is encountered in the source sentence.

`lemma:` contains the lemma of the token.

[4] A dataframe is an efficient container object, effectively a table with labelled rows and columns. The dataframe is implemented in Python using pandas (Reback et al. 2021). The algorithms furthermore rely on `networkx` (Hagberg, Schult and Swart 2008) to efficiently process the dependency parses as graphs.

`pos:` contains the part-of-speech tag (POS tag) of the token in the context of the sentence.

`deprel:` contains the dependency relation between the token and its head. If it does not have a head, the deprel is `root`.

`feats:` contains morphological features of a token, such as singular number or third person. The complexity of these features varies from language to language.

Additionally, the attribute `translation` is added, which contains the word in the target language with which the source word was aligned. If a source word is aligned to multiple target words, all alignments are added, in which case the order of the target words is retained. For example, if the English preposition *around* is – correctly – aligned to the Dutch circumposition *om ... heen*, the `translation` field of *around* would be the list `[om, heen]`, and not `[heen, om]`.

As Kroon et al. (2020) already observed (i.e. Chapter 3), though, UD's categories (POS, morphological tags and syntactic relations) may be too coarse-grained to extract syntactic differences between languages with high precision. For example, verbs are not tagged for transitivity, but the transitivity of verbs is related to some specific morpho-syntactic differences between languages, chief among which is perhaps ergativity, in which the subject of an intransitive verb takes the same form as the object of a transitive verb, which is distinct from the subject of a transitive verb.

In order to detect differences with higher precision later, the UDPipe parses and tags are 'enriched' by adding some additional annotations that can be deduced from the trees. Among these enrichments verbs receive an additional tag for transitivity. Whenever a word in the source language is tagged as a verb, the algorithm automatically adds the sub-label `Trans` to the POS tag if it has a daughter node in the dependency tree with the dependency relation `obj` (used to denote the direct object relation between a nominal word and an active verb) or `nsubj:pass` (used to denote the subject relation between a nominal word and a passive verb).[5] Whenever a word is labelled as a verb but does not have any daughter node with one these relations, the algorithm automatically adds the sub-label `Intrans`. This is done so as to be able to distinguish between transitive and intransitive verbs in later stages.

Furthermore, for words that have the `conj` relation to their mother node in the dependency tree, denoting a conjunction relation, the dependency relation of their closest ancestor node that does not have the `conj` relation is percolated down and added as an additional relation (which in practice usually is their mother node's dependency relation, except for in nested summations). For example, in Genesis 1:1 (*"In the beginning God created the heavens and the earth."*), *the heaven* and *the earth* are conjoined. In UD *earth* receives the

[5] In cases such as *He was given a book*, the verb *given* also receives the sub-label `Trans`, because it has both an `obj` and an `nsubj:pass` daughter.

tag `conj`, being in a conjunction relation with *heaven*, while *heaven* is in an object relation to its mother node, *created*. The enrichment is then achieved by percolating the `obj` relation down, such that *earth* now has the relation `obj:conj`. A similar approach is deployed in Odijk et al. (2017), who describe that PaQu counts every conjunct in a subject conjunction as a subject of the verb, as well, a strategy they also deem reasonable and very useful. Percolating relations down opens up the possibility to distinguish between conjoined words, while still identifying their actual syntactic function. For the purposes of this research, it is mostly relevant for verbs, which give rise to a variety of syntactic differences between languages concerning conjunction – for instance, some languages may readily use participles instead of conjoining finite verbs, or may express specific instances of conjunction with a specific verbal form, such as the *te* form in Japanese.

On top of these CoNLL-U attributes and 'enrichments', a few more contextual and structural attributes that are of special interest in the detection of syntactic differences are derived from the trees: parents, children and crossings. These are explained below.

Parents and children

In order to connect possible morpho-syntactic differences to structural context, words receive two more attributes: one containing the POS tag and the dependency relation of its parent in the dependency parse; and one containing the POS tag and the dependency relation of all its children in the dependency parse. Having access to these structural contexts was deemed relevant, because knowing, for example, which personal pronouns are children of a verb, either as a subject or an object, can give a linguist all the necessary information to detect verbal paradigms or object agreement; or having direct access to a determiner's or adjective's parent's dependency relation can be telling in whether determiners or adjectives agree with their heads.

The algorithm distinguishes, however, between open and closed word categories when extracting the information of parents and children, which helps with the generalization over word classes while still retaining specificity regarding function words. Additional to the POS tag and the dependency relation, a parent or child's lemma is also extracted if its POS tag is a closed class. In this distinction, the algorithm follows the UD programme's line in their classification of open and closed word classes.[6] For illustration, consider the sentence *The woman walks in the garden* (see Figure 4.2); the word *garden*'s parent would be extracted as `VERB|root`, while its children would be extracted as `[in|ADP|case, the|DET|det]`. This allows the linguist to better distinguish

[6] The closed word class in Universal Dependencies contains the following POS tags: ADP (adpositions), AUX (auxiliaries and modals), CCONJ (coordinating conjunctions), DET (determiners, including articles and demonstratives), NUM (numerals), PART (particles), PRON (pronouns, whose subclassifications are encoded as features), and SCONJ (subordinating conjunctions).

between analytical and synthetic representations of grammatical features.

Crossings

With word order differences being a specific point of interest in comparative syntax, detecting crossing constituents or words between languages is desired, if not necessary. The alignments are therefore combined with the linguistic annotation of the source language in order to discover word order differences pertaining to specific morpho-syntactic attributes.

This is done by first checking whether there are crossings among the alignments within a sentence pair. Each alignment technically consists of a pair of indices (i, j), in which i refers to the ith word in the source language utterance (this i is identical to the `id` attribute) and j to the jth word in the target language utterance. For a pair of alignments (i, j) and (p, q), if $i < p$ (so, the ith word is on the left of the pth word in the source language), there is a crossing if $j > q$ (so, the jth word is on the right of the qth word in the target language). Similarly, there is also a crossing if $i > p$ and $j < q$. Note that if either $i = p$ or $j = q$ there is a many-to-one alignment; these cases are not considered to be crossings.

If a crossing is discovered for a pair of aligned words i and p, this is recorded for the words with `id` i and p. Each word in the source language is thus given a set of `id`s of other words in the same sentence, of which the alignments appear on the other side of the word's alignment in the target language.

For example, Figure 4.3 shows a sentence pair of English and Latin: *The woman walks in the garden* vs. *Femina in horto ambulat*. In this example English is taken as the source language, meaning that the sentence is parsed in UD, while Latin is taken as the unannotated target language. In the example, the following alignment pairs are found: $(2, 1)$, $(3, 4)$, $(4, 2)$ and $(6, 3)$. Of these alignments $(3, 4)$ and $(4, 2)$ cross, because $3 < 4$ and $4 > 2$. The alignments $(3, 4)$ and $(6, 3)$ also cross, because $3 < 6$ and $4 > 3$. Replacing the indices by the actual words, this in other words means that *walks* appears on the left of English *in*, while *ambulat* appears on the right of Latin *in*; and *walks* appears on the left of *garden*, while *ambulat* appears on the right of *horto*. It is then temporarily recorded for each English word whether it crosses and with what: the 3rd word (*walks*) crosses with the 4th word (*in*) and the 6th word (*garden*), the 4th word (*in*) crosses with the 3rd word (*walks*), and the 6th word (*garden*) crosses with the 3rd word (*walks*).

For each of these words, the shortest paths between it and the words with which it crosses are calculated, and are added as an attribute in the dataframe. In so doing each word (i.e. step) in the shortest path is retrieved as the dependency relation of the word to its mother, indexed with the depth of the step, which corresponds to the number of downward steps between the word and the start point of the path (upward steps are represented with a negative index). These paths are then sorted on the linear order of the words within the utterance. To illustrate this, let us consider Figure 4.3 again. *Walks* crosses

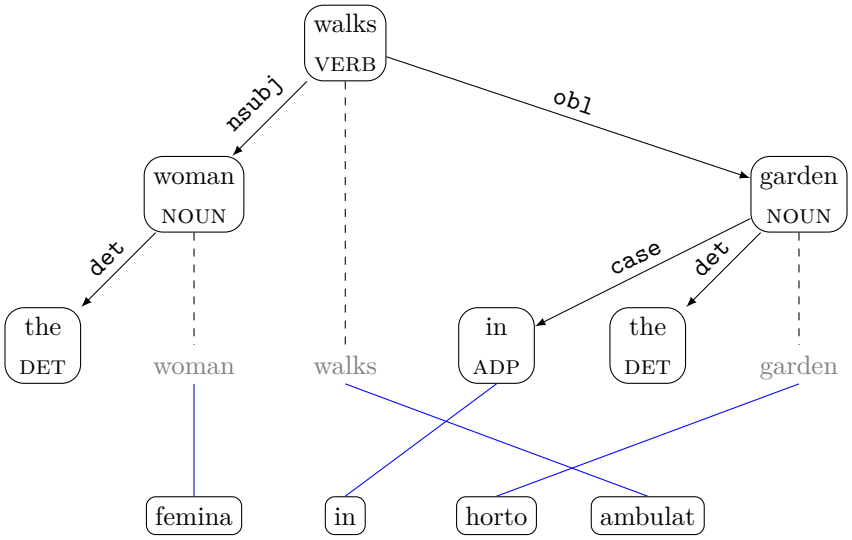

Figure 4.3: An English-Latin sentence pair. English acts as the source language, having full linguistic annotation, with a dependency tree in UD. Word alignments are indicated in blue. It can be seen that there are two crossings: $(walks, ambulat)$ crosses (in, in) and $(garden, horto)$.

with *in*. The shortest path from *walks* to *in* is *walks*—*garden*—*in*. Every word being retrieved as the dependency relation, this path becomes `root`—`obl`—`case`. Adding depth to this path renders it $\texttt{root}^0$—$\texttt{obl}^1$—$\texttt{case}^2$, meaning that from `root` it is one step down to `obl`, and two steps down to `case`. Finally sorting it on the linear order of the words in the utterance, would render it $\texttt{root}^0$—$\texttt{case}^2$—$\texttt{obl}^1$. So, as a `root`, *walks* crosses with a granddaughter `case` node, which is a daughter of an `obl`, and both of them are to its right. Crossing paths like these are useful: they show the linear order of the words and due to the depth indices the dependency structure is still retrievable.

For specific cases, however, a slightly different strategy is followed. First, in the case of crossings between nodes n_1 and n_2 that are siblings or where n_2 is a descendant[7] of a sibling of n_1, the shortest path is calculated up to the lowest common ancestor[8] in the tree and then down to n_2. For example a subject and an object swapping places, which are each other's siblings as they are both child nodes of the `root`, would render the crossing $\texttt{nsubj}^0$—$\texttt{root}^{-1}$—$\texttt{obj}^{1*}$, specifying the one step up from the subject to the root verb and then one step down from there again, indicated with the asterisk. It would also render the

[7] A descendant of node n in the dependency tree is any other node that is a child node of n or a descendant of a child node of n.

[8] An ancestor of node n in the dependency tree is any other node that is a parent node of n or an ancestor of a parent node of n. The lowest common ancestor is the ancestor node shared between two (or more) nodes that is lowest in the tree.

reverse, with the `obj` as starting point. This example could indicate that the sentence in the source language has an SVO order, while the target language, swapping subject and object, has an OVS order.[9] When this particular crossing is encountered often, one could hypothesize that the target language has the base order OVS in general.

Secondly, let us consider Figure 4.3 once more. Four crossings are found in total:

1. $\texttt{root}^{0}$—$\texttt{obl}^{1}$
2. $\texttt{root}^{0}$—$\texttt{case}^{2}$—$\texttt{obl}^{1}$
3. $\texttt{root}^{-1}$—$\texttt{obl}^{0}$
4. $\texttt{root}^{-2}$—$\texttt{case}^{0}$—$\texttt{obl}^{-1}$

Notice that 3. and 4. are the mirrors of 1. and 2., respectively. The only differences are the start and end points in the paths; in 1. and 2. one walks down from `root`, in 3. and 4. one walks up from `obl` and `case`. Now, 2. and 4. could be considered superfluous. Seeing as *walks* already crosses with *garden*, it can be expected that *walks* would cross with all of *garden*'s children nodes, too. In order to simplify the output, a crossing between word i and one of its descendants d is ignored if i also crosses with d's direct mother node in the tree. Inversely, a crossing between word i and one of its ancestors a is ignored if i also crosses with a's direct daughter node in the tree. As for crossings between word i and its siblings – or its, i.e. the sibling's, descendant – s, they are ignored if i also crosses with s's direct mother node, similar to crossings with ancestor nodes.

This reduction of the output was deemed reasonable. However a caveat: it is not necessarily a given that, if there is a crossing between two words, the words also cross with each other's children. Especially relevant in the case of extrapositions or any other form of discontinuity, consider a sentence pair such as English-Dutch *I saw a man who lives in Amsterdam : ik heb een man gezien die in Amsterdam woont*, in which the main verb (*saw : gezien*) appears to the left of the object in English, but to the right of it in Dutch. However, the relative clause (which is a daughter node of the object noun) appears to the verb's right in both languages. In this case, there are two crossings: *saw* crosses with the determiner *a* and the object noun *man*, of which the former is ignored because *a* is a child node of *man*, which also crosses with *saw*. The result is that only the crossing between *saw* and *man* is outputted, however this does not imply that there is a crossing between *saw* and all of *man*'s children: the relative clause does not cross. The output reduction therefore still retrieves the relevant crossings, but the discontinuity of the phrase and the interfering

[9] Not necessarily, however, since we do not know anything about the syntactic structure of the target language. It could be that the source language has an active sentence, while the target language has a passive sentence, in which case the order of the participants would be swapped, as well.

material are not highlighted.[10] A user must be aware of this behaviour, as it may cause for extrapositions to go unnoticed.

It similarly holds true that it is not necessarily a given that, if two words do not cross, there is also no crossing between them and any of the other's children. However, these cases do not cause any issue with the algorithm. Consider for example colloquial Russian *čto ona krasivuyu videla devušku* lit. 'that she (a) beautiful saw girl'. In this example the main verb interferes between the adjective and the object noun. Aligning this sentence with its English translation on word level shows that *saw* and *girl* do not cross, because the relative order of the Russian words to which they were aligned (*videla* and *devušku*) is the same: the object follows the verb. However, this does not imply that there is also no crossing between *saw* and any of *girl*'s descendants. In fact, *saw* crosses with *beautiful*, because *saw* appears to the left of *beautiful* while in the Russian sentence the order of the equivalent words (*videla* and *krasivuyu*) is reversed. No output reduction takes place, however, because *beautiful*'s mother node *girl* does not cross with *saw*, and the crossing between the verb and a daughter node of the object is correctly retrieved as $\texttt{ccomp}^0$—$\texttt{amod}^2$—$\texttt{obj}^1$.

As a final remark on crossings, a word's crossings are split into three categories: ancestor crossings, containing crossings with words that are its ancestors; descendant crossings, containing crossings with words that are its descendants; and sibling crossings, containing crossings with words that are its siblings or descendants of its siblings. This is necessary in order to be able to quickly distinguish between the types of crossings, and to see what kind of material a word crosses with.

4.2.3 Discovering features

After preprocessing the data and extracting attributes from the dependency parses, the dataframe is ready to be explored, and morpho-syntactic features of the target language and differences between it and the source language can be extracted from it. This is done with the help of three different tools that the author developed for this purpose: the Data Grouper for Attribute Exploration, the Generalization Tree Inducer and the Affix-Attribute Associator. The Data Grouper for Attribute Exploration, or DGAE, gives a breakdown of how often each morphological feature, crossing and translation (i.e. each meta-data attribute) occurs by *grouping key*. A grouping key can be any attribute or combination of attributes, such as POS tag, dependency relation or the combination of POS tag and dependency relation. These breakdowns can quickly provide insight in the prevalence of, e.g., determiners or adpositions in the source language that are not aligned to a word in the target language, indicating the absence of articles or the presence of cases (e.g. aligning a target language without articles to English, leaves the vast majority of articles to be

[10] In fact, even if the output reduction was not performed, the discontinuity of the phrase and the interfering material would not be highlighted.

unaligned, and quickly accessing the information that indeed very many articles are left unaligned, allows for the linguist to draw conclusions about the existence of articles in the target language). The Generalization Tree Inducer, or GTI, creates a tree based on the conditional entropy of attributes in order to better explore the co-occurrence of attributes. Though, whereas for a decision tree the most favourable split is the one that gives the highest information gain, which should lead to the correct classification as quickly as possible, this tool builds a tree by considering the most favourable split to be the one with the lowest information gain, which should lead to better generalization as opposed to identification. Finally, the Affix-Attribute Associator, or AAA, attempts to discover productive affixes in the target language and to relate them to morpho-syntactic attributes of words in the source language. All tools are explained in more detail below.

Data Grouper for Attribute Exploration

Simple yet insightful exploration of the data can already be done by means of a tool that gives attribute frequency breakdowns of the data or parts of the data. Splitting up the data, or rather grouping the observations, based on the value of an attribute can lead to the discovery of high co-occurrences between attributes. The attribute on which a split or grouping is based shall be referred to as a *grouping key*. As said above, a grouping key can be any attribute or combination of attributes, such as POS tag, dependency relation or the combination of POS tag and dependency relation, over which the data is partitioned. Patterns may arise when taking the dataframe and grouping all words by a specific attribute. For instance, grouping the data by POS tag should quickly show that pronouns are very likely not to be aligned to a word in the target language if the target language has pro-drop and the source language does not.

The author developed a tool that does exactly this: DGAE. While DGAE allows the user to group by any attribute or combination of attributes, grouping by POS tag, dependency relation or the combination of the two is probably the most useful in the case of discovering morpho-syntactic features of a language. For example, a (toy) dataframe such as the one in the top in Figure 4.4 can be grouped on the value of the dependency relation column, resulting in three smaller dataframes. In the middle, smaller dataframe a clear pattern can be observed: all `obj` nouns end in *-m* in Latin.[11] Per group, DGAE then gives a frequency breakdown of which attributes, including translations and crossings, occur with it, as shown for the `nsubj` group. It additionally gives the 20 most frequent attribute bundles in the group – i.e. which specific combinations of attributes occur most frequently – for better insight in the attribute distribution within the group, but this is not shown in the Figure.

[11] This is only the case because all words are singular in this toy example. Plural objects tend to end in *-s* or *-a* in Latin.

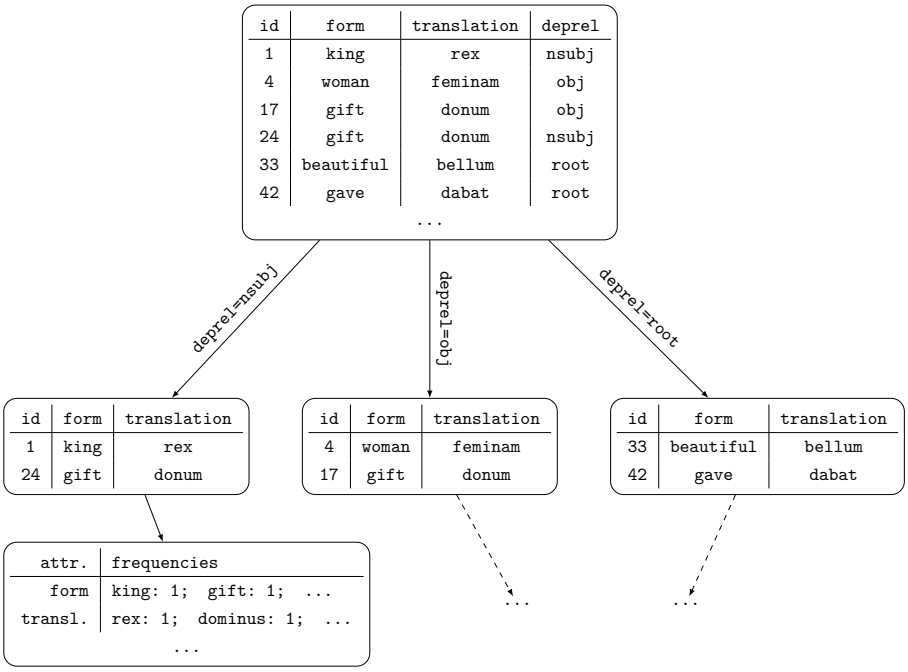

Figure 4.4: An example of grouping a dataframe by the value in the `deprel` column, short for dependency relation. A clear pattern emerges for the `obj`: all *translations* end in *-m*. DGAE then gives frequency breakdowns of attributes of the partitions, illustrated for the leftmost partition.

Some attributes can contain multiple values, such as the `feats` attribute that contain morphological features of a word as tagged by UDPipe. For these multi-valued attributes, DGAE does not count the frequency of unique feature bundles, but of the separate features instead. So, if the dataset consists of the two Latin words *anni*, the genitive singular – `(Gen, Sing)` –, and *annorum*, the genitive plural – `(Gen, Plur)` – of *annus* 'year', the frequency breakdown would record that `Gen` occurs twice, and `Sing` and `Plur` both once.

Generalization Tree Inducer

The author also developed GTI. The goal of GTI is to structure the data, in order to explore it in more detail and to be able to investigate whether certain attributes often co-occur. The data are grouped over the attributes in an attempt to generalize.

In machine learning, decision trees are a popular choice in classification tasks, where they predict the value of a target variable (such as the language in which a sentence was written in the case of language identification) on a set of observed features. They iteratively partition the data over the observed

features in order to arrive at groups in which as many items as possible have the same target value. GTI was built on this property of iteratively grouping and structuring the data, with two differences.

First, whereas decision trees canonically aim at partitioning the data based on the "most distinctive" feature, GTI aims at partitioning the data based on the least distinctive feature. This is done because the goal is not to obtain groups with a homogenous target variable, but to obtain groups with homogenous features.

In decision trees, the "distinctiveness" of a feature is usually described in terms of their influence on the *entropy* of the target variable in the partitionings. First introduced by Claude Shannon (Shannon 1948), entropy is an information theoretic term, and is often interpreted as the expected surprisal over an outcome of an event, or the amount of chaos in a system.[12] The higher the entropy, the less certain one is over the outcome of an event, meaning that there is much variation in the value of a variable. This "distinctiveness" of a feature, then, is the amount by which it reduces the entropy of the target variable – in other words, how much more it makes the outcome of the target variable homogenous.[13] GTI therefore partitions the data over the feature that reduces the entropy the least.

Secondly, in the task at hand, there is no formal target variable that needs to be predicted. In GTI, the role of target variable is therefore filled by a unique identifier for each observation (in casu, a token plus its attributes).[14] Effectively, the result is that GTI tries to group words into as large as possible groups.

The expected behaviour of this algorithm is then that it would detect "stable" features that show little variation. For instance, it can be expected that it would partition the data on POS tag very early. With the help of GTI, one can expect to find groups of words with many common features, which helps to structure the data.

However, to help the researcher explore the data more efficiently, GTI allows for the data to be pre-partitioned, for example by grouping words by POS tag. GTI is then run on each POS tag separately, which allows for the gener-

[12] The entropy H of variable X (with possible outcomes $x_1, ..., x_n$, which occur with probability $P(x_1), ..., P(x_n)$) is defined as

$$H(X) = \sum_{i=1}^{n} P(x_i) \log P(x_i)$$

[13] The amount by which it reduces the entropy is also known as *information gain*, which is defined as the difference between the entropy of a system and the entropy of a system given the outcome of another variable or the value of a feature:

$$IG(X, a) = H(X) - H(T|a)$$

[14] In fact, in this case the information gain of a feature is equal to its entropy. GTI therefore partitions the data based on the value of the attribute that has the lowest entropy.

alization over, for example, all nouns. This method produces a large file with nesting levels of indentation to represent the hierarchy in the generalization tree. Additional to the indentations, the file also contains the 20 most frequent feature bundles inside a partition (per indentation level), and it lists the (non-zero) entropies of the remaining attributes, for better exploration. Output can be suppressed to discard low-frequency data from which it is difficult to draw reliable conclusions, but this is optional and the parameters can be chosen by the user; the default settings do not print partitions that contain fewer than 5 observations or contain less than 1% of the observations of the partition one level higher (i.e. are a partition based on an attribute value that has a less than 1% probability).

It should be remarked that for multi-valued attributes, such as the `feats` attribute, the entropy calculated is the joint entropy of the technically multivariate distribution. That is to say, the entropy is calculated using the probabilities of the unique feature bundles, and not using the probabilities of the separate features. So, if the dataset consists of the two Latin words *anni*, the genitive singular – `(Gen, Sing)` –, and *annorum*, the genitive plural – `(Gen, Plur)` – of *annus* 'year', the entropy of the `feats` attribute would be 1, as calculated with the probability of `(Gen, Sing)`, 50%, and the probability of `(Gen, Plur)`, 50% – and not with the probability of `Gen` (50%), `Sing` (25%) and `Plur` (25%) separately. However, partitioning *is* done over the separate features. This allows for easier generalization over all singular nouns, for example.

Affix-Attribute Associator

Finally, the author developed AAA, that aims to generate hypotheses about which character sequences, or strings, could be affixes in the target language, and to associate them to morpho-syntactic attributes in the source language. It extracts all string pre- and suffixes (including full words) from the target language, without length restrictions, and all attribute subsets from the source language, in which it maintains a minimum frequency on both the strings and the attribute subsets in order to suppress the looming combinatorial explosion. The default minimum frequencies are 100: both strings and attribute subsets must occur at least 100 times in order to be included in the set of generated affix hypotheses. This minimum frequency is a parameter to be chosen by the user, though.

As for the attribute subsets, recall that some attributes can contain multiple values, such as the attribute `feats`. In extracting attribute subsets, AAA considers the words' full attribute bundles, in which the multi-value attributes have been flattened (i.e. the "brackets have been removed"). For example, Latin *anni* 'year' has `feats` attribute `(Gen, Sing)` as well as `lemma` *annus*. Its standard attribute bundle would be `[lemma=annus, feats=(Gen, Sing)]`, however flattening it would result in `[lemma=annus, feats=Gen, feats=Sing]`, in which all values are on the same level. From these full, flattened attribute bundles,

AAA extracts all non-empty subsets,[15] but only those that exceed the minimum frequency. This extraction is very prone to cause an exponential explosion, as the number of subsets is equal to $2^n - 1$, in which n is the number of attributes in the attribute bundle. Limiting this process is therefore very important. Imposing a minimum frequency on the attribute subsets (and therefore on the attributes themselves), as AAA does, already helps, but it is furthermore made sure that the algorithm does not extract subsets that contain the exact same observations. That is to say, if for example all words that are genitive happen to be singular as well, the algorithm will not extract both subsets `[feats=Gen]` and `[feats=Gen, feats=Sing]`, but only the latter. This drastically reduces the runtime, in practice. In the process of extracting attribute subsets, AAA furthermore ignores crossings and forms. Crossings, namely, tend to explode the number of subsets and are highly unlikely to be meaningfully associated to an affix in the target language; and forms (i.e. (inflected) forms in which words are encountered in the source language) have a strong tendency to associate to very long potential affixes, if not entire words, which does not benefit the desired generalization.[16]

AAA detects associated string-attribute subset pairs by means of *pmi*, or pointwise mutual information. Often used in natural language processing for finding collocations, pmi is an information theoretic measure of association, quantifying the amount of information learned about an outcome (e.g. it has rained) through observing the outcome of another random variable (e.g. the streets are wet).[17] To illustrate in terms of collocations, *Puerto* and *Rico* very often occur together in a corpus, which is reflected by a fairly high pmi between them. This means that the one word can fairly certainly be predicted when the other has been observed; when *Puerto* is observed, chances are very high that the next word is going to be *Rico*, and vice versa.

Pmi is calculated by

$$pmi(x;y) = \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

in which $p(x,y)$ is the joint probability of outcome x and outcome y occurring at the same time; $p(x)$ is the probability of outcome x; and $p(y)$ the probability of outcome y.

[15] In the case of *anni*, that would be: `[lemma=annus, feats=Gen, feats=Sing]`, `[lemma=annus, feats=Gen]`, `[lemma=annus, feats=Sing]`, `[feats=Gen, feats=Sing]`, `[lemma=annus]`, `[feats=Gen]`, and `[feats=Sing]`.

[16] Depending on the needs of the user, it can parametrically be specified what attributes need to be ignored or not. If it is so desired that crossings or forms are not ignored, they can be included in the attribute subsets.

[17] Pointwise mutual information is, then, the mutual information between two specific outcomes. One could compare this to the difference between self-information, which is about one outcome, and entropy, which is the expected self-information over all outcomes – pmi is about two specific outcomes, while the mutual information is the expected pmi. Mutual information is another name for information gain.

However, as can be seen from the formula, pmi is symmetric, that is to say, $pmi(x;y) = pmi(y;x)$. This is not ideal for our purposes, as some affixes may represent multiple distinct morpho-syntactic attribute subsets of a word; homomorphs. AAA therefore weights the pmi with the probability of the string conditioned by the attribute subset; that is, how likely it is that a string is encountered given that the attribute subset is known. This conditional probability is asymmetric. For each string-attribute subset pair AAA therefore calculates the following association value:[18]

$$A = P(string|attribute\ subset) \times pmi(string;\ attribute\ subset)$$

All string-attribute subset pairs are then sorted on this A, which is based on the probability of a word in the source language having attribute subset s and the word in the target language to which it was aligned having string, or potential affix, a. The higher A, the stronger the association between attribute subset s and potential affix a. It is then hypothesised that a may be an affix in the target language, associated to the attribute subset in the source language.

4.3 Evaluation

For the evaluation of the proposed method and developed tools, an experiment was run in which the researcher has linguistic knowledge of the source language, for which automatic parsers and taggers are available, while the researcher had no linguistic knowledge of the target language, in order to arrive at results as unbiased as possible. In order to gain insight into what kind of differences can be found with the tools, as well as what kind of differences cannot, the researcher compiled a list of morpho-syntactic hypotheses about features of the target language, and specifically differences between the source and the target language, based on the output of the tools. Meanwhile, a linguistic expert on the target language independently compiled a list of characteristic differences between the two languages that are prominent in the linguistic literature. These two lists were then cross-checked: which features that were found by the author were indeed correct features of target language's grammar; which hypotheses on features formed by the author were not correct; and which features were not found by the author that the expert listed as characteristic of the target language? These categories effectively correspond to true positives, false positives

[18] This is actually identical to a summand of the Kullback-Leibler Divergence between the probability distribution of the strings $Q(string)$ and the probability distribution of the strings conditioned by the attribute subset $P(string)$ in

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} p(x) \log_b \left(\frac{P(x)}{Q(x)} \right)$$

In this case, the Kullback-Leibler Divergence would be the information gain achieved if the conditional distribution P is used instead of the non-conditional distribution Q. This summand, then, represents the weighted part of the information gain for a specific string if the attribute subset is known.

and false negatives, respectively. True negatives cannot be considered, because they would correspond to features missed by the author that the expert had not listed as characteristic of the target language.

To this end, the language pair English-Hungarian was chosen, in which English served the role of the source language. Hungarian was chosen because the author had no linguistic knowledge of it. Dr. Lipták of Leiden University, who is a native speaker of Hungarian and a linguist specialized in Hungarian syntax, acted as the independent expert, and compiled the list of characteristic differences.

An English and a Hungarian Bible were used as corpora (see below). The English Bible was parsed and tagged in Universal Dependencies (Nivre et al. 2016) with UDPipe (Straka and Straková 2017), using the ParTUT model (Sanguinetti and Bosco 2015),[19] while the two Bibles were aligned on word level using `eflomal` (Östling and Tiedemann 2016), as discussed above.[20]

Limitations of this evaluation procedure will be discussed in Section 4.5.

4.3.1 Data

The corpus we use for the evaluation is, as mentioned, the Bible, specifically the English and the Hungarian Bible. In this Chapter it was opted not to use the Europarl corpus (Koehn 2005), in contrast to the previous two Chapters, because of specific shortcomings and complications inherent to the Europarl corpus that were noticed in the previous Chapters, such as the extensive presence of headlinese, a high average sentence length, some cases of code switching and untranslated utterances, and misaligned sentence pairs. The Bible is then convenient in that it is sufficiently large for many of our purposes, available in many languages, void of headlinese, monolingual, and implicitly parallellized through the way it is structured. However, the Bible is often archaic in its language and may not be representative of the way the language is spoken today. Hence, existing NLP tools and models may not be very suitable for Biblical language, and one must be very aware of the possibility of errors in parses and tagging. Nevertheless, we deemed the Bible to be a good choice, because of its ready availability and size, and because the only part of our approach that is dependent on language-specific models is the parser, which we did not expect to perform too poorly on the English Bible with the ParTUT model (Sanguinetti and Bosco 2015), because it was trained on texts that were collected from several legal and other formal texts, such as the Universal Declaration of Human Rights, the Europarl corpus and Wikipedia, which most closely resembled the formal, archaic, Biblical English of the available UD models.

[19] The model is available at `https://github.com/UniversalDependencies/UD_English-ParTUT`

[20] Given the nature of the task, it is impossible to measure the quality of the alignments a priori. We can therefore only report on the performance of `eflomal` on other languages in terms of alignment error rate (AER; Och and Ney 2003), which for closely related languages ranges between 7.6 and 10.6, while for less closely or unrelated languages ranges between 17.3 and 46.7 (Östling and Tiedemann 2016: Table 2).

In particular, we use the English and Hungarian parts of the Bible corpus by Christodoulopoulos and Steedman (2015),[21] which is a corpus consisting of over 100 different Bibles in `xml` format, annotated for book, chapter and verse, making alignment a straightforward task. The (standard) English Bible in the corpus is the King James Bible (KJB) from 1611, while the Hungarian Bible is the Vizsoly Bible (VB) from 1590, in a way the Hungarian equivalent of the KJB. Both Bibles are still widely considered the "classic" translation. Even though the age of the KJB would push the usefulness of the parser model to its limits, it also assured that it was, just like the VB, directly translated from Latin, Greek, Hebrew and Aramaic, allowing us to safely assume that the verses are, for the large part, syntactically comparable.[22] While a more recent version of the English Bible (the World English Bible; WEB) is included in the corpus, we chose not to use that, precisely for this reason: the WEB has been simplified more over the centuries than the KJB, diverging further from the syntactic structures in Latin, Greek, Hebrew and Aramaic and therefore in the VB.

	verses		words		
	original	shared	tokens	types	used for
EN	31102	28972	852606	12371	test & dev.
HU	31298		683690	61036	test
NL	29098		838324	21605	dev.
CS	31102		680938	39648	dev.

Table 4.2: The number of verses in the original Bibles, as well as the number of verses shared between the four versions. The number of words in the shared verses in terms of tokens and types is also listed per language. Hungarian is used for testing (i.e. running the experiment), Dutch and Czech only for the development of the tools, and English is used for both testing and development.

During development of the tools we used the Czech and the Dutch Bibles as well, and while the Czech version is from 1380 (so likely to be directly translated from Latin, Greek, Hebrew and Aramaic, too), the age of the Dutch Bible is not listed by Christodoulopoulos and Steedman (2015), however it seems to be from the late 1900s, probably 1987.[23] Because we wanted to make sure that the results are comparable between language pairs, we only used verses that

[21] The corpus is available on `https://github.com/christos-c/bible-corpus`

[22] Kroon et al.'s (2019) filter for syntactic comparability (see Chapter 2) was not deployed in this approach, because all filters that were developed build on existing NLP tools for both languages, except for the sentence-length filter. Seeing as the assumption that no linguistic knowledge or tools were available for the target language is an important aspect in this research, and that the sentence-length filter did not yield satisfying results, it was opted not to use a filter in this research.

[23] The age of the Dutch Bible corroborated our assumption about syntactic comparability between Bibles of similar ages. Although not quantified, the word alignments between English,

are present in all four versions of the Bible (i.e. only verses with IDs that were present in all four parts of the Bible corpus), which resulted in 28972 verses, with 852606 tokens in English and 683690 tokens in Hungarian. A full overview of the number of verses, tokens and types can be found in Table 4.2.

4.4 Results

After analysing the data and the output of the files, a list of 43 morpho-syntactic hypotheses about features of Hungarian was compiled by the author based on the output of the tools – a summary can be found in Table 4.3. Meanwhile, Dr. Lipták independently compiled a list of 32 differences between English and Hungarian that are characteristic of Hungarian and are prominent in the linguistic literature on Hungarian (henceforth the AL list). In this section we will discuss all of the (correct or wrong) hypotheses and morpho-syntactic features in detail.

In general it was observed that the majority of the hypotheses on morpho-syntactic features were correct, with 37 out of the 43 being a feature of Hungarian grammar. Two hypotheses were only half correct, painting an incomplete picture or overgeneralizing slightly. Another two raised further questions about Hungarian, about which Dr. Lipták was unsure whether they are or are not features of the Hungarian language. Only two hypotheses were actually incorrect.

Furthermore, out of the list of 32 prominent and characteristic differences on the AL list, eight were correctly discovered, while one hypothesis was contradicting a difference on the AL list. The rest –23– were missed, and are listed in Table 4.4. Each missed difference will be discussed in full in Subsections 4.4.1 to 4.4.5 below. In summary, these differences were mainly missed due to the information structure of a sentence not being annotated, or to the genre of our corpus.

In order to illustrate the process of forming hypotheses, we will begin this section by taking the hypotheses and missed differences concerning articles and demonstratives as an example, and discussing them in more detail, explaining our reasoning behind the interpretation of the data in Subsection 4.4.1.

We shall continue the section by discussing the remaining hypotheses and missed morpho-syntactic differences briefly, divided over three subsections: hypotheses and differences concerning the nominal domain (Subsection 4.4.2),

Hungarian and Czech were much better than the alignments between English and Dutch, reflected in the number of wrongly aligned words, implausible crossings and the number of unaligned words between English and Dutch. It is also reflected in the sizes of several output files, which were much larger for English-Dutch than for the other language pairs. This shows that the Dutch output can be summarized and compressed much less well; the entropy of the Dutch aligned data is much higher than for the other language pairs. We can probably conclude that the syntactic structures between the KJB and the Dutch Bible are therefore much less similar than between the KJB and VB, leading to wrong alignments, zero alignments, crossings and noisy data in general.

No.	Hypothesis	Correct?
	Nominal domain	
H1	articles	+
H2	articles come before NP	+
H3	articles do not inflect for case	+
H4	only definite articles	−
H5	licensing difference for articles	+
H6	nom. mods before and after NP	+
H7	case, expressed on noun	+
H8	locative cases	+
H9	accusative on nouns: -t	+
H10	accusative on pronouns: -t	+
H11	vowel harmony: front-back	+
H12	agglutinative	+
H13	no gender anywhere	+
H14	possessives optional: suffixes	+
H15	possessives prenominal	+
H16	3sg and 3pl same possessive	+
H17	adjectives both before and after noun	+/−
	Verbs and constituent order	
H18	free(r) word order	+
H19	transitive and intransitive verbs same position	+
H20	SV word order	+
H21	VO word order	+
H22	SVO word order	+
H23	relative order of constituents mostly same as EN	+
H24	adverbials mostly postverbal	+
H25	adv. clauses: same position EN	+
H26	adv. clauses: stricter word order than main clause	?
H27	pronouns positionally the same as nouns	+
H28	subject pro-drop	+
H29	verbs inflect for all persons: present	+
H30	verbs inflect for all persons: past	+
H31	much fewer auxiliary verb (temporal, aspectual, modal)	+
H32	synthetic passive	+
H33	no infinitival marker to	+
H34	adverbial negation	+
H35	negation comes before negated	+
H36	adverbs precede verb	+
H37	fewer pro-adverbs used	?
H38	zero copula	+/−
H39	copulae in general before predicate	+
	Other	
H40	both prepositions and postpositions	−
H41	adpositions declined for person	+
H42	coordinating conjunctions before conjunct	+
H43	subordinating conjunctions before conjunct	+

Table 4.3: A summary of the hypotheses formed about morpho-syntactic features in Hungarian, and whether they are correct or not. A plus sign means the hypothesis was correct; a minus that it was incorrect; and a question mark that the hypothesis has not yet been confirmed nor rejected.

No.	Difference
	Nominal domain
M1	demonstratives inflect for case
M2	numerals select singular noun
M3	demonstrative and article must co-occur
	Verbs and constituent order
M4	object pro-drop (singular)
M5	any number of constituents before verb
M6	wh-phrase: before finite verb in main clause
M7	wh-phrase: before finite verb in embedded clause
M8	wh-phrase: more than one before verb possible
M9	yes/no question: same word order as declarative
M10	embedded yes/no question: same word order + -e
M11	only N phrase before finite verb
M12	verbs agree with definiteness of object
M13	infinitive sometimes agree with subject
M14	singular agreement with subject with numeral
M15	verbal particle: can be before verb
M16	verbal particle: if after, free word order
M17	verbal particle: can be before auxiliary
M18	verbal particle: if before, can be reduplicated
M19	verbal particle: only before, if not only N phrase before
M20	verbal particle: idem, if not wh-phrase before
M21	verbal particle: idem, if not negation before
M22	verbal particle: idem, if not progressive aspect
	Other
M23	negative concord language

Table 4.4: A summary of the missed differences between Hungarian and English on the list compiled by Dr. Lipták.

those concerning verbs and constituent order (Subsection 4.4.3), and other hypotheses (Subsection 4.4.4).

Additionally, it was observed that our method was successful in detecting morpho-phonological features of Hungarian, including hypotheses about specific case endings, such as *-t* for accusative. Since automatic affix detection is an important goal in the field of comparative syntax, we will conclude the section by discussing all hypotheses and differences pertaining to affixes in subsection 4.4.5.

During the discussion of hypotheses and differences, we shall refer to them by their code as found in Tables 4.3 and 4.4 for easy reference.

4.4.1 Articles and demonstratives

In total five hypotheses about morpho-syntactic features of Hungarian pertain to articles and demonstratives (**H1**–**5**), four of which proved to be correct, while one was false. Meanwhile, one morpho-syntactic difference on the AL list pertains to demonstratives, which was missed (**M1**).

The hypothesis that Hungarian has articles (**H1**) turned out to be correct. To illustrate how this hypothesis was formed, Figure 4.5 shows (a fraction of) the output of the DGAE for words tagged as determiners (DET) in English – a tag that includes articles in UD – that have a `det` (determiner) relation to their head in the dependency tree. The output shows us that there are 80341 instances of such words in the English Bible. Under `form` it lists that 59116 of these words were an instance of the word *the*, and 7660 and 1582 were *a* and *an*, respectively, amounting to 9242 instances of `('DET', 'det')` having the lemma *a*. It also shows further breakdowns of features, such as that 68976 instances of `('DET', 'det')` were tagged as `'PronType=Art'` by UD – as having the pronoun type 'article'.

Importantly, under `translation`, it can be seen that 19500 instances of `('DET', 'det')` did not receive an alignment to a Hungarian word. Though a large number, it is significantly less than the number of words tagged as an article in English. This in turn means that if it were only articles that did not receive an alignment, then still $49476 (= 68976 - 19500)$ articles were aligned to a Hungarian word, amounting to at least 71.7% of all English articles having an alignment in Hungarian. This led to the correct hypothesis **H1**, that Hungarian has articles.

The DGAE output in Figure 4.5 also shows frequency breakdowns of crossings. Under ancestor crossings, it can be found that 40083 instances of `('DET', 'det')` do not cross with their ancestors, which suggests that the relative order of a determiner and its ancestors in English is the same as the relative order of the aligned-to Hungarian words.[24] Note that there can only be a crossing if a word has received an alignment: if it was not aligned, DGAE will

[24] Remember that a crossing between word i and its ancestor a are only considered – and taken up in DGAE's output – if i does not cross with a's direct daughter node. See Section 4.2.2.

('DET', 'det'): 80341	
form	('the', 59116), ('a', 7660), ('all', 3492), ('this', 1874), ('an', 1582), ...
lemma	('the', 59116), ('a', 9242), ('all', 3492), ('this', 2651), ('that', 1199), ...
pos	('DET', 80341)
deprel	('det', 80341)
feats	('PronType=Art', 68976), ('Definite=Def', 59345), ('Number=Sing', 13528), ('Definite=Ind', 9631), ('PronType=Dem', 4054), ...
translation	('a', 28600), (None, 19500), ('az', 17116), ('minden', 1926), ('e', 1309), ...
ancestor cross	('det^{0}', 40083), (None, 19500), ('det^{0}-nmod^{-1}', 2292), ('obl^{-2}-det^{0}-nmod^{-1}', 1889), ('nmod^{-2}-det^{0}-nmod^{-1}', 1480), ...
descendant cross	('det^{0}', 60833), (None, 19500)
sibling cross	('det^{0}', 50210), (None, 19500), ('case^{1*}-det^{0}-obl^{-1}', 1124), ('case^{1*}-det^{0}-nmod^{-1}', 594), ('det^{0}-obl^{-1}-nmod^{1*}', 356), ...
children	(None, 80264), ('that|fixed', 19), ('one|nummod', 13), ('be|cop', 7), ('of|case', 5), ...
parent	('NOUN|nmod', 19983), ('NOUN|obl', 18717), ('NOUN|obj', 12825), ('NOUN|nsubj', 9525), ('NOUN|root', 4178), ...

Figure 4.5: An example of the DGAE output. Displayed is a fraction of the results for English words tagged as determiners and that have a `det` relation to their head.

count that instance towards a `None` crossing. This allows us to quickly see that 65.9%($= \frac{40083}{80341-19500}$) of all aligned English determiners are on the same (i.e. left) side of their head as the Hungarian words they were aligned to. Similar to the reasoning that led to **H1**, we can see that if it were only articles that crossed with their ancestors, then still 48218($= 68976-(80341-40083-19500)$) articles showed up on the same side of their head as in English. This amounts to at least 69.9% of all Hungarian articles occurring on the left of their head, leading to hypothesis **H2** – Hungarian articles come before the NP – which turned out to be correct.

Furthermore, **H3** was formed based on the fact that there would be only four forms, which suggest a common stem *a-* under `translation`, of which only two are listed in Figure 4.5: *a*, *az*, *annak* and *ama*. The latter two, however, were much less frequent, with 361 and 224 occurrences, respectively, suggesting perhaps noise or another lemma. If articles were marked for case, it could be expected that there would be a higher entropy among the aligned-to Hungarian words, especially those that suggest a common stem. During further exploration of the GTI (output not shown) it was indeed noticed that there are only two Hungarian words clearly associated to definite articles (*a* and *az*) and only three Hungarian words clearly associated to indefinite articles (*a*, *az* and *egy*). Additionally, there was no pattern noticeable in the form of the article and the grammatical function (i.e. dependency relation) of the head of the English article, which one would expect if case is marked on articles. This led to the correct hypothesis that articles do not inflect for case. The fact that demonstratives, also tagged as `DET`, do inflect for case (**M1**)[25] can be found with the help of GTI, which shows that English demonstratives are aligned to a group of Hungarian words sharing a common stem, while the different endings show a noticeable correlation between the dependency relation of the determiner's parent node in English, suggesting that the determiners are inflected for case.

The output in Figure 4.5 shows that 9631 English determiners are tagged as having the `'Definite=Ind'` feature, meaning that they are indefinite, specifically indefinite articles. Notice that 9631 is more than the number of occurrences of the lemma *a*; this is because the label indefinite also includes the word *another*, which is not analysed as having the lemma *a*. However, with 9631 there is no clear candidate for a translation among the aligned Hungarian words; Hungarian aligned-to words either occur much more often, or much less often than indefinite articles in English. This immediately prompted further investigation with GTI, with which it was possible to observe that out of the 9631 English words tagged as an indefinite article, 4214 were aligned, constituting only 43.8%. Furthermore, as also mentioned above, there were only three Hungarian words clearly associated to indefinite articles: *a*, *az* and *egy*. However, we had already seen that *a* and *az* probably correspond to definite articles, and *egy* is rather infrequent with only 1146 occurrences, i.e. 11.9% of all English indefinite articles. It was therefore concluded that *egy* was either noise or the

[25] It turns out that *annak* is actually an inflected form of the demonstrative *az* 'that'.

cardinal number *one*, and that Hungarian only has definite articles *a* and *az* (**H4**). The fact that *a* and *az* are so often aligned to indefinite articles additionally led to the hypothesis that there is a usage difference between English and Hungarian articles (**H5**).

While **H4** proved to be incorrect with Hungarian having both definite (*a* and *az*) and indefinite articles (*egy*, which also serves the purpose of the cardinal number *one*), **H5** was correct: Hungarian does not use indefinite articles in existential and 'have' constructions and before predicate nouns, and indefinite articles can be dropped before subjects, objects or adverbials directly preceding the verb, while in all these cases an indefinite article must be present in English (Rounds 2009: 83). Although we believe that the data do point towards hypotheses **H4** and **H5** despite one of them having been proved incorrect, we also believe that the important conclusion of this showcase is that a linguist can start asking basic questions about characteristic morpho-syntactic features of a language – such as whether a language has both definite and indefinite articles – and explore the data with the help of our tools to form meaningful hypotheses on them.

For the sake of brevity, the remaining hypotheses and differences will be discussed in somewhat less detail. While **H1**–**5** served as an example to illustrate how the output of our tools are interpreted, the remaining hypotheses and missed differences will showcase the wide range of morpho-syntactic domains our tools can detect differences in.

4.4.2 Other hypotheses concerning the nominal domain

Additionally, 12 other hypotheses were formed pertaining to the nominal domain, 11 of which proved to be correct, while one was only half correct. Meanwhile, four morpho-syntactic differences on the AL list pertained to the nominal domain, of which two were discovered correctly, while the other two were missed.

It was observed in the DGAE output that an English word with a UD `nmod` relation to its heads appears without crossing it in 52.3% of the cases. It was specifically observed that in about half of the cases an `nmod` that occurs to the right of its head in English appears to the left of its head in Hungarian. Seeing as the `nmod` relation is used to denote the relationship between nominal dependents and another noun or noun phrase, corresponding functionally to an attribute (i.e. a nominal modifier; in the case of English, a prepositional complement) or a genitive complement, it was hypothesized that attributes and genitives can come both before and after their head in Hungarian (**H6**), although it is hard to identify from the output when it comes before and when it comes after its head.

It was correctly hypothesized that Hungarian has grammatical case, marked on the head noun (**H7**). This was most prominently suggested by the fact that 55.6% of all English prepositions did not have an alignment in Hungarian, as found with DGAE. This characteristic feature of Hungarian was also on the

AL list.

Further investigation in GTI showed that it was most frequently the prepositions *of*, *in*, *unto*, *to*, *with*, *from*, *upon*, *by*, *into*, *at* and *on* that did not receive an alignment, leading to hypothesis **H8**, about the presence of a genitive, inessive, dative,[26] allative, instrumental or sociative, elative or ablative, superessive, adessive, illative, and perhaps a temporal case in Hungarian – or at least something similar. Indeed, Hungarian does have all of these grammatical cases, except for the genitive, however the genitive is expressed by either a nominative noun that precedes its selecting head or a dative that follows it; in either way, possession is not expressed with a preposition.

Touching briefly on **H9** and **H10** – all hypotheses on morphology will be discussed in more detail in Subsection 4.4.5 – it was correctly hypothesized that Hungarian also has an accusative case ending in *-t*, which is marked on nouns and pronouns. Hypotheses **H9** and **H10** were formed by interpreting the results of AAA and GTI, in which the ending *-t* was highly associated to Hungarian words that were aligned to English nouns and pronouns that have an `obj` relation to their head, indicating a direct object relation. Importantly, this *-t* does not appear on Hungarian words aligned to subjects of intransitive verbs, indicating that Hungarian is not an ergative language: this conclusion allows us to form hypotheses on subjects and objects in Hungarian in Subsection 4.4.3. The discovery of the ending *-t*, which led to **H9** and **H10**, will be discussed in more detail below.

Gender was correctly hypothesized to be completely absent in Hungarian (**H13**). It was observed in DGAE and GTI that English lemmas *he* and *she* often received the same translations in Hungarian: *ő* or a suffixed form of that. There were furthermore no indications of gender being present on nouns, as there did not seem to be specific sets of affixes only occurring with one group of nouns, and not another – nor were there any attribute bundles found with AAA that are associated to two distinct endings. In fact, it may be enough to notice that there is no gender in pronouns in order to conclude that there is no gender in nouns: Greenberg's linguistic universal number 43 states that if a language has gender categories in the noun, it has gender categories in the pronoun (Greenberg 1963), although the number of languages Greenberg studies is limited. The absence of gender in Hungarian was also listed on the AL list.

Possessive pronouns were found not to be aligned to a Hungarian word in 34.9% of the cases. Based on alternations observed in GTI between suffixed and unsuffixed nouns, which seemed to be correlated to the English noun having a possessive pronoun as a child in the dependency tree, it was tentatively concluded that Hungarian suffixes the possessed noun with a personal possessive ending, making possessive pronouns redundant (**H14**). Indeed, Hungarian only uses possessive pronouns for emphasis or contrast (Rounds 2009: 140). It

[26] Bearing in mind that our corpus is the Bible – the preposition *unto* is often used as a dative construction in the KJB, e.g. Genesis 3:2: "And the woman said **unto** the serpent, We may eat of the fruit of the trees of the garden:"

was furthermore found in the data that a possessive pronoun, whenever it is expressed, precedes the noun, leading to the correct hypothesis **H15**. Additionally it was observed in the data and correctly hypothesized (**H16**) that there is no distinction between singular and plural third person possessive pronouns in Hungarian.

Hungarian words aligned to English adjectives were found not to cross with their head noun in 65.2% of occurrences and not to cross with any child nodes, i.e. modifiers, in an overwhelming 98% of occurrences. It was subsequently hypothesized that adjectives are mostly prenominal but can occur postnominally, but that the structure of the adjective phrase is the same as in English (**H17**). The latter part of the hypothesis was formed, seeing as the word order inside the adjective phrase was mostly the same, and there were hardly any crossings observed among the children of adjectives. However, while the latter part of **H17** is correct, Hungarian adjectives can only occur prenominally. Later inspection of the data showed that many adjectives were wrongly tagged as such, with many occurrences of *thy*, *unto* and *Lord* receiving the tag ADJ. The personal pronoun *I* was often interpreted by the tagger as the Roman numeral one, which was then interpreted as *first* and tagged as an adjective. The noisy nature of the ADJs highly influenced the numbers and consequently led to a partly wrong conclusion; however, we believe that thorough inspection, especially in the GTI output, could have laid bare this tagging error.

The fact that a noun phrase containing a numeral has a singular head noun in Hungarian (**M2**) was missed. Currently, there are a few complications in the data processing and output formatting that would prevent a linguist from forming a hypothesis about the grammatical number of a head noun in a noun phrase containing a numeral, even when they are specifically researching this question. Due to the way the data are represented in the dataframe, numerals can only "see" the POS of their parent nouns and what dependency relation they have to it – and not the Hungarian word their parent noun is aligned to, which is necessary to be able to see that it is singular.[27] Due to the way the output of our tools is formatted (and the way it suppresses infrequent attributes), numerals cannot be accessed and easily investigated as children of nouns, as they are so infrequent[28] that they are washed away among the much more frequent determiners or adjectives (or even `None`), making them "invisible". These complications led to **M2** currently being missed, however we believe that if a linguist could narrow down on numerals as children of nouns more easily, our tools would work well and provide linguists the information needed to form meaningful hypotheses about the grammatical number of head nouns in a noun phrase containing a numeral.

Lastly, **M3** was not found: a demonstrative and a definite article necessarily co-occur in a Hungarian noun phrase (e.g. *ez a hely* lit. 'this the place'). In our current setup it is not possible to find this: when only the demonstrative and

[27] In order to be able to notice that a word is singular, the linguist would first need to form a hypothesis about nominal paradigms.

[28] Only 6073 out of the 737319 tokens in the English Bible were tagged as a numeral: 0.8%.

the noun are aligned to an English word[29] we cannot discover in our dataframe that *a* or *az* was there in the Hungarian sentence. Adding concordances or adjacent words of aligned-to words in the target language could perhaps allow the linguist to find features such as these.

4.4.3 Verbs and constituent order

22 hypotheses were formed that concern the verbal domain or constituent order. Of these hypotheses, 19 were correct, while one was an overgeneralization. Two hypotheses have not yet been confirmed or rejected, as they require further research. Dr. Lipták compiled 25 differences between English and Hungarian that pertain to the verbal domain or the constituent order, of which 6 were correctly discovered, but 19 missed.

Unlike Japanese or Bantu languages, English does not encode the information structure of a sentence with morphemes. The result is that the dependency tree as produced by UDPipe is not annotated for the information structure of the sentence in any way, and that information-structural knowledge can therefore not be mapped onto Hungarian sentences. However, as many languages rely much more heavily on word order to encode the information structure than English does, a linguistic user of our tools can venture the hypothesis that the target language does, too. In that case, investigating crossings can provide valuable insights into the freeness of word order and consequently the information structure of the target language.

It was observed that in 39.1% of all occurrences of a verb there was a crossing between it and one of its arguments (i.e. one of its descendant nodes, which include complements, auxiliaries and adverbs), indicating a different word order than in English, while in 60.9% there was no crossing. It was thus hypothesized that Hungarian word order is much freer than in English (**H18**), because these statistics suggest that Hungarian does not systematically have the same or a different word order than English. These crossing frequencies are similar for both transitive and intransitive verbs, leading to the correct hypothesis that both types of verbs behave similarly in this respect (**H19**).[30]

Subjects were hypothesized to precede the verb in general (**H20**), observing that Hungarian words aligned to English subject nouns occur on the same side of the verb as English subject nouns in 71.8% of occurrences, which is before the verb. Similar numbers were found for object nouns, which come on the same side of the verb, i.e. after it, in 77.3% of occurrences, leading to the hypothesis that Hungarian has a standard VO order (**H21**). Together,

[29] This is what happens. The aligner learns that *a(z)* is to be aligned to *the*, and that *ez* is to be aligned to *this*, however *the* is absent. Indeed, *ez* and *a(z)* often occur together in Hungarian, but only when the demonstrative modifies a noun; if it is used predicatively, the article is absent in Hungarian, too. The aligner therefore does not learn to align *this* to both *ez* and *a(z)* at the same time, but instead leaves the article unaligned.

[30] We don't know whether transitive and intransitive verbs take up different sentence positions in any language – but at least we know it does not make a difference in Hungarian.

these numbers led to the hypothesis that Hungarian is primarily SVO (**H22**). However, while it is (or may be) quantitatively correct that SVO is the most prominent constituent order in Hungarian, all orders can occur: in about one in four sentences, Hungarian subjects and objects occur on the other side of the verb than in English, confirming the hypothesis of a freer word order (**H18**) and that of information structure being encoded through syntactic movement, as well.

Indeed, Dr. Lipták later confirmed that word order in Hungarian is in principle free and wholly determined by the information structure of a sentence. Hungarian word order is characterized by four sentence positions: Topic–Preverb–Verb–Rest. While the topic position may be empty, it is usually filled by the subject (hence the SV order) but can be filled by other constituents, too. The preverbal position neutrally contains a verbal complement, an adverbial or a coverb, prefixed to the verb, but can also contain the focus of the sentence, such as wh-words, negation or otherwise stressed phrases. Importantly, whenever the preverbal position is taken up by the focus, any other material that would have gone into the preverbal position is moved after the verb (to Rest), creating the possibility for all word orders to arise (Rounds 2009: 254; cf. also Kiss 2002).

It was furthermore hypothesized that the relative order of subjects, objects and other constituents is predominantly the same as in English (**H23**). This conclusion proved to be correct, and was made based on the fact that object nouns rarely cross (13.1%) with their sisters in the dependency tree. This means that the position of subjects, nominal adverbials (such as locative, temporal or directional (possibly prepositional) complements, excluding adverbial clauses; i.e. a noun or pronoun receiving the `obl` relation to its head), adverbs and auxiliaries relative to the object is the same in English as it is Hungarian in 86.9% of the cases. That is, if a Hungarian subject is present in the sentence (or rather, has received an alignment to an English subject) it will appear on the left of the object in the majority of the cases, while nominal adverbials are on its right, whenever they are there. However, if the order is different, then it is mostly the adverbial that comes to the left of the object, while the subject rarely comes to the right. This could be read as that whenever a subject is overt in a Hungarian sentence, it will often be the topic, and therefore on the left of the object. Meanwhile, nominal adverbials are usually to the right of the object, but can be topicalized, ending up on the left of the object, as well. And, if a subject and an object are both present in a sentence where an adverbial is fronted, the subject would still appear on the left of the object in the majority of the cases.

Adverbials that are tagged as nouns or pronouns by UDPipe with an `obl` relation to their heads (such as locative, temporal or directional complements, excluding adverbial clauses) were indeed found to be mostly postverbal, leading to correct hypothesis **H24**. 20.8% of English adverbials cause a crossing with its ancestor verb, meaning that about one in five Hungarian adverbials are on the other side of the verb compared to English. This is further supported by the

fact that English adverbials preceding the verb have their translation appear to the right of the verb in Hungarian in 32.9% of the cases, while English adverbials following the verb have their translation appear on the other side of the verb in 15.5% of the cases: Hungarian adverbials are therefore more likely to follow the verb.

Hungarian adverbial clauses, on the other hand, were correctly hypothesized to be in the same position as in English (**H25**). DGAE shows that an overwhelming 93.6% of all adverbial clauses in Hungarian appear on the same side of the main verb as in English. It is not entirely clear why, but it could be the result of translation, where word order is kept constant throughout translations, or it could show a correlation between the weight of an adverbial clause and its position.

We furthermore observed that the verb of an adverbial clause crosses with one of its arguments in 63.7% of all adverbial clauses (i.e. the argument appears on the other side of the verb compared to English), less than the amount of verb-argument crossings in a main clause. We interpreted this as perhaps being the result of a slightly less free word order in embedded clauses, leading to hypothesis **H26**. This stricter word order in adverbial clauses could again be a statistical anomaly, with Dr. Lipták not being aware of any such restriction on the order of words in adverbial clauses, but it is nonetheless a valid reason to investigate Hungarian adverbial clauses in more detail. As of yet, the correctness of **H26** is unknown.

Pronouns were correctly hypothesized to behave similarly to nouns, positionally (**H27**). Pronoun subjects, objects and adverbials appear in the same positions as nouns do, showing similar crossing statistics. This means that, in general, pronoun objects appear after the verb – and are not proclitics as in French – and that pronoun subjects and adverbials appear before and after the verb, respectively.

The AL list mentions that Hungarian allows for the dropping of subject pronouns as well as singular object pronouns. While it was indeed correctly hypothesized that Hungarian has subject pro-drop (**H28**) – by noting that English subject pronouns were not aligned to a Hungarian word in 47.6% of the cases – a hypothesis was not formed on pro-drop including singular objects (**M4**). Nonetheless, Hungarian object pro-drop could have been found, by observing in DGAE and GTI that English singular object pronouns often have no translation either.

More than half, 58.5%, of English auxiliary and modal verbs (POS tag AUX in UD) were not aligned to a Hungarian word, leading to the correct hypothesis that auxiliaries are much less frequent in Hungarian and that Hungarian is less analytical than English (**H31**). Indeed, Hungarian has a synthetic past tense, without a *have*-like auxiliary, does not have an analytical continuous, and can express the future with a present tense, similar to Finnish and Dutch, for example. However, the future can also be expressed with the auxiliary *fog*, but this was not found in DGAE or AAA, with no form of *fog* being among the top-

20 most aligned-to Hungarian words for English auxiliaries.[31] As for modals, Hungarian expresses *can* or *may* with the suffix *-hat/-het*, also adding to the unaligned auxiliaries in English.

The fact that Hungarian does not have passivization nor a passive auxiliary was on the AL list. It was found in DGAE that passive auxiliaries, which are tagged with a distinct relation to their head verb in UD, are not aligned in 64.6% of the cases, with other translations seeming to be noise,[32] suggesting that there is indeed no passive auxiliary. This discovery was however generalized to the hypothesis that Hungarian has a synthetic passive voice (**H32**). Dr. Lipták pointed out that this is not true for modern Hungarian, where the third person plural is used instead of passives. However, more archaic Hungarian, such as in the Bible, does still have passivization to some degree, making **H32** correct for this specific corpus. It must be noted, though, that it would be very difficult, if not impossible, to detect whether a language has passivization with our tools. Although AAA does retrieve (with rather low association scores) two suffixes that, to the best of our knowledge, are unique for the archaic passive forms in Hungarian (*-tik* and *-tott*), the attribute bundles with which they are associated do not contain any information about it being a passive, and only tell us that the endings are associated with an English past participle.[33] DGAE and GTI also do not grant good insight in the presence of passives in Hungarian.

It was observed in DGAE and GTI that 50.2% of all infinitival markers *to* are not aligned to a Hungarian word, with other alignments mostly containing *hogy* 'that, (in order) to' (29.5%). This led to hypothesis **H33**: Hungarian does not have an infinitival marker such as English *to*, which turned out to be correct.

English *not* was aligned in 98.1% of its occurrences, most frequently to *nem* (62.4%), *ne* (21.4%), *sem* (4.5%), *meg* (2.1%) and <,> (1.7%). It was consequently concluded that negation in Hungarian is not done with verbal morphology on the main verb. It was therefore hypothesized that Hungarian negation is done with adverbs or particles (**H34**), which turned out to be correct. The possibility of Hungarian negation being expressed through a negative auxiliary verb was ruled out, because one would expect for *not* to be aligned to more forms containing the same stem, as well as a less skewed distribution over the aligned forms corresponding to multiple grammatical persons, since we had seen in **H29** (discussed in 4.4.5) that Hungarian verbs inflect for all persons.

[31] In fact, the top-20 most aligned-to Hungarian words for auxiliaries contained many non-auxiliaries, indicating noise – with the most striking being that the comma (i.e. <,>) was the most common Hungarian alignment among English auxiliaries, but only amounted to 6% of all aligned cases. This even more strongly corroborates the conclusion of H31.

[32] Similar to non-passive auxiliaries, the most common Hungarian alignment was the comma (i.e. <,>), but it only amounted to 2.1% of all aligned cases. Other alignments (including articles, conjunctions and more punctuation) were even less frequent, suggesting that the bulk of the Hungarian alignments of English passive auxiliaries are noise.

[33] In fact, nine other endings are associated to past participles, and although these endings could be passive suffixes, they can also be active past tenses.

It was also correctly hypothesized that negation precedes the negated, specifically the verb (**H35**), by observing in GTI that *not* does not cross with its ancestor in 71.5% of its occurrences. While negation never comes left adjacent of the finite verb or auxiliary in English, it does usually come left of the main verb,[34] which is the root of the dependency tree. It can therefore be deduced from the absence of a crossing that negation comes before the verb in Hungarian, and precedes any negated word or phrase. The positioning of negation is on the AL list as well, specifically pointing out that negation is left adjacent to the finite verb or auxiliary, contrary to English.

The hypothesis that other adverbs also precede the main verb in Hungarian (**H36**), turned out to be correct, too. This was tentatively concluded based on the fact that adverbs preceding their heads in English are much more common (73.6% of all adverbs in English come in a position before their heads), while it was observed that whenever an adverb precedes its head in English, its alignment would be to the right of the head in Hungarian in 14.6% of the cases (i.e. 10.7% of all adverbs) and whenever an adverb follows its head in English, its alignment would be on the left of the head in Hungarian in 31.3% of the cases (i.e. 8.3% of all adverbs). These relative frequencies, as illustrated in Table 4.6, then show that adverbs are more likely to precede their head, primarily main verbs, in Hungarian.

		Hungarian		
		left	right	total
English	left	62.9%	10.7%	73.6%
	right	8.3%	18.1%	26.4%
	total	71.1%	28.9%	100.0%

Table 4.6: The distribution of the relative positions of adverbs in English and in Hungarian, as deduced from crossing frequencies. The conclusion is that Hungarian adverbs tend to come to the left of their heads (**H36**).

It was hypothesized that pro-adverbs, such as *then* and *so*, are not as abundant in Hungarian (**H37**). This has not yet been confirmed or rejected; neither Dr. Lipták nor a grammar of Hungarian could provide an answer to the matter. It came to be hypothesized as it was observed in DGAE that almost one in five adverbs did not receive an alignment, and further inspection in GTI showed that it was mostly due to such small adverbs.

It was noted that English copulae were not aligned to a Hungarian word in 43.8% of occurrences. This led to hypothesis **H38**: that Hungarian allows for zero copula. This is, however, a slight overgeneralization, as Hungarian only allows for the dropping of the third person forms of *van* 'to be', and only if

[34] Except for archaic constructions such as *I know not*, which, admittedly, are present in the Bible.

the predicate is a noun or an adjective. This restriction on zero copulae in Hungarian could have be found in GTI, though.

When copulae are overt in Hungarian (i.e. when an English word tagged as copula is aligned to a Hungarian word), however, they were found to cross in 31.1% of the cases. This led to the hypothesis that Hungarian copulae can come both after or before the predicate, though in general before (**H39**). It could not be found what causes these crossings, though it turns out that copulae come before the predicate if the preverbal position is taken up by e.g. negation (Rounds 2009: 254).

Several morpho-syntactic features or differences between Hungarian and English on the AL list were not found. It was not found that in Hungarian any number of constituents can come before the verb (**M5**). This is simply because our tools do not collect statistics on the number of constituents preceding or following the verb or any other head, although this could easily be implemented.

To continue, three differences concerned question phrases (**M6**–**8**), or wh-phrases, and another two concerned yes-no questions (**M9**–**10**). These differences were unfortunately not found because wh-words are not separately tagged in UD; while interrogative personal pronouns do receive a `feats` tag that distinguish them from other pronouns, other wh-words, such as *where*, *whence* and *how*, do not. It is therefore difficult to detect any morpho-syntactic features pertaining to question phrases when using UD tagging. Furthermore, sentences are not individually tagged for sentence function (declarative, interrogative, exclamative or imperative), which would be very beneficial for the detection of morpho-syntactic features pertaining to yes-no questions. Finally, questions are not all too frequent in the Bible, making the corpus somewhat unsuitable for the detection of differences with regards to questions.

The difference that *only N* phrases must be left adjacent to the finite verb or auxiliary in all types of clauses was not found (**M11**). This is due to the fact that the difference is rather fine-grained and can easily be missed if one is not looking for this difference in particular. Furthermore, the word *only* only occurs 255 times in the KJB (or at least, the section that we used), making *only N* phrases highly infrequent, and even so infrequent that they do not show up in the GTI, which limits its output. Though, even if the construction was more frequent, our tools do not automatically correlate the position of the noun in Hungarian (in terms of crossings) to the fact that it contains an aligned instance of *only*, making it difficult to spot this pattern.

Another important difference, **M12**, was also not found: Hungarian present and past tense finite verbs show agreement with the definiteness of the object; verbal paradigms depend on whether the object is definite or indefinite. Seeing as nouns are not tagged for definiteness (although articles are), it is very hard to detect a pattern in GTI, and even impossible for AAA to correctly associate specific verbal suffixes to the definiteness of the object.

It was also missed that infinitives in Hungarian sometimes agree with subjects (**M13**), which happens when an infinitive is used with an impersonal

verb such as *kell* 'must'.[35] This cannot be found because this peculiarity of Hungarian is solely dependent on a Hungarian context and cannot be clearly related to an English construction. While one may be able to find that *kell* means 'must' in GTI, it cannot be found what forms the English infinitives are aligned to in those constructions.

Furthermore, verbs show singular agreement with a noun phrase that contains a numeral (**M14**).[36] This cannot be found directly, because it is not possible in the way data are represented to see grandchildren nodes, i.e. daughter nodes of daughter nodes in the dependency tree in English, which is necessary in order to be able to see that the verb's subject (which is the verb's daughter) is modified by a numeral, represented as the subject's daughter. However, even if our tools returned statistics on grandchildren, singular agreement with a subject modified by a numeral can only be found if one already has a hypothesis about the paradigm of the verb in the target language, as otherwise it will be very challenging to notice a pattern.

Lastly on verbs, there are several differences between English and Hungarian pertaining to verbal particles on the AL list – eight, in fact. Verbal particles in Hungarian, also called preverbs in Hungarian linguistics, comprise resultative, terminative and locative elements that telicise the verb (see Ladányi 2015 for a recent overview), while Dr. Lipták took English verbal particles to be particles that associate with phrasal verbs, such as *away*, *down*, *forth* and *up*, whose functions are in many cases similar to those of Hungarian preverbs. Hungarian preverbs can come both before and after their verb. Dr. Lipták lists the following differences:

M15 A verbal particle can be left adjacent to its verb in Hungarian, while in English it cannot.

M16 If a verbal particle follows its verb in Hungarian, it can show up in any position between the constituents following the verb, while in English it can only come in fixed positions.

M17 A verbal particle can occur before an auxiliary in Hungarian, while in English it cannot. However, not all auxiliaries allow for this.

M18 A verbal particle that is left adjacent to its verb can be reduplicated in Hungarian if it is shorter than 3 syllables long.

M19 A verbal particle can only be left adjacent to its verb in Hungarian if the verb is not preceded by an *only N* phrase.

[35] With impersonal verb, we mean a modal verb without arguments.

[36] Of course, the fact that numerals select singular nouns in Hungarian (**M2**) was already missed, and because **M2** was missed, **M14** was highly unlikely to be found, too. Typologically it is not necessarily surprising that Hungarian verbs show singular agreement with a noun phrase that contains a numeral, however there also exist languages, such as Russian, that show plural agreement with a noun phrase that contains a numeral (larger than one), despite (some) numerals selecting a singular noun; and therefore a difference like **M14** would ideally be found.

M20 Idem, if the verb is not preceded by a question phrase.

M21 Idem, if the verb is not preceded by sentential negation.

M22 A verbal particle cannot be left adjacent to its verb if the clause has progressive aspect.

None of these differences were found, however. This is probably due to several reasons, including orthography: whenever a verbal particle is left adjacent to its verb in Hungarian, it is attached to the verb as a prefix, while if it follows the verb it is not. The problem is that prefixes – or any affixes, for that matter – are not analysed as a separate token by `eflomal`, the aligner that we used. This results in frequencies of preverbs (that are not attached to the verb) being heavily underrepresented, and that the co-occurrence of English verbal particles and Hungarian preverb tokens may be too low for them to be consistently aligned to each other. In turn, this leads to many English verbal particles to be unaligned incorrectly.

Another reason includes tagging of verbal particles in English. UD treats Germanic verbal particles as adpositions or adverbs, making it very hard to distinguish verbal particles from other adpositions or adverbs in DGAE or GTI. In other words, there is no simple way to identify verbal particles in the output of our tools, and therefore to draw any conclusions concerning them.

Yet, even if the alignment and tagging problems were solved, the ability to correctly detect any morpho-syntactic features concerning verbal particles in Hungarian hinges on the assumption that all English verbal particles will always have a translation in Hungarian and vice versa, which may very well not be the case as this is lexical to a significant degree. There are examples of English so-called phrasal verbs of which the verbal particle does not have a (preverbal) translation in Hungarian, e.g. *ask around* vs. *kérdezősködik* and *call up* vs. *telefonál*. Conversely, there are many examples of Hungarian preverbs that have no direct translation to an English verbal particle, e.g. *return* vs. *visszatér*, which contains the prefix *vissza-* 'back' and can be separated from the verb. The fact that the presence of a verbal particle in both languages is lexically determined to a large extent, makes it very hard to detect them in the target language, seeing as the linguistic annotation of the source language is mapped onto the target language: if there is no verbal particle in the source language, it is impossible to see if it is present in the target language.

Detecting morpho-syntactic features of Hungarian with regards to verbal particles therefore proved very difficult. Setting aside the non-distinctive tagging of verbal particles, **M15** to **M17** were not found because of the high frequency of unaligned English verbal particles, leading to the impossibility to detect the position of the preverb in Hungarian. **M18** was not found for the additional reason that our tools do not correlate or associate the presence of words (or affixes) in the target language with other words in the target language, making it hard if not impossible to see that the verbal particle can be reduplicated. Ideally, the aligner aligns the English verbal particle to both real-

izations of the preverb, but this does not happen in practice. **M19** to **M22** are not found for the same reason as why it was not found what causes the copula to end up after the predicate: no correlation can be found between the position or any feature of a word in the target language and the presence of a specific type of phrase that is not the word's head or modifier.

4.4.4 Other hypotheses

Four more hypotheses were formed about Hungarian morpho-syntax, that do not necessarily fall under the nominal or verbal domain or under constituent order. One of these four was incorrect, while one difference on the AL list was missed.

It was hypothesized that Hungarian has both prepositions and postpositions (**H40**). Out of the aligned adpositions in English 55.9% had no crossing, meaning their Hungarian alignment shows up on the same side of their head in about half of the cases. **H40**, however, is incorrect: Hungarian only has postpositions, a difference also on the AL list. Although we believe that the numbers did suggest the presence of both prepositions and postpositions, the numbers were misleading. The group of English words that received the ADP tag also includes conjunctions (e.g. *for* and *as*) and, as mentioned, verbal particles. Verbal particles, in particular, end up after the verb in English, and whenever there is an alignment to a Hungarian word, that word will also be after the verb (seeing as Hungarian verbal particles preceding the verb will be prefixed to it in writing), leading to the absence of a crossing. As for the conjunctions, they always come before the verb, in both languages, leading to the absence of a crossing, as well. Further investigation laid bare problems with alignment, as many prepositions in English were aligned to Hungarian articles and other determiners.

Coordinating conjunctions were correctly hypothesized to precede the second conjunct (**H42**), as English conjunctions did not cross with their head in 85.3% of the aligned cases.[37] Similar numbers were found for subordinating conjunctions, which do not cross with their head verb in 88.4% of the aligned cases, and do not cross with their siblings, including subjects and objects, in 86.9% of the aligned cases – this led to the correct hypothesis **H43** that subordinating conjunctions mostly end up in the same position in Hungarian as in English.

Finally, it was missed that Hungarian is a negative concord language (**M23**). This feature of Hungarian could only be detected if English *not* was aligned to multiple Hungarian words at the same time (a one-to-many alignment), but this was only very rarely observed; or by correlating the presence of a Hungarian

[37] In UD the first conjunct is the head of the clause, while all other conjuncts depend on it via the `conj` relation. If, then, for example, *John and Mary* is the object in a sentence, only *John* would have the `obj` relation to the verb, while *Mary* would be a daughter node of *John*, having the `conj` relation. The conjunction *and*, in turn, would be a daughter node of *Mary* via the `cc` relation.

negation to the presence of another word or affix in the Hungarian sentence, which, as mentioned, is not done by our tools.

4.4.5 Hypotheses on affixes

Seeing as automatic affix detection is an important goal in the field of comparative syntax, we will conclude this section by discussing all hypotheses and differences pertaining to affixes in here. Specifically, we will explain how these hypotheses were formed.

The most important tools for affix detection are the AAA and the GTI. While the AAA tries to associate affixes with attribute bundles and retrieves a list of candidate affixes in the target language, the GTI can be used to further explore the data with these candidate affixes in mind.

Revisiting **H9** and **H10**, the accusative ending in Hungarian was hypothesized to be *-t* because AAA retrieved *-t* as being associated to the attribute bundle `(deprel=obj, parent=VERB|Trans, pos=NOUN)`. This means that the affix *-t* is very common in Hungarian words that are aligned to English nouns that have a direct object relation to their head verb. In fact, this association is the highest association found; see Figure 4.6 for the top-20 affix-attribute associations. Further inspection in the GTI, which gives more detailed breakdowns of attributes than DGAE, showed that (nearly) all English object nouns were aligned to Hungarian words ending in *-t*, leading to the correct hypothesis that *-t* is the accusative ending in Hungarian (**H9**), because, as mentioned above, this *-t* does not seem to occur with subjects of intransitive verbs. Despite the fact that AAA associates *-t* most strongly with nouns, and not pronouns, further exploration in the GTI suggested that pronouns do also often bear this accusative suffix *-t* (as shown in *őt* 'him/her', *őket* 'them', *melyet* 'which' and its plural form *melyeket*, *azt* 'it, that', *minket* 'us', among others), leading to the correct hypothesis **H10**: that *-t* is also the accusative marker for pronouns.[38]

To discuss the AAA output in slightly more detail – Figure 4.6 furthermore shows many associations between attribute bundles and the prefix *mond-* or the word *monda*. It is clear from the AAA output that this Hungarian prefix or word is associated with an indicative mood, a finite verb form, being an intransitive verb, and also a third person and a past tense in English. An association with the English word *and* being a daughter node of the verb is also found. Further inspection of the data strongly suggests that *mond-* in fact means 'to say' (which turned out to be correct), a verb that is very common in the Bible, especially in the third person, past tense and with the word *and* being a daughter node.

The Hungarian affix *meg-* is reported by AAA to be somewhat highly associated with English transitive verbs in the past tense. Despite inspection of

[38] However, it turns out that there are two exceptions to this: *engem* 'me' and *teged* 'you (sg.)'. These two forms nonetheless do appear in dialectal Hungarian as *engemet* and *tegedet*, respectively, although they are not attested in the Bible translation that was used in this research.

attribute bundle	affix	weight
(feats=(Tense=Past), feats=(VerbForm=Part))	-tt	0.116655
(deprel=obj, parent=VERB|Trans, pos=NOUN)	-et	0.125745
(feats=(VerbForm=Inf))	-k	0.126978
(feats=(Tense=Past), pos=VERB|Intrans)	mond-	0.127453
(deprel=obl, feats=(Number=Sing), pos=NOUN)	-l	0.127497
(feats=(Tense=Past), feats=(VerbForm=Part))	-k	0.128595
(children=NOUN|nmod, children=the|det, pos=NOUN)	f-	0.130742
(children=and|cc, feats=(Mood=Ind), feats=(VerbForm=Fin))	monda	0.132111
(feats=(PronType=Prs))	nék-	0.132633
(children=of|case, pos=NOUN)	-nak	0.135104
(feats=(Mood=Ind), feats=(Person=3), feats=(Tense=Past), feats=(VerbForm=Fin))	monda	0.138214
(feats=(Number=Plur), parent=NOUN)	-k	0.142976
(feats=(Mood=Ind), feats=(VerbForm=Fin), pos=VERB|Intrans)	monda	0.144797
(children=of|case, pos=NOUN)	-k	0.145477
(feats=(Tense=Past), pos=VERB|Trans)	meg-	0.145837
(feats=(Number=Sing), parent=VERB|Trans, pos=NOUN)	-t	0.155972
(children=and|cc, feats=(Mood=Ind), feats=(VerbForm=Fin))	mond-	0.168884
(feats=(Mood=Ind), feats=(Person=3), feats=(Tense=Past), feats=(VerbForm=Fin))	mond-	0.172108
(feats=(Mood=Ind), feats=(VerbForm=Fin), pos=VERB|Intrans)	mond-	0.225624
(deprel=obj, parent=VERB|Trans, pos=NOUN)	-t	0.41646

Figure 4.6: The top-20 affix-attribute bundle associations as retrieved by AAA. The higher the weight, the higher the association is between the affix and the attribute bundles.

the data it could not be narrowed down what this prefix means, however it was later revealed in Rounds (2009) that *meg-* is an aspectual prefix called a preverb, which are also discussed above.

In Figure 4.6 it can also be observed that the ending *-k* is associated with several attribute bundles. Indeed, many Hungarian words that are aligned to English nouns that have a child preposition *of* end in *-k* (because a genitive construction in Hungarian is expressed using the suffix *-nak* – which can also be found in Figure 4.6 – and *-nek*, or using a zero suffix; Rounds 2009), and to English plural nouns (the nominative plural ends in *-k*, possibly with a preceding linking vowel; Rounds 2009). Though the AAA output shows that the ending is also associated with English past participles and infinitives, Hungarian past participles and infinitives do not end in *-k* (Rounds 2009). This association as returned by AAA can be explained through the fact that Hungarian main verbs are aligned to English main verbs, and English main verbs are often non-finite, with auxiliaries showing finite verbal morphology. Indeed, Rounds (2009) confirms that many Hungarian finite verbal forms (which would be aligned to English non-finite, main verbs; a result of Hungarian having much fewer auxiliaries than English, see **H31**) end in *-k*: plural forms all end in *-k*, as well as the first person singular in certain forms. The fact that AAA wrongly (or at least incompletely) retrieves the suffix *-k* as what can be interpreted as a genitive suffix and as what can be interpreted as a non-finite verbal suffix, can therefore be explained by the confusion of multiple alternating suffixes (such as *-nak* and *-nek*), as well as different suffixes that share an attribute in the English annotation and happen to both end in *-k*.

Among the top-20 affix-attribute bundle associations is also the ending *-l*, which is associated to singular nouns that have an `obl` relation to their head, used to denote non-core (oblique) arguments or adjuncts. While *-l* is not a case suffix in Hungarian nominal morphology in itself, several case endings end in *-l*: the elative, delative, adessive, ablative, instrumental and sociative cases are all denoted with a suffix that ends in *-l* (Rounds 2009). Such noun phrases would typically receive an `obl` relation in UD. Similar to what happens with *-k*, *-l* is retrieved because of the confusion of multiple, longer suffixes that share an attribute in the English annotation and happen to all end in *-l*.

AAA found several suffix pairs associated with the same attribute bundles that had an alternating vowel. For example, *-nak* and *-nek* both appear three times in the AAA output: once associated with `(children=of|case, pos=NOUN)` (also seen in Figure 4.6); once with `(deprel=nmod, feats=(Number=Sing), parent=NOUN, pos=NOUN)`; and once with `(deprel=nmod, parent=NOUN, pos=NOUN)`. In a similar fashion, *-ból* and *-ből* are associated to the same attribute bundles, as are *-ban/-ben* and *-tok/-tek*. The fact that this alternation did not seem to be caused by any other morpho-syntactic or lexical feature, such as gender (cf. *embernek* 'man' vs. *királynak* 'king') led to the correct hypothesis **H11**: that Hungarian shows a systematic form of vowel harmony, most

likely to be front-back vowel harmony.[39]

The stacking of suffixes, such as seen in *őt* 'him/her' vs. *őket* 'them', *melyet* vs. *melyeket* 'which', as well as in *ember* 'man' vs. *embernek* 'of/to (the) man' vs. *emberek* 'men' vs. *embereknek* 'of/to (the) men', led to the correct hypothesis that Hungarian is agglutinative (**H12**), as case suffixes, such as *-nak/-nek* and *-t*, are stacked onto the plural suffix *-k*, sometimes with a linking vowel. AAA found a few stacked affixes, among which *-(o)kat* (associated with plural nouns, as well as object nouns; the suffix indeed corresponds to plural object nouns) and *-knak* (equivalent to plural *-k* + dative *-nak*).

Concerning verbal morphology, it was correctly hypothesized that verbs inflect for all persons in both present and past tense (**H29** and **H30**), a morpho-syntactic difference between English and Hungarian also on the AL list. This feature of Hungarian was found by observing very rich morphology in DGAE and GTI; English verb lemmas were aligned to a plethora of different Hungarian words, which occurred in different forms with distinct endings and prefixes. By identifying the subjects of the English verbs in the attribute bundles (subjects are children of the verb in UD), these endings could be clearly matched to a grammatical person. As such, the verbs *mond-* 'to see' and *tud-* 'to know' can be observed in several forms in the present and past tense, where all persons receive distinct endings; see Table 4.8. Of the observed endings in the Table, AAA correctly discovered *-ának* as being associated with third person plural past tense indicative. It also found *-om* though with an incomplete attribute bundle associated to it. AAA additionally found front-vowel counterparts of two listed suffixes: *-em* and *-ünk*, though both with incomplete attribute bundles as well.

It must be noted that the columns in Table 4.8 turned out to contain multiple paradigms: while *mondom* is indicative, *mondjak* is subjunctive, for instance. However, the observed forms still prove that verbs decline for all persons, in both present and past (although the second person plural in the past tense was not observed for both verbs; it still seemed a safe – and indeed correct – assumption that it would receive a suffix distinct from all other persons). Also note that the first person plural either seems to receive *-unk* or *-juk* in the present tense: this is due to Hungarian verbs agreeing with the definiteness of the object, a missed difference also discussed above (**M12**).

With AAA it was also hypothesized that infinitives in Hungarian end in *-ni* (which is correct) and that *-á/-é* is a frequent past tense suffix, possibly third person singular (correct). Furthermore, *-tt* was hypothesized to be a past participle – however, the real ending turned out to be *-ott/-ött*, and turned

[39] There are more possible explanations for the alternation of suffixes, such as dissimilation, gender or other word classes, or simply multiple noun declinations, that should in principle be tested. However, when reviewing the data, I noticed there seemed to be a correlation between the presence of certain vowels in the stem and the vowel in the suffix, but I never quantified this correlation. In forming the hypothesis, I may have been somewhat guided by my limited knowledge of Finnish, a language related to Hungarian, of which I know it has vowel harmony.

subject	***mond-* 'to see'**		***tud-* 'to know'**	
	PRS	PST	PRS	PST
I	mond-om, mond-jak	mond-ám	tud-om	
thou		mond-ál	tud-od	
he/she/it	mond-ja	mond-a		
we	mond-unk	mond-ottuk	tud-juk	
ye			tud-játok	
they	mond-ják	mond-ának	tud-nak	tud-ják

Table 4.8: Observed forms of the Hungarian verbs *mond* 'to see' and *tud* 'to know', in present and past. Note that the columns contain multiple paradigms.

out to be also used in the finite past tense. The ending *-ék* was hypothesized to be third person past indicative, but that is not entirely correct: the third person singular does not show any such ending, while the plural does but with an additional *j*, *t* or *n* before it. The ending *-ék* is therefore likely to be a result of the algorithm trying to generalize over *-jék*, *-ték* and *-nék*.

head	***nék-* 'unto'**	***ellen-* 'against'**
me	nék-em	ellen-em
thee		ellen-ed(?)
him/her	nék-i	ellen-e
us		ellen-ünk
you (pl.)	nék-tek	ellen-etek
them	nék-ik	ellen-ük(?)
NOUN		ellen

Table 4.9: Some postpositions in Hungarian decline for person, such as *nék-* 'unto' and *ellen-* 'against'. Listed are some attested forms.

Furthermore on adpositions, it was correctly hypothesized that some postpositions in Hungarian decline for person (**H41**). This was found by observing that the prepositions *unto* and *against* were aligned to multiple Hungarian words, depending on the head of the preposition.[40] It was thus observed that all Hungarian aligned words started with *nék-* 'unto'[41] and *ellen-* 'against', and have different endings for each different pronominal head (which are reminiscent of verbal endings) as shown in Table 4.9. The preposition *against*

[40] In UD nouns and pronouns are the heads of prepositions, because it follows the convention that all functional words are dependent on content words. This is done in order to parse sentences more uniformly cross-linguistically.

[41] *Nék-* is an archaic or dialectal variant of modern *nek-* 'to, for'.

furthermore shows that it does not receive an ending if its head is a noun.

4.5 Discussion

4.5.1 On the results and subjectivity

The results discussed in the previous section show that our tools are effective and useful in the detection of morpho-syntactic features of a language. It was observed that the large majority of the hypotheses that were formed by analyzing the output of the DGAE, GTI and AAA are correct. Not only do the hypotheses formed have a high precision, the output of the tools even gave rise to two questions about Hungarian syntax, the answers to which are to the best of our knowledge as of yet unknown: both H26 (there is a stricter word order in Hungarian subordinated clauses than in main clauses) and H37 (Hungarian uses fewer pro-adverbs than English) remain to be confirmed or rejected. On the other hand, several differences on the AL list were not found, indicating that our tools do not detect every difference. However, many of these missed differences can be attributed to either the interpretation of the output by the linguist (e.g. M4), lacking annotations (e.g. M6–8) or the processing and formatting of the data by the tools (e.g. M2 and M3), for all of which there is room for improvement.

Of course, there is no objective measure of performance of our tools. In this research, we tried to overcome this lack of a formal test set by compiling a list of hypotheses based on the output of the tools, while an expert of Hungarian independently compiled a list of characteristic morpho-syntactic differences between Hungarian and English (the AL list). Both lists are far from complete, and many more differences could have been discovered (and hypotheses formed) with the help of our tools, and many more differences exist that were not on the AL list. While we think we have sufficiently shown that our tools can successfully aid a linguist in the detection of syntactic differences between a source and a target language, the evaluation carried out in this chapter does not give a complete overview of the full range of possibilities and, especially, the shortcomings of the proposed method and presented tools. Ideally, a more objective measure or a dataset should be developed to more adequately grasp the performance of tools for the automatic detection of syntactic differences between languages, but it is not clear at present how this could be achieved.

As mentioned, many of the missed differences can be attributed to the interpretation of the linguist. In our tools, we have left substantial room for the linguist to interpret results. While the advantage is that the linguist can use any prior knowledge about the language or its family, or more general linguistic expertise that they may possess in order to form more informed hypotheses, this can lead to bias. We have seen this happen in the forming of H11, in which it was hypothesized that Hungarian has vowel harmony. Although it turned out to be correct, the conclusion may have been guided by the author's knowledge

of Finnish, a language related to Hungarian, which also has productive front-back vowel harmony, and was drawn too quickly, as there are other plausible explanations of the vowel alternation that was observed in a few suffixes. Other hypotheses may have been somewhat steep as well, but the interpretation of the output is sometimes difficult, in which case linguistic knowledge can aid the user to arrive at the forming of a hypothesis – whether correct or wrong, a hypothesis should always lead to closer inspection of the data.

It can be argued that the interpretation of the output should be made less subjective, by having a computer interpret (a part) of the results and automatizing the generation of hypotheses. One can think of a list of questions about the target language that a linguist will always ask and the tools should minimally be able to answer, but while it will reduce the subjectivity of the results, one will only get answers to questions directly posed to the algorithm beforehand. That is to say, the algorithm will only discover differences for which it was expressly programmed to look, and the output will only be interpreted by the algorithm in ways it was expressly programmed to do so. We believe that a good balance can be struck between the freedom for subjective interpretation on the one hand and the more computer-driven generation of hypotheses on the other, though whatever the tendency in the balance struck, the expertise and subjective interpretation of the linguist will always be there: either the linguistic bias will be present in the interpretation of the output, or the linguistic bias will be put in the design of the algorithm.

4.5.2 Other remarks on the methodology

Several other factors that influence which hypotheses are or can be formed can be identified, apart from the interpretation of the output. First, it was observed that the choice of source language and target language influences the results, despite the tools having been designed to be language-independent. Due to the unilateral mapping of linguistic annotation from the source language onto the target language based on word alignments, the user may fail to detect any morpho-syntactic features that concern unaligned words in the target language. For example, English allows for the dropping of the conjunction *that* in relative and subordinating clauses. Hungarian, however, does not allow for the dropping of its equivalents *hogy* 'that (conj.)' and *(a)mely* 'that, which', but if English *that* is absent no linguistic annotations are mapped onto *hogy* or *(a)mely* through alignment, and the Hungarian words are in fact completely absent in our tools' output, leading to this difference being undetected. Similarly, differences can remain undetected when a word type and its equivalent in the target language occur in a completely complementary distribution. As a fictive example, it could have been the case that the English infinitival marker *to* only occurred after aspectual verbs, while a Hungarian equivalent infinitival marker only occurred after modal verbs. In that case, the linguist would be led to form a hypothesis such as H33 (that Hungarian does not have infinitival markers at all), because English *to* would never be aligned.

Giving a frequency overview of all unaligned target-language words will most likely not provide further information, because there would be no linguistic information or annotation mapped onto them; the linguist would not know what each word means and in what context it was encountered. It would therefore be very hard to conclude anything about unaligned target-language words, and to form hypotheses about morpho-syntactic differences based on them.

This 'blind spot' could perhaps be remedied in several ways. Choosing two languages that are closely related could maximize the number of words in the source language being aligned, securing a high quantity of linguistic annotation being mapped onto the target language. Similarly, one could argue to choose a source language that is highly analytical, which could ensure that as many words in the target language as possible are aligned to a morpheme in the source language. Yet another remedy would be to run the entire experiment twice, with two different source languages. The right choice of two (or in fact, more) complementary source languages (e.g. one language that has reflexive verbs and one that does not) can diminish the size of the blind spot. We believe the latter remedy is the most straightforward and feasible option when there is no linguistic knowledge of the target language at all.

When linguistically annotated corpora or automatic taggers and parsers for the target language do exist, the linguist can also consider to run the experiment twice, but with the source and the target language swapped. Words in the target language that do not receive an alignment in the first run will be linguistically annotated in the second run, allowing for the linguist to form hypotheses. However, annotated corpora or taggers and parsers for the target language were assumed not to exist for the purpose of this research. Additionally, adding annotations for the target language may have negative effects, especially when the annotations are not perfect: Kroon et al. (2020) report that the quality of the annotations led to noisy, hard to interpret results and to the detection of differences in annotation guidelines.

Secondly, the user chooses a few parameters that are passed to the tools, the choice of which may influence results, as well. For instance, in our experiment the GTI output is suppressed by not outputting partitionings of the data if they are smaller than 1% of their parent partition or if the partition contains fewer than five words. While it is meant to control the overflow of output and to suppress noise, it also can also result in some infrequent phenomena not being retrieved by GTI. One of the issues why M2 (Hungarian noun phrases containing a numeral have a singular head noun) was missed, is this suppressing of the output, as numerals are relatively rare. Only 6073 out of the 737319 tokens in the English Bible were tagged as a numeral, amounting to only 0.8%. This suppression threshold, however, leads to a trade-off, as increasing it may lead to more infrequent phenomena being missed, while lowering it may retrieve more noise, which could increase the number of incorrect hypotheses formed.

A last factor that can influence the results is the matter of the genre of the corpus. As with any linguistic research, our tools and method are subject to

the genre of the input corpus, and can only detect differences that are extant in the data. In the case of English and Hungarian Bibles, it will not be found that Hungarian has a distinct second person singular and second person plural pronoun, a difference with modern English. This is because in the KJB the now somewhat archaic pronoun *thou* is still frequently used for singular, while *you* is exclusive to plural (where *ye* is also abundantly used). Similarly, M6, M7, M8, M9 and M10 were all missed (which all have to do with questions) partly because direct questions are not very frequent in the Bible. A final example of the influence of corpus genre on our results is H32, which expresses our hypothesis that Hungarian has a synthetic passive voice. Dr. Lipták pointed out that modern Hungarian does not have a passive voice at all, but in the Bible, which is written in more archaic Hungarian, there still exists a synthetic passive, making our hypothesis only true for this specific corpus.

4.5.3 Points of improvement and future research

Some specific points of interest for future research and the improvement of results can also be identified. Perhaps the most prominent possible improvement is the implementation of automatic outlier detection. By for instance automatically retrieving combinations of attributes that are unexpectedly frequent, the linguist will be aided by being pointed towards possible differences for which they may not have been looking (e.g. Dutch verbs in a subordinating clause are "unexpectedly" frequently occurring with a crossing with the object when compared to English, directly leading a linguist to Dutch's SOV order in subordinating clauses). In turn, this would increase the number of differences found as well as leave less room for subjective interpretation, which would play into the balance between automation and interpretation discussed above.

On that note, it would be very helpful if co-occurrences of attributes were reported in the output. As of now, our tools only output frequencies of single attributes. While this is already very useful, unusually frequent co-occurrences can lead a linguist to forming more informed hypotheses. In order to suppress the output somewhat, because the number of combinations of attributes quickly explodes, one could perform some statistical test and only return the most statistically significant or those that exceed some threshold.

Furthermore, it can be insightful to track adjacencies in the target language. That is to say, the linguist can discover more differences pertaining to (phonological) context or possibly to target-language words that were left unaligned, when the words directly adjacent to the aligned-to word in the target language are also present among the source-language word's attributes. For instance, the difference in usage between Hungarian *a* and *az* 'the' can only be discovered when the word directly following it is somehow accessible in the output of our tools; only then can it be observed that *a* precedes only words beginning with a consonant and *az* only words beginning with a vowel. Moreover, it would allow the linguist to discover that demonstratives and articles must co-occur in Hungarian (M3).

Deriving more information from the dependency tree in the source language can also be beneficial. In our current approach it was already derived that a verb is transitive or intransitive, but it could similarly be derived that a verb is ditransitive, or that it takes a complement in a specific case form, which could lay bare more differences between two languages. Additionally, it could be useful to automatically derive from the dependency tree that a verb is third person when its subject is non-pronominal.

It was observed in the English-Hungarian experiment that our AAA tool may not be ideal for agglutinative languages. While it already retrieved some useful potential affixes in Hungarian, many affixes turned out to be incomplete or noise. We think this may be the case because it was designed only to consider prefixes and suffixes that include the beginning or the end of the word. The result of this is that if suffixes are stacked in the target language – for instance the Hungarian plural marker *-ak* and the inessive marker *-ban* – AAA will calculate an association value between *-ban* and the attribute `(children=(in|case))`, and between *-akban* and the attribute bundle `(children=(in|case), feats=(Number=Plur))`, but not between *-ak* and the attribute `(feats=(Number=Plur))`, thus underrepresenting the frequency and the association value of the plural marker. Ideally AAA also considers affixes that do not necessarily contain the word boundary, as well as even discontinuous affixes (such as the Hungarian superlative circumfix *leg*⟩…⟨*bb*), however the number of affixes to consider would grow exponentially, making the current algorithmic design unfeasible. Given that AAA is already subject to an exponential blow-up as a result of considering all potential attribute sub-bundles, AAA in particular should be improved by increasing its computational efficiency, especially when discontinuous affixes and infixes are to be considered as well.

Another very interesting potential improvement would be to tag adverbs for their type, such as modal, temporal, aspectual or even more detailed. As of now, adverbs are indiscriminately tagged in UD, but distinguishing between different subtypes would make it possible to automatically test the hierarchy of clausal functional projections as proposed by Cinque (1999) with our tools, and to detect any differences in use or relative order of adverbs between the source and the target language.

Similarly, tagging verbs or sentences for aspect would allow our tools to successfully detect the Hungarian coverbs, such as *meg-*, along with associating it with their aspectual attribute.

On the subject of improving tagging and parsing, the used parser model is of course not fully appropriate for use on the Bible. Additionally, any improvements in aligning will benefit the proposed method, as alignments obtained with `eflomal` (Östling and Tiedemann 2016) were far from perfect and newer neural approaches such as SimAlign (Jalili Sabet et al. 2020) only marginally improve on older models in exchange for higher computational requirements and a very steep increase in run-time. However, despite imperfect parses, tags and alignments, we have found many correct hypotheses on Hungarian morpho-syntax,

underlining the power of our method and tools. One could only speculate on the quality and quantity of the hypotheses and detected differences when the corpus were perfectly annotated.

4.6 Conclusion

In this chapter I have explored the possibility of detecting morpho-syntactic differences between an annotated source language and an un-annotated target language by using bitext alignment in order to map the annotation of the source language onto the target language and to derive several morpho-syntactic features of the target language. It was shown that our tools can be used effectively to form many correct hypotheses on differences between English and Hungarian in several syntactic domains and to extract potential affixes in Hungarian. Despite some room for improvement, I believe this research can pave the way for future research towards a pipeline for automated comparative-syntactic research.