”

Utrecht University

Faculty of Humanities

Bachelors degree in Artificial Intelligence

Thesis: Computational Linguistics

Exploring the Correlation between Dialogue Acts and Tense Usage in Germanic Languages

Jack Chen
4427737

Supervisor:
Henriette Swart

Second Reader:
Meaghan Fowlie

Academic year 2022/2023

.

Contents

1 Introduction

Machine translation has made significant progress in the last decade. Services like Google Translate and DeepL Translation have played a significant role in bringing quick and accessible language translation to anyone with access to the internet. These tools provide impressively accurate and rapid translations for all commonly used languages (Wu et al., 2016). And as our dependency on automated translation services and AI language-model chatbot services such as OpenAI's ChatGPT grows (Nicolescu et al., 2022), it becomes critical to continue the development of these systems. The models behind these programs fundamentally rely on a near-perfect interpretation of user input and a strong comprehension of the semantics of the input text. While Chat-GPT and other exciting modern architectures have exhibited Modern Natural Language Processing's amazing capabilities, there is still much more to explore and improve upon. The translation of dialogues for example, as opposed to narrated text, remains a challenging task that requires further improvement (Wang et al., 2016; Li et al., 2016).

1.1 Problem Statement

An important distinction that is just made, is the difference between narrative texts and dialogue. Narrative texts employ descriptive language to provide a 'story-like' experience and express the author's or narrator's point of view. Dialogues, on the other hand, are talks between characters, conveyed in direct speech, expressing character interactions, feelings, and ideas. These two differ not only in their easily noticeable distinctive semantics but also in grammar usage.

Germanic languages are known for their similarities in syntax, morphology, and vocabulary. However, when it comes to the usage of tenses in dialogue, these languages exhibit noticeable differences that have yet to be wholly understood and investigated (Harbert, 2007). While some Germanic languages use tenses to mark temporal relationships between events, others use tenses to convey subtle differences in the speaker's perspective, attitude, and intention.

These differences become the core of several issues that are found when developing systems capable of dialogue translation. The lack of a deterministic understanding of how different languages employ verb tenses to convey the same meaning hinders the establishment of a strong correlation for accurate translation predictions (Baker, 1995). Identifying an external factor that can assist in determining translations of verb tenses more reliably holds the potential to greatly enhance the quality of machine learning training data.

In addition to the aforementioned challenges, the development of dialogue translation systems is hampered by the scarcity of databases containing dialogue in textual form (Wang et al., 2016; Li et al., 2016; Baker, 1995). Transcribing

spontaneous conversations is a time-consuming task, and even with automated methods, substantial manual correction is required, resulting in high costs. On top of that, there is an even greater scarcity of parallelized, truly natural and spontaneous dialogues. This is due to the increasing difficulty of maintaining parallel speech without the need for manual human translation, which is fundamentally subjected to biases (Wang et al., 2016; Li et al., 2016; Baker, 1995).

1.1.1 Time in Translation

This study is part of the Time in Translation project, which aims to investigate the meaning and form of words, sentences, and discourse across closely related languages and dialects (de Swart, 2016). The Time in Translation project, led by Henriëtte de Swart and Bert Le Bruyn, tries to capture the semantics of one of the most difficult tense-aspect categories prevalent across languages: the perfect tense, which is signified in English by the auxiliary 'Have' followed by a past participle.

1.1.2 Research Questions

The goal of this study is to expand the Time in Translation project by studying correlations between communicative function and tense usage between Germanic languages. Recent research within the project had suggested the existence of these correlations (Tellings et al., 2023). Specifically, interest is taken in the Present Perfect tense, which is known to be a synchronically unstable category because it exhibits inconsistent and variable patterns in a specific language system at a specific point in time. By exploring how different Germanic languages employ verb tenses to convey meaning, the project continues the efforts to achieve a comprehensive understanding of tense usage in dialogue.

The scope of the experiment was limited to categorically related Germanic languages in order to improve the validity of conclusions. This method assures that any discrepancies in outcomes provide more evidence for any hypotheses, as extracting conclusions from two dramatically different languages introduces potentially arbitrary factors. English, Dutch, and German were chosen, due to the authors' familiarity with these languages and their availability in the dataset provided by the Time in Translation Project.

Dialogue Act Annotation is an annotation method that represents the communicative function of dialogue text and will be used in this research. It offers several advantages, including its ease of use and wide acceptance. Its user-friendly and intuitive nature enables quick annotation operations, while its broad acceptance ensures the interoperability, comparability, and accessibility of annotated datasets across various study areas and applications.

The annotated data will be analyzed to identify patterns and relationships between dialogue acts, and tenses. By incorporating Dialogue Act into the learning data and observing its impact on verb tense translation accuracy. Rhetorical linkages have been found to have a crucial influence on the temporal perception of narrative texts. In this study, it is explored whether DAA can be the counterpart to that for dialogue.

This research is part of the Time in Translation Project and aims to contribute to assessing potential solutions for improving dialogue translation. The primary goal of the study will be to answer the following questions:

"How does incorporating Dialogue Act Annotation in the learning data impact verb tense translation accuracy, specifically for the use of the Present Perfect tense, in Dutch, English, and German?"

"What is the correlation between tense usage and communicative function in different Germanic languages during spoken interactions?"

1.1.3 Hypotheses

Including a Dialogue Act Annotation in the training data provides additional context regarding the communicative function and intent of the dialogue. By integrating information about the communicative function, the translation model can get a better grasp of the intended meaning behind various verb tenses, which indicate the speaker's communicative objectives. This increased contextual information may result in more accurate translation outcomes.

Therefore, it is hypothesised that incorporating Dialogue Act Annotation into the training data will result in higher verb tense translation accuracy for conversations between Dutch, English and German. It is also hypothesized that there exists a strong correlation between tense usage and communicative function in different Germanic languages during spoken interactions. The hypothesis suggests that specific tense choices in these languages serve distinct communicative purposes, such as expressing attitudes, intentions, or temporal relationships between events.

1.2 Contribution to knowledge

This study aims to achieve advancements in the field of dialogue translation by analyzing the usage of unstable and problematic tenses in Germanic languages such as the Perfect tense. The goal is to discover correlations between a sentence's communicative function and the tense usage of its verbs. Finding these links may aid in gaining a better grasp of the complexities of using difficult tenses. Understanding the relationship between a sentence's communicative intent and

the tense choices made within it allows the opportunity to create more effective dialogue translation systems.

1.2.1 Relevance for Artificial Intelligence

The insights gained from this work will have various practical applications in systems of Natural Language Processing, which is a branch of artificial intelligence that focuses on natural language communication between humans and computers. Namely, the improvement of translation models will benefit from these discoveries. Difficulties with tense processing in languages appear to be an issue that NLP systems are not yet fully capable of resolving. These NLP systems rely on the strength of huge data to circumvent this issue, but their performance is still limited, particularly with dialogue, because dialogue training data is more limited (Wang et al., 2019.)

If it can be shown that including the communicative aspect of sentences in dialogue text through Dialogue Act Annotation enhances tense translation in training data, it could be suggested that using Dialogue Act Annotation during translation will provide comparable results. As a result, if including Conversation Act Annotation improves both tense and conversation translation, it would be advantageous to investigate accurate and automated ways of recognizing and annotating conversation Acts based on these contextual clues. This has the potential to have a significant impact on dialogue translation in language models.

2 Theoretical Background

The research is a continuation of the previous works of two other studies. Namely Isa Buwalda's "Perfect use in dialogue contexts from Harry Potter and the Philosopher's Stone (2020)" and Vera Karssies' "Similarities in spontaneous speech and dialogue in Harry Potter and the Philosopher's Stone (2020)". Both studies compared Harry Potter and the Philosopher's Stone to dialogue from the Switchboard Corpus.

The findings from Buwalda's study, which investigated Perfect use in dialogue contexts from Harry Potter and the Philosopher's Stone, revealed similarities in the distribution of the Present Perfect between the two datasets, suggesting that written dialogue in novels can be used as a proxy for spoken language.

Karssies' study, which also looked for similarities in spontaneous speech and dialogue in Harry Potter and the Philosopher's Stone, looked at overall tense usage rather than just Perfect usage, showing a difference in verb usage in novels and naturally occurring speech. In addition to that, the research concludes that Harry Potter contained more questions than the Switchboard corpus. The research concluded that the differences in verb usage could be attributed to the nature of the corpora, such as longer dialogues in novels that embed bits of storytelling. However, these findings highlight the need to consider the limitation of the Harry Potter corpus when analyzing language usage.

Other data analysis from the Time in Translation project has shown that the parallel database shows a significant amount of variation in the frequency of verb tense usage across English, Dutch and German. In that study, Tellings et al. (2023) suggest that there might be a connection between the distribution of the Present Perfect over declarative and interrogative sentences and its pragmatic use. Using Translation Mining, the authors analyzed the meaning of the HAVE-PERFECT in a parallel corpus of "Harry Potter and the Philosopher's Stone" and its translations and establish the Present Perfect as an indexical tense-aspect category that appears exclusively in dialogue. By linking the proposed information management roles of the Present Perfect to moves in the language game, they find different distributions of its use across sentence types, supporting the existence of a cross-linguistically structure in sequences of Perfect sentences.

2.1 The Perfect form in Germanic Dialogue

2.1.1 The Perfect Form

The perfect tense is a grammatical structure that enables speakers and writers to establish a temporal connection between the past and present. The Present Perfect conveys a past event, with the added current relevance. The perfect tense is often formed in Germanic languages by using an auxiliary verb, such as "have" or "have been," followed by the past participle of the primary verb. This

auxiliary verb denotes the present tense, whereas the past participle denotes the accomplished action or occurrence. The auxiliary verb and past participle combine to form a grammatical structure that is used in dialogue when past actions hold significance in the present situation. By employing the perfect tense, speakers can describe events or activities that have occurred before the current moment but have ongoing effects or relevance to the present.

For example, in English, one could say, "I have studied German for five years." This structure, where the present tense auxiliary verb "have" is followed by the past participle "studied", indicates that the "I" in the statement began learning German in the past and currently possesses knowledge or expertise in the language. This construction is used to emphasize the duration of the study and the resulting proficiency in German.

In Dutch, both "hebben" (have) and "zijn" (to be) can be used as auxiliary verbs for the perfect tense. You can, for example say "Ik heb een boek gelezen" (I have read a book) or "Ik ben naar de winkel gelopen" (I have walked to the store). In both sentences, the speaker shows that they have finished an action by employing the perfect tense, stressing that the action occurred prior to the current instant while having lasting consequences with regard to the present. The choice of the auxiliary verb depends on the verb conveying the action. When the action itself is central, the perfect tense is conjugated with "hebben" (to have). When the change of location or direction (with the intended goal) is central, they are conjugated with "zijn" (to be).

In German, the perfect tense is also formed by combining the auxiliary verb "haben" (have) or "sein" (be) with the past participle. For example, "Ich habe das Buch gelesen" or "Ich bin nach Hause gegangen" (I have returned home). The distinction of auxiliary verb usage is largly identical to Dutch. However, there are some verbs in German that can take either "haben" or "sein" as their auxiliary verb, depending on the context and the intended meaning. For example:

"Ich habe gelegen." *("I have lain.")*
"Ich bin gelegen." *("I have been lying.")*

Both phrases convey a similar meaning, but "Ich habe gelegen" focuses more on the completed action of lying down, while "Ich bin gelegen" emphasizes the resulting state of having been lying down.

2.1.2 The Synchronically Unstable Perfect Tense

One aspect supporting the correlation between tense usage and communicative function in Germanic languages is the synchronically unstable nature of the perfect tense. All three languages employ the Perfect Tense to convey accomplished actions with present significance. However, there exist disparities in the frequency and specific contexts of its use.

For instance, in English, incorrect usage of the present perfect can be observed in sentences such as:

*** "I have eaten a sandwich yesterday."** *(present perfect)*

Many individuals would instinctively perceive this construction as unnatural and would naturally use the simple past tense:

"I ate a sandwich yesterday." *(simple past)*

The mistake lies in combining the present perfect tense, which emphasizes a connection between past actions and the present moment, with a specific time reference like "yesterday." The present perfect is typically used to discuss actions that happened at an unspecified time in the past or have a continuing relevance to the present, rather than stating a specific time frame.

When translating the sentence to Dutch while aiming to preserve the inherent meaning, two valid options emerge:

"Ik at gisteren een broodje." *(ovt)*
"Ik heb gisteren een broodje gegeten." *(vtt)*

Both instances are acceptable and commonly used, with the vtt slightly more prevalent. However, the simple past tense is more extensively employed in everyday conversations in English compared to the present perfect tense.

This study aims to establish a more deterministic approach to distinguish these differences in tense selection when considering communicative function through dialogue acts.

2.2 Dialogue Act

Dialogue Act Annotation (DAA) is a technique used to categorize and identify the different types of dialogue acts that occur in a conversation. Dialogue acts represent the various functions that dialogue serves, such as asking questions, making statements, giving orders, expressing emotions, or conveying information. In the context of this research, the focus of annotation is on distinguishing between declarative and interrogative sentences. By employing DAA to annotate dialogue data, the aim is to gain insights into the potential advantages of incorporating dialogue act information into machine learning translation for dialogues. Under Dialogue Act types, the following types were considered:

Statement-opinion: This dialogue act involves making a statement that includes personal opinion or subjective information. It goes beyond presenting

objective facts and expresses a point of view or evaluation. Examples of statement-opinion dialogue acts include:

"The poor toilet 's never had anything as horrible as your head down it"

"We have much to be thankful for."

Statement-non-opinion: This dialogue act involves making a statement that is presented as a fact or objective information without expressing personal opinion or subjectivity. Examples of statement-non-opinion dialogue acts include:

"No, sir - house was almost destroyed but I got him out all right"

"It was only a dream"

Declarative yes-no-question: These dialogue acts are declarative sentences that seek a yes-or-no response from the listener. They are structured as statements but carry the intention of asking a question. Examples of declarative yes-no-question dialogue acts include:

"Vol- sorry - You-Know-Who was at Hogwarts?"

"Snape was trying to save me?"

Tag yes-no-question: These dialogue acts are short yes-no questions that are added as tags or additions to declarative statements. They serve to seek confirmation or agreement from the listener. Examples of tag yes-no-question dialogue acts include:

"But they were our kind, weren't they?"

"I really don't think they should let the other sort in, do you?"

Yes-no question: These dialogue acts consist of direct questions that require a simple yes or no answer. They are structured as interrogative sentences and typically start with an auxiliary verb or the main verb to form a question. Examples of yes-no question dialogue acts include:

"And did he - did he seem interested in Fluffy?"

"Do you think they'll attack us if we cross the room?"

wh-question: This dialogue act involves declarative sentences that begin with a wh-word (e.g., who, what, where, when, why, how) to ask for specific information. They seek more detailed answers beyond a simple yes or no. Examples of declarative wh-question dialogue acts include:

"How exactly do they sort us into houses ?"

"Who's that teacher talking to Professor Quirrell?"

Rhetorical-Question: This dialogue act involves asking a question not to obtain information but to make a point or emphasize a statement. Rhetorical questions are usually self-evident or have an obvious answer. They are used to engage the listener and make them reflect on the topic being discussed. Examples of rhetorical question dialogue acts include:

"What are you doing?"

"Unless you'd like to tell us and save us the trouble?"

By annotating dialogue data with these dialogue act types, the research aims to provide valuable insights into how incorporating such information into machine learning translation models can enhance dialogue-based translation tasks.

3 Methodology

3.1 Data Collection

All data used in this research stems from a database from the Time in Translation project, which consisted of sentences extracted from chapters 1, 16 and 17 of "Harry Potter and the Philosopher's Stone" in multiple languages. Each dataset in the database focused on one language, paired in parallel with the English original.

For this research, only the Dutch, German, and English databases were used. Each entry in each dataset revolved around a core verb, with various information associated with its sentence. Notably, each sentence containing a verb was annotated to include the core verb's tense in both included languages of the dataset.

3.1.1 Data Preprocessing

To prepare the database for model compatibility, preprocessing steps were performed by using a Python script.

Initially, all entries where core verbs were part of non-dialogue sentences were removed. Harry Potter and the Philosopher's Stone, similar to most written novels, consists of both dialogue text and discourse text. However, as previously discussed in section 1.1, non-dialogue sentences are not relevant to the research and were therefore removed. Whether a sentence belonged to a dialogue or non-dialogue context was established by examining the context of each entry and its target verb (highlighted in the dataset with asterisks). Dialogue sentences were properly transcribed with opening quotation marks, although closing quotation marks were not present in a consistent structure. Therefore, to filter out non-dialogue sentences, the following approach was implemented:

For each entry and its target verb, an iteration through each character in the context sentence was performed, while keeping track of the quotation marks encountered. Once the target verb was reached, the iteration was halted and it was checked whether the target verb appeared to the right of an opening quotation mark. If the number of quotation marks before the asterisk was uneven, the target word was certainly to the right of an opening quotation mark. Furthermore, it could be assumed that the quotation marks were later closed, within the same context sentence or beyond, somewhere after the target word. If the number of quotation marks before the asterisk was even, it was determined that the last quotation marks were closing quotation marks, implying that the target verb was not part of a dialogue phrase. These sentences were removed accordingly.

Furthermore, only entrees with relevant verb tenses that were already annotated in the database were taken into account. For the research, the following

tenses were used: "simple present," "simple past," "present perfect," "present continuous," "past perfect," "past continuous," and "present perfect continuous." As part of the preprocessing steps, any entries associated with other verb tenses were removed.

Finally, all three languages (English, Dutch, and German) were combined into a single database file. The process of merging was based on the unique "source ID" column, which was partially repeated across languages. Certain columns were renamed to improve clarity.

3.2 Dialogue Act Annotation

Following the data preprocessing, the remaining data were annotated according to the existing Dialogue Act Annotation guidelines. In section 2.2 each of these dialogue acts was thoroughly discussed. These are the dialogue act together with their respective annotation code.

- Statement-opinion (sv)
- Statement-non-opinion (sd)
- Declarative Yes-No-Question (qyd̂)
- Tag Yes-No-Question (qyĝ)
- Yes-No-Question (qy)
- Wh-Question (qw)
- Rhetorical Question (qh)

1	source id	source tense	de_tense	nl_tense	DAA - Simple	DAA	DAA - Extra
82	55998	simple present	Präsens	ott	q	qy^g	
83	56000	present continuous	Präsens	ott	nq	sv	
84	56005	simple past	Präteritum	ovt	q	qy	
85	56006	simple present	Präsens	ott	q	qy	o*
86	56017	simple present	Präsens	ott	nq	sd	
87	56022	simple present	Präsens	ott	nq	sd	
88	56043	simple present	Präsens	ott	nq	sv	
89	56050	simple past	Perfekt	ovt	nq	sd	
90	56060	simple present		ott	nq	sd	
91	56069	simple past	Präteritum	ovt	nq	sd	
92	56070	simple past	Präteritum		nq	sd	
93	56073	simple past	Präteritum	ovt	q	qh	y*
94	56074	simple present	Präsens	ott	q	qh	o*

Figure 1: Sample of the annotated data

To ensure the validity and accuracy of the annotation, an inter-annotator agreement was reached with two individuals who annotated 100 entries. In the inter-rater reliability assessment of the annotation process, a high agreement rate of 95% was achieved. When calculating Cohen's kappa, a measure of inter-rater reliability, the result was approximately 0.9. This demonstrates the consistency and reliability of the annotation process by indicating an almost perfect agreement among the annotators. These guidelines ensured consistent and accurate annotation of dialogue acts, allowing for effective analysis of dialogue translation in later stages of the research.

The final dataset contains 883 data points, each consisting of the English verb tense and one or both of the Dutch and German-translated verb tenses.

3.3 Translation Model

3.3.1 Random Forests

Random Forests is a common ensemble learning approach, particularly in machine learning. It is intended to produce predictions by merging the outcomes of several decision trees. This method goes through a process in which it builds a group, or ensemble, of decision trees. Each tree in this group is trained on a random subset of the available training data and uses just a random subset of the available features.

Random Forests tries to introduce variety and reduce overfitting by training each tree on a different subset of the data and considering only a subset of features. The various forecasts of each tree are then pooled or aggregated to generate the Random Forests algorithm's final prediction.

The Random Forests technique was used in this specific application to develop a translation/classification model that focuses on translating verb tenses between Germanic languages, specifically from English to Dutch and German. Model frameworks from the Python package sklearn were used to create this model.

3.3.2 One-Hot Encoding

Depending on the type of information represented, variables in data analysis and machine learning can take several shapes. Numeric variables and categorical variables are the two most prevalent forms of variables. Numeric variables are numbers that can be employed directly in mathematical computations. Categorical variables, on the other hand, are made up of separate values or categories that do not have an intrinsic numerical meaning. This means that categorical variables like colour, gender, or nationality offer a barrier when dealing with machine learning algorithms because these algorithms often require numerical input to be handled effectively.

To address this issue, a commonly utilized approach known as One-Hot Encoding was used to convert categorical data into a numeric representation suited for machine learning algorithms. This conversion is accomplished by producing binary dummy variables for each category within a categorical variable. Each dummy variable relates to a specific category and takes the value 1 if the original variable does, or 0 if it does not.

Certain variables in the context of the particular dataset underwent One-Hot Encoding. This encoding approach was utilized to encode all of the variables necessary to train the model for predicting tense translations. It was necessary to encode the source tense variable when translating English tenses to Dutch tenses. Furthermore, the DAA column was also encoded to evaluate its impact on the prediction model

3.3.3 K-Fold Cross Validation

Machine learning techniques frequently encounter the problem of unintended under- and over-fitting of data. K-Fold Cross Validation is a common strategy for dealing with this problem. This strategy solves the problem by only evaluating performance on data that the model was not exposed to during training. A robust evaluation of the model's performance over the full dataset is done by partitioning the dataset into K subsets or "folds" and completing K 'training and evaluation' iterations.

During each iteration of K-Fold Cross Validation, one set is assigned as the validation set, while the rest are used for training. This method allows for a comprehensive evaluation of the model's performance across multiple data partitions, reducing the influence of data variability and possible bias. The results from all iterations are then combined to provide a final assessment of the model's capabilities.

In this particular study, a value of 5 was used for k during model training. K-Fold Cross Validation proved especially advantageous given the constraints of a smaller dataset. Limited datasets often lead to overfitting, where the model becomes too specialized to the training data, resulting in poor generalization. However, by implementing K-Fold Cross Validation, the available data was effectively utilized by systematically rotating the validation set, producing a more reliable and generalized estimation of the model's performance.

Furthermore, the entire process was performed ten times to increase the dependability of the experiment's findings. This repeated execution was designed to reduce the influence of any potential outlier results, resulting in a more robust and accurate assessment of the model's performance.

3.4 Evaluation Methodology

3.4.1 Selection of Evaluation Metrics

In order to evaluate the performance of the dialogue translation classification model, it was necessary to select appropriate evaluation metrics. The chosen

metrics were designed to capture various aspects of translation quality, ensuring a comprehensive analysis of the model's effectiveness.

Accuracy is a fundamental metric that measures the overall correctness of the model's predictions. It is calculated as the ratio of correctly classified instances to the total number of instances. By assessing accuracy, the model's ability to correctly classify dialogue translations can be determined.

Each model's accuracy is calculated as a whole, but recall and precision (and f1-score) are calculated for individual classes. This is because, when working with non-binary variables like many classes or categories, it is crucial to evaluate the model's performance over all classes rather than just one.

Macro averages provide a way to calculate an average performance metric by considering each class individually and then averaging the results. To calculate macro-averages, the performance metric (e.g., precision, recall) is computed for each class separately. Then, the average of these metrics across all classes is calculated, giving equal weight to each class. This ensures that each class contributes equally to the overall evaluation, regardless of its size or prevalence in the dataset.

Precision is a metric that measures the proportion of correctly classified positive instances out of all instances that were classified as positive by the model. In the context of dialogue translation classification, precision indicates the model's ability to accurately identify and classify dialogue translations as belonging to a specific category. A high precision score signifies that when the model predicts a translation to be of a certain category, it is likely to be correct.

Recall, also known as sensitivity, measures the proportion of correctly classified positive instances out of all actual positive instances in the dataset. In the context of dialogue translation classification, recall signifies the model's ability to identify and capture all relevant translations belonging to a specific category. A high recall score indicates that the model is effectively capturing and classifying dialogue translations of a certain category without missing many instances.

While precision and recall are useful in and of themselves, they do not provide a whole picture of the model's performance. The F1 score combines precision and recall into a single rating that balances the two. It is derived as the harmonic mean of precision and recall and provides an overall evaluation of the model's ability to detect positive examples within a class. The F1 score is a number between 0 and 1, with 1 representing flawless precision and recall.

3.4.2 Statistical Analysis Approach ***

To conduct the statistical analysis and compare the translation performance across different language pairs or model variations, an analysis of variance (ANOVA) test was employed. The ANOVA test allows us to determine if there

are significant differences between the means of multiple groups.

For each iteration of each model version, multiple evaluation results were obtained. These results were collected in a structured manner, ensuring that data points were organized by language pair and model version.

The ANOVA test aims to evaluate the null hypothesis (H0) that there are no significant differences in translation performance across the language pairs or model variations, against the alternative hypothesis (Ha) that at least one of the groups differs significantly from the others.

The ANOVA test involves partitioning the total variation in the data into two components: variation between groups and variation within groups. The test statistic, known as the F-statistic, is calculated by dividing the between-group variation by the within-group variation.

The calculated F-statistic is compared to the critical value from the F-distribution based on the chosen significance level (e.g., $= 0.05$). The resulting p-value is used to make a decision regarding the null hypothesis. If the p-value is below the significance level, the null hypothesis is rejected, indicating that there are significant differences in translation performance between at least two groups.

4 Results

4.1 Exploratory Data Analysis

4.1.1 Dialogue Act & Tense Distribution

The final dataset contains 883 data points, each consisting of the English verb tense and one or both of the Dutch and German-translated verb tenses. The distributions of tense usage in each of the three languages, as well as the frequencies of each dialogue act are shown in Figure 2:

Tense (English)	Count
simple present	507
simple past	175
present perfect	119
present continuous	58
past perfect	8
past continuous	8
present perfect continuous	7

(a) Tense Distribution of the English dataset

DAA (English)	Count
sd	480
sv	231
qw	62
qh	35
qy	26
qyĝ	25
qyd̂	23

(b) Dialogue Act Distribution of the English dataset

Tense (Dutch)	Count
ott	470
ovt	105
vtt	81
imperatief	8
ovtt	6
vvt	6
infinitief	5
ottt	2

(c) Tense Distribution of the Dutch dataset

Tense (German)	Count
Praesens	390
Perfekt	91
Praeteritum	58
Konjunktiv	8
Plusquamperfekt	2
Imperativ	2
Futur I	2
Konjunktiv II	2
Partizip	1

(d) Dialogue Act Distribution of the German dataset

Figure 2: Dialogue Act & Tense Distribution

4.1.2 Dialogue Act to Tense Correlation

Figure 3 shows a heatmap that was created to analyze the association between dialogue acts and tense usage in order to examine the possibility of dialogue acts as a feature for tense translation. The emphasis will be on the highlighted area delineated by a red border. Positive correlations between value pairs are represented by darker red shades, whilst negative correlations are represented by darker blue shades. A light grey tint will be used to signify neutral or non-existent correlations.

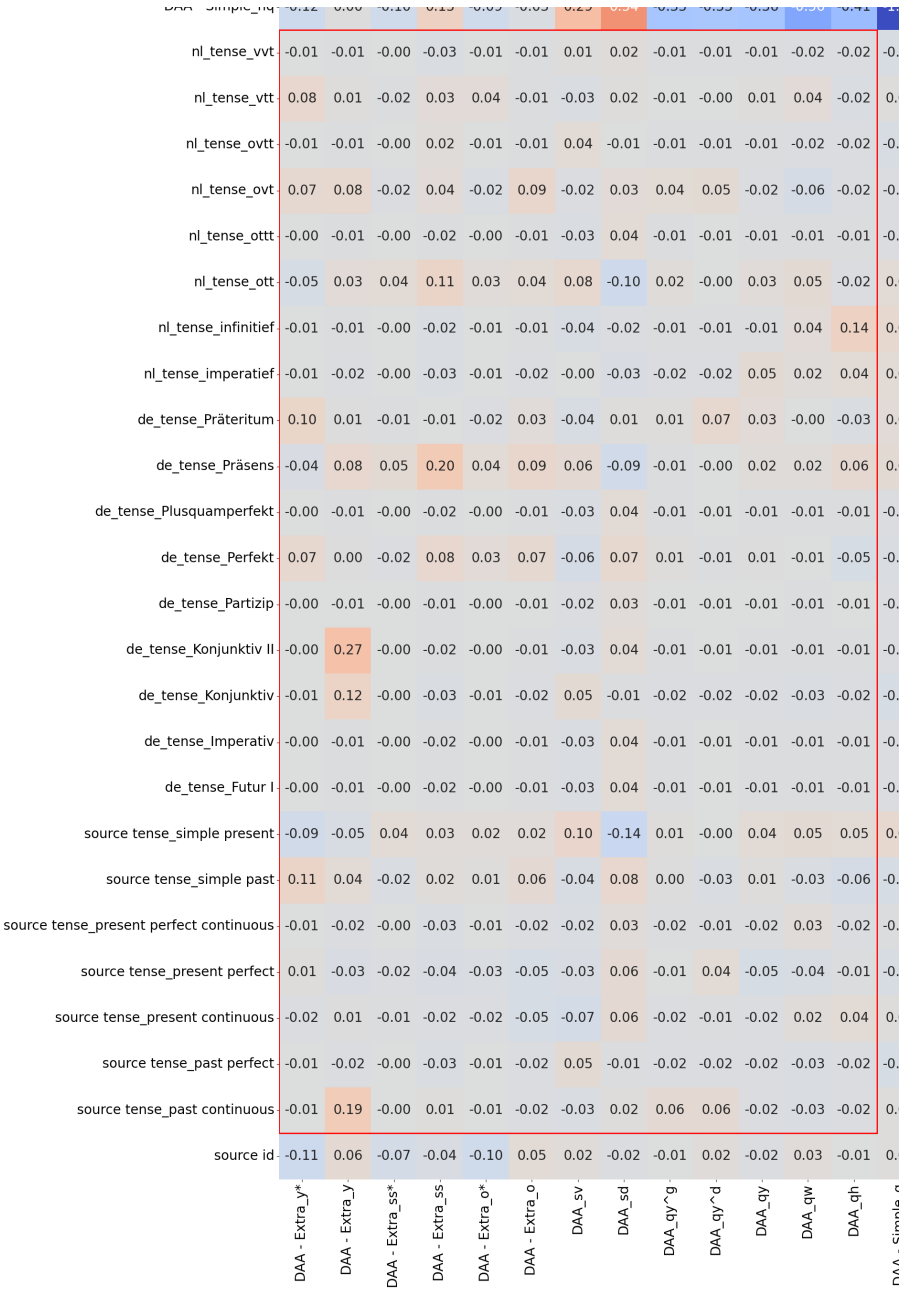


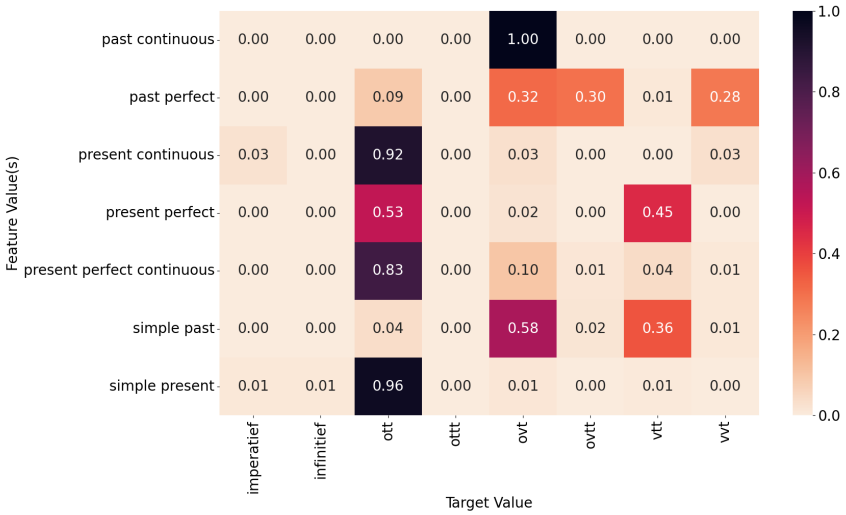
Figure 3: Dialogue Act to Tense Correlation

4.2 Model Predictions

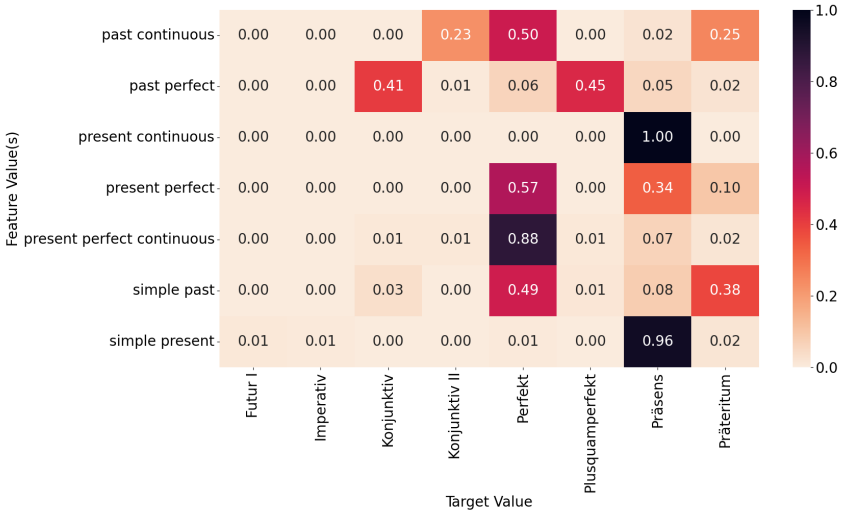
For both English-to-Dutch and English-to-German language pairs, four separate models were built, each with its own set of features. To ensure the reliability of the findings, training and evaluation were conducted using five different folds for each model. Furthermore, to obtain consistent results, each of the eight models was trained and evaluated a total of ten times, and the mean results were calculated subsequently. Figure 4, 5 and 6 show the correlation maps of each model that is meant to represent the probability distribution of the random forest model. Each figure is split into the English-to-Dutch model (a) and the English-to-German model (b). The final/fourth model for both languages is not visually presented due to the complexity of its feature set. Its results from statistical analysis are however incorporated.

4.2.1 Model Predictions: Control

The initial model, as shown in figure 4, acted as a control, incorporating only the source tense (English) as a feature while excluding any additional classification features.



(a) Random forest probability distribution: English-to-Dutch



(b) Random forest probability distribution: English-to-German

Figure 4: Random forest probability distribution: Control

The objective of the upcoming analyses are to observe whether previously deterministic tenses maintain their determinism even when paired with dialogue acts and whether previously uncertain tenses become (more) deterministic when combined with dialogue acts. Particularly, the focus lies on the latter scenario. These English tenses are classified as severely problematic if their best translation prediction exhibits a certainty below 0.65.

For the English-to-Dutch tense translation, the following English tenses are classified as severely problematic:

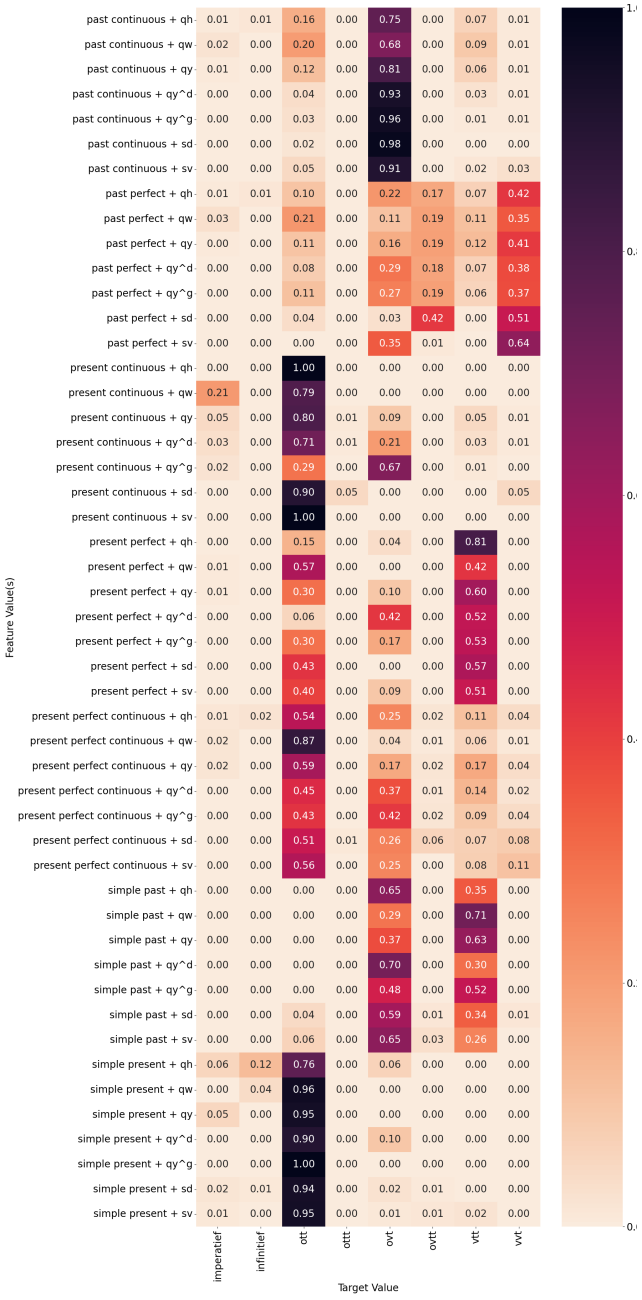
- Past Perfect
- Present Perfect
- Simple Past

For the English-to-German tense translation, the following English tenses are classified as severely problematic:

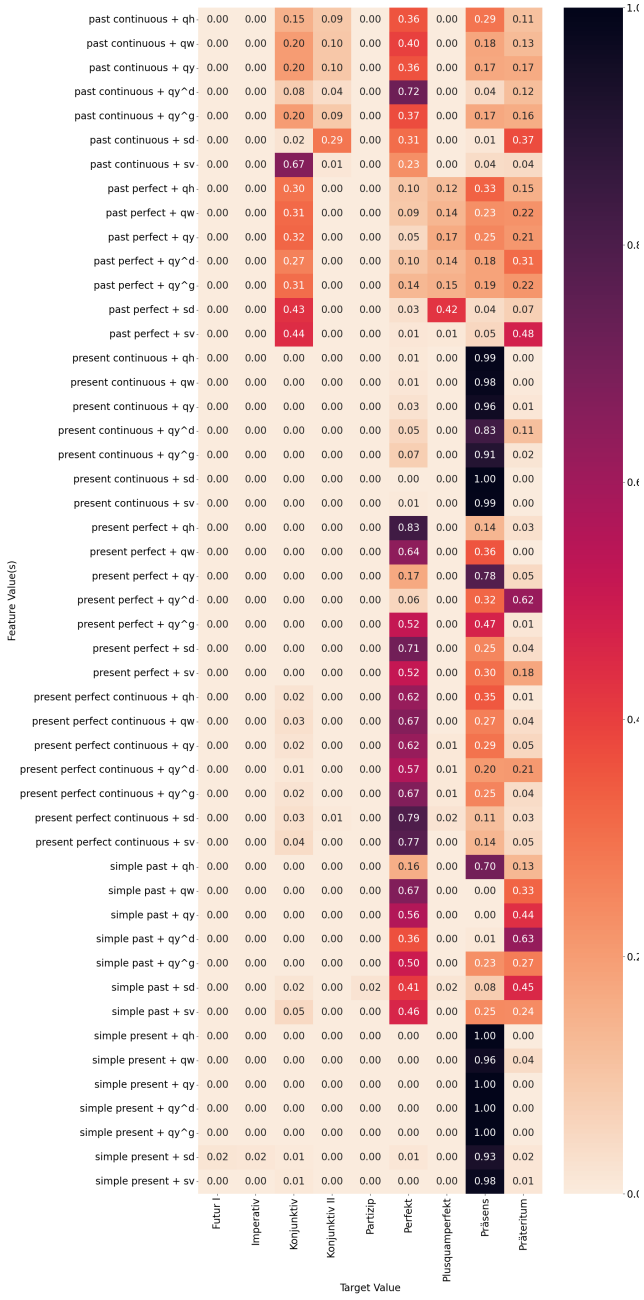
- Past Continuous
- Past Perfect
- Present Perfect
- Simple Past

4.2.2 Model Predictions: DAA

The following two figures, 5a and 5b, depict a similar probability distribution but with the inclusion of dialogue acts in the model. Each row in the table corresponds to a specific pairing of source tense and dialogue act, displaying their respective probability distributions.



(a) Random forest probability distribution: English-to-Dutch



(b) Random forest probability distribution: English-to-German

Figure 5: Random forest probability distribution: Source Tense + DAA

4.3 Adaptations on DAA

The following section discusses variations on the Dialogue Act Annotation process. These modifications were implemented with the goal of addressing hypothesized issues that arose during the annotation process.

4.3.1 Model Predictions: DAA Plus

The annotation process encountered a notable issue related to sentences containing multiple clauses. A clear guideline was lacking when it came to dealing with such sentences. For instance, consider the sentence: "Why do you think he wanted to referee your next match?" This sentence consists of two clauses. The main clause is "Why do you think," and the sub-clause is "he wanted to referee your next match?" Each clause has its own representation in the database, as they contain distinct verbs, often in different tenses.

However, when considering the dialogue act, it becomes uncertain whether the sub-clause should be disregarded. If we ignore it, we categorize it under the dialogue act of the main clause, which feels inappropriate. The main clause, "Why do you think," is a question, while the sub-clause, "he wanted to referee your next match?," is a statement. This research operates under the assumption that tense usage differs when distinguishing between questions and statements.

To address this issue, an additional feature was introduced for each sentence entry containing multiple clauses. This feature aims to identify the role of the main and sub clauses within the entire sentence.

- origin-clause (o)
- sub-clause (y)
- separate sentence (ss)
- special origin-clause (o*)
- special sub-clause (y*)
- special separate sentence (ss*)

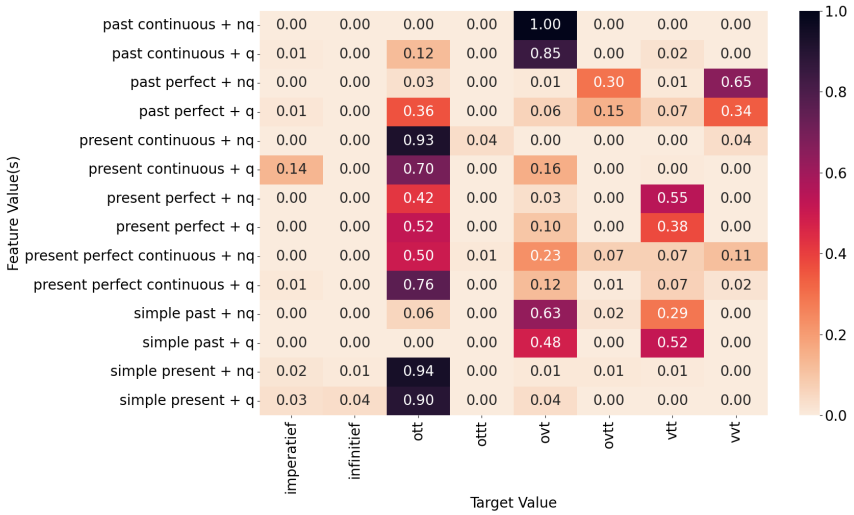
To annotate the main and sub-clauses in the data entries, both "o" and "o*" were utilized when the main verb belonged to the main clause of the sentence. On the other hand, "y" and "y*" were used to annotate the sub-clauses. The label "ss" and "ss*" were applied to all clauses within a sentence if they were grammatically separate and lacked a direct semantic connection. In all three cases The asterisk symbol denoted an additional sub-feature, which indicated a notable distinction in the dialogue act between the main and sub clauses, specifically when one was a question and the other was a statement.

Unfortunately, we currently do not have a probability distribution visualization available for this model, as the figure would be too large to display and unreadable. However, during the evaluation of the performances between models, the performance of this model is included.

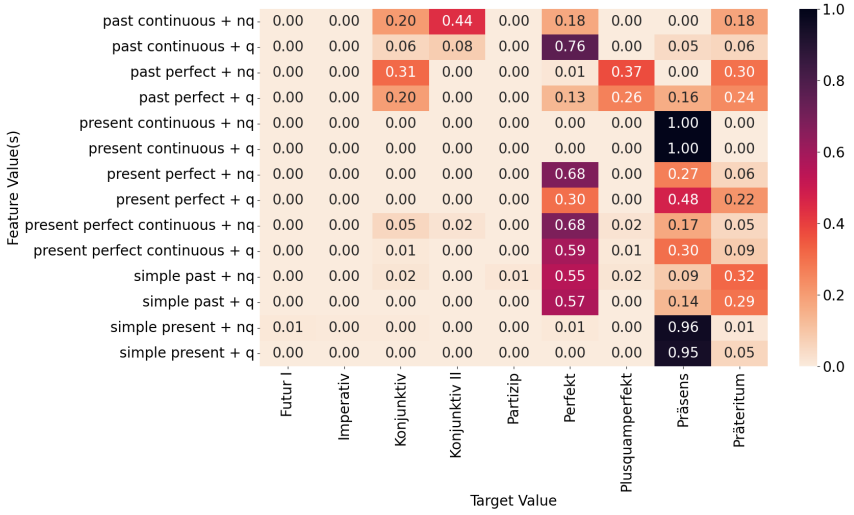
4.3.2 Model Predictions: DAA Simple

Contrasting the more complex DAA variant, a simplified version of the Dialogue Act Annotation was employed to investigate a hypothesis that the complete Dialogue Act Annotation might be overly complex and prone to overfitting the data sample. The following two heatmaps illustrate the probability distribution obtained from this simplified approach.

In this variant, we focus on categorizing dialogue acts solely based on the distinction between questions and statements. If this simplified version yields comparable results to the conventional DAA, it would be highly favored due to its effortless implementation, faster execution, and reduced cost. Ultimately, our objective is to assess the necessity and potential drawbacks of incorporating a more complex Dialogue Act Annotation approach.



(a) Random forest probability distribution: English-to-Dutch



(b) Random forest probability distribution: English-to-German

Figure 6: Random forest probability distribution: Source Tense + DAA-Simple

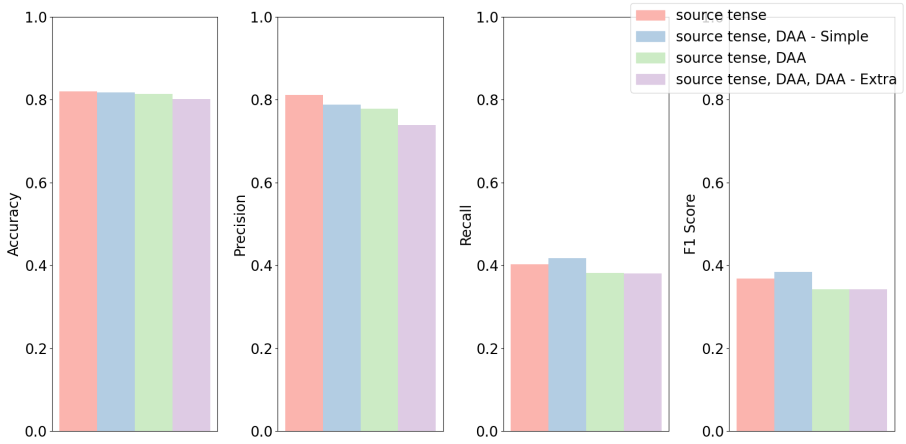
4.4 Model Performance

In contrast to the initial correlation heatmap (Figure 3) that displayed discouraging results by solely pairing dialogue acts with tenses, certain improvements are observed that suggest a potential positive impact of dialogue acts as a feature. Notably, in both languages, the tenses identified as highly problematic show some enhancements in terms of determinism. Although the results still exhibit a level of

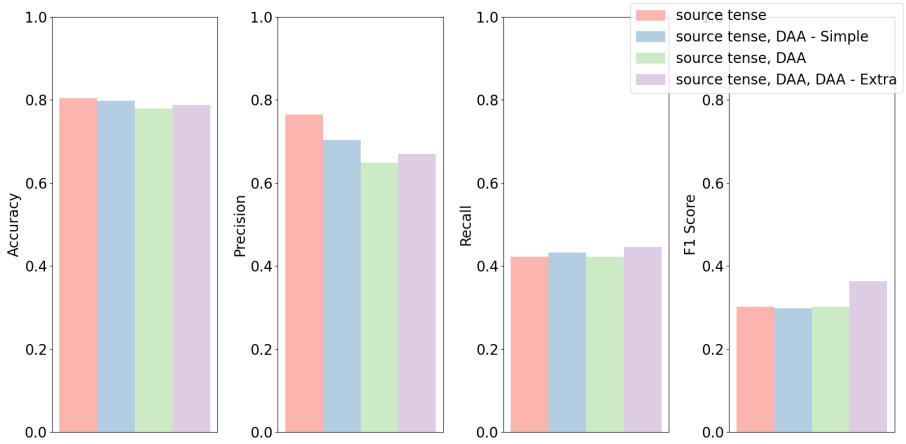
uncertainty, subtle patterns begin to emerge when the two features are combined.

4.4.1 Evaluation Metrics

Examining these models in isolation is insufficient. While they provide visual insights into the internal mechanisms of the trained classification model, they do not directly indicate its performance. The practical results need to be analyzed to assess the performance, using scoring metrics obtained by running the designated test sets through the models.



(a) Model Performance Comparison Source to NL



(b) Model Performance Comparison Source to DE

Figure 7: Model Performance Comparison

4.4.2 Significance

The following tables present the ANOVA results for four different tense translation/classification models. The first and third table show the analysis of variance for the accuracy scores of the English-to-Dutch and English-to-German tense translation/classification models. The second and fourth tables provide similar information but focus on the F1 scores for the same translation/classification models. These tables provide valuable insights into the statistical significance and variability of the different tense translation/classification models across the two languages.

Source	sum$_s q$	df	F	PR(F)
C(Group)	0.001637	3.0	18.889936	1.576992e-07
Residual	0.001040	36.0	NaN	NaN

(a) ANOVA Results: English-to-Dutch - Accuracy

Source	sum$_s q$	df	F	PR(F)
C(Group)	0.013864	3.0	5.445415	0.003424
Residual	0.030553	36.0	NaN	NaN

(b) ANOVA Results: English-to-Dutch - F1 Score

Source	sum$_s q$	df	F	PR(F)
C(Group)	0.001844	3.0	9.861896	0.000069
Residual	0.002244	36.0	NaN	NaN

(c) ANOVA Results: English-to-German - Accuracy

Source	sum$_sq$	df	F	PR(F)
C(Group)	0.042162	3.0	13.312372	0.000005
Residual	0.038006	36.0	NaN	NaN

(d) ANOVA Results: English-to-German - F1 Score

Figure 8: ANOVA Results

5 Discussion

5.1 Exploratory Data Analysis

5.1.1 Dialogue Act Tense Distribution

As seen in Figure 2, the dataset is clearly imbalanced, as certain tenses are significantly more prevalent than others across all three languages. Similarly, the dataset's distribution of dialogue acts is very skewed, indicating the occurrence of both rather rare and more common dialogue acts. These insights on tense distribution may not directly contribute to answering the research question, but they provide a comprehensive overview of the filtered dataset and reaffirm observations made in previous works of the Time is Translation project (Buwalda. 2020, Karssies. 2020).

Upon examining the "problematic" tenses identified in section 4.2.1, it becomes apparent that the majority of instances in our dataset do not actually fall within this 'problematic' tenses category. Therefore, they are not really relevant to addressing our established problem. This has the effect of both skewing the learning model as well as the resulting evaluation metrics. This should be taken into account when making conclusions off of the results.

5.1.2 Dialogue Act to Tense Correlation

The heatmap (Figure 3) analysis of the correlation between dialogue acts and verb tenses aimed to investigate the potential of dialogue acts as a feature for tense translation. The heatmap's weak grey colouring is immediately noticeable, reflecting a pessimistic perspective on the potential of dialogue act and its use as a feature for tense translation. There is no dominantly positive or negative correlation between any of the dialogue acts and any tense across all three languages. This is a preliminary sign that suggests a lack of improvement in the determinism of tense translation through dialogue act alone. Despite this discouraging result, there is still a chance that the dialogue act will show correlations when combined with additional features.

This finding suggests that dialogue acts alone may not be sufficient to improve the determinism of tense translation. The absence of clear patterns in the correlation heatmap indicates that other factors or features might play

a more significant role in tense translation. However, it is worth noting that when combined with additional features, such as the source tense, dialogue acts might still show connections that can enhance the translation process. Further exploration of additional features and their interactions with dialogue acts is recommended.

5.2 Model Predictions

5.2.1 Model Predictions: Control

A correlation between the source tense and the translated tense can be seen by analyzing the probability distributions represented in the figures 4a and 4b, which illustrate the control model that depended only on the source tense as a feature. It becomes evident that certain tense translation classifications are straightforward due to their deterministic nature. For instance, the simple present tense exhibits a deterministic pairing in both English-to-Dutch and English-to-German translations. In contrast, tenses that offer difficulties in translation can also be identified because they lack a clear relationship, such as the present perfect tense. Furthermore, while some English tenses are deterministically translated from English to Dutch, this determinism does not apply when translating from English to German. This discrepancy is particularly noticeable in the case of the past continuous tense, which further indicates the differences in synchronic instability of tenses in these languages.

5.2.2 Model Predictions: DAA

From the probability distribution maps in Figure 4, some improvements in the determinism of tense translation can be seen from the inclusion of dialogue acts as a feature in the models. Previously problematic tenses exhibited enhanced determinism when combined with dialogue acts, although their uncertainty still largely persisted.

In light of the uncertainties surrounding these predictions, it is crucial to exercise caution when drawing conclusions based on assumptions. Our dataset, as demonstrated in section 4.1.1, is limited in size and exhibits a skewed distribution. Consequently, even slight modifications to the training set may result in the emergence of overfitting tendencies that display positively in the probability map, as the probability map does not take into account the number of predictions trained or tested. Due to the uneven distribution of the dataset with 883 instances, which is further broken into around 50 unique combinations, when employing Source Tense and DAA, certain combinations will only occur a few times if at all within the whole database.

5.3 Adaptations on DAA

5.3.1 Model Predictions: DAA Simple

The outcomes obtained from the simplified model and simplified dialogue acts demonstrate moderate improvements compared to the control model (Figure 6). They are not as substantial as the advancements achieved with the complex dialogue acts. The observed correlations fall somewhere between the control model and the advanced dialogue act model.

In both English-to-Dutch and English-to-German translations, the simplified dialogue act annotation does not reveal any substantial indications of heightened determinism in the classification of the present perfect. Although the limited and biased database could, again, potentially affect the model's performance, this concern should be mitigated to a large extent by the model's simplicity, which only has around 15 feasible combinations.

5.4 Model Performance

5.4.1 Evaluation Metrics

The performance of all models, regardless of the dialogue act annotation variation used, was deemed unsatisfactory based on the scoring metrics. In general, as the models became more complex, their effectiveness diminished, leading to lower accuracy, precision, recall, and f1-scores (Figure 7). These findings emphasize the difficulties in accurately translating verb tenses across different languages and underscore the necessity for additional research and advancements in model development.

The DAA Plus model demonstrates a noticeable decrease in overall accuracy when additional features are incorporated. However, there is a significant improvement in the f1-score specifically for the English-to-German translation classification (Figure 7b). Similarly, the DAA - Simple model shows minor enhancements in its F1 score for English-to-Dutch translation (Figure 7a). These improvements in the macro-average f1-scores can be attributed to their above-average macro-average recall score. This indicates that the models are effectively identifying a larger proportion of relevant instances, thereby increasing their ability to correctly classify data points belonging to different classes.

Nevertheless, considering the aforementioned factors once more, it remains important to consider that our dataset is insufficient to accommodate certain intricate features effectively. With a mere 300 potential combinations using the features of the DAA Plus model, any differences compared to the other models could be considered arbitrary. In contrast, the enhancements of the DAA - Simple model appear to be more realistic and warrant further investigation.

5.4.2 Significance

The results obtained from conducting the ANOVA tests (Figure 8) revealed a statistically significant decline in the performance of all three DAA models when compared to the control model. Furthermore, this decline was observed in both English-to-Dutch (Figure 8ab) and English-to-German (Figure 8cd) tense translation classification tasks. By subjecting the data to statistical analysis, it became evident that the observed decrease in performance was not due to mere chance or random variation.

6 Conclusion

6.1 Hypotheses & Research Questions

The main objective of this study was to investigate the impact of incorporating Dialogue Act Annotation (DAA) in the learning data on verb tense translation accuracy, specifically for the use of the Present Perfect tense, in Dutch, English, and German. Additionally, the study aimed to explore the correlation between tense usage and communicative function in different Germanic languages during spoken interactions. The following research questions guided the investigation:

1. How does incorporating Dialogue Act Annotation in the learning data impact verb tense translation accuracy, specifically for the use of the Present Perfect tense, in Dutch, English, and German?

2. What is the correlation between tense usage and communicative function in different Germanic languages during spoken interactions?

Results from the analysis revealed surprising findings that contradicted the initial hypotheses of the study. Contrary to expectations, the incorporation of Dialogue Act Annotation (DAA) in the learning data did not yield any significant improvements in verb tense translation accuracy, particularly regarding the usage of the Present Perfect tense (or other tenses) in Dutch, English, and German. The absence of any discernible impact raises questions about the effectiveness of DAA as a method for enhancing tense translation skills.

Moreover, the study aimed to explore the correlation between tense usage and communicative function in different Germanic languages during spoken interactions. However, the analysis did not reveal any significant correlation between tense usage and communicative function across the languages examined. This unexpected result challenges the assumption that tense usage is strongly linked to the communicative intent or function in spoken interactions.

6.2 Implications and Significance

One possible explanation for the lack of improvement could be attributed to the complexity of verb tense usage and translation. The Present Perfect tense, in particular, poses challenges for language learners due to its multifaceted nature and nuanced distinctions compared to other tenses. The incorporation of DAA may not provide sufficient guidance to accurately translate and appropriately use this tense in different communicative contexts. It is possible that the complexity of tense usage requires more advanced linguistic resources beyond DAA to facilitate accurate translation.

The absence of a clear correlation between DAA and tense usage suggests that other factors, such as cultural influences, individual speaker preferences, or specific lexical choices, may play a more significant role in determining tense usage during communication. It is possible that the complexity and variability

of communicative functions, combined with the flexibility of tense usage in Germanic languages, contribute to the absence of a direct correlation. Further research is needed to explore these factors and their impact on tense selection in different communicative contexts.

In light of the conclusions drawn from the study, it is crucial to emphasize the incompleteness of the research itself. The study relied on a constrained dataset in the context of complex classification models, which posed significant limitations. Nonetheless, it is crucial to recognize the presence of numerous unexplored viewpoints and methodologies that could been explored using the used database. These limitations and unexplored avenues will be elaborated upon in the subsequent sections and highlight the potential for future research to delve more profoundly into the subject matter.

6.3 Limitations of the Study

Despite the significant findings presented in this study, it is crucial to acknowledge the limitations that exist. Primarily, as extensively discussed earlier, the dataset utilized exhibited strong biases and was relatively small in size. These factors could potentially compromise the generalizability of the study's results. The imbalanced distribution of tense usage and dialogue acts within the dataset might have introduced a skewed learning pattern to the model, subsequently impacting its overall performance. For instance, the present perfect tense, which was among the most frequently occurring tenses, and one of the tenses most interesting to the research, only appeared 119 times in the dataset, other tenses were represented in single digits. Consequently, any conclusions drawn from this study are highly susceptible to overfitting, given the limited occurrences of these tenses.

Another potential issue is the source of the dataset. Using dialogue from Harry Potter and the Philosopher's Stone (2020), a fictional book, and its translations, as the dataset, can cause various issues. Firstly, focusing solely on dialogue from one book introduces bias, as it relies on the language choices of a single writer. Moreover, translations of literary works are subject to interpretation and can deviate from the original text. Specifically, in the case of tense usage, when there are no definite translations for some, such as the present perfect. Additionally, using a work of fiction, especially a fantasy series like Harry Potter, as a dataset carries the risk of misrepresentation of grammar in real-world dialogue. The narrative being inherently fictional may cause the intricacies of real-world dialogue to not be captured.

Lastly, another limitation is the reliance on annotated data. The accuracy and consistency of the annotations can significantly impact the performance of the trained model. If the annotations contain errors or inconsistencies, the model may learn incorrect patterns or make incorrect predictions based on that flawed information. While an inter-annotators agreement was successfully performed, it did not span the entire dataset. As will be discussed in the following section,

using more simplified annotation systems will be benificial in both cost and accuracy.

6.4 Future Research Directions

This study lays the foundation for future research in the domains of dialogue translation and tense analysis, suggesting several avenues for further exploration. Expanding on specific focus points within this research alone can lead to the discovery of numerous new findings that were beyond the scope of this study.

For example, to streamline the analysis process, future research could exclude verb tenses that are already known to be deterministic, such as the simple present. By focusing on problematic cases and omitting already 'solved' deterministic tenses, researchers can direct their efforts toward addressing the 'problematic' instances and reduce the burden of annotation and analytical requirements.

In order to delve deeper into the influence of Dialogue Act Annotation (DAA) on translation accuracy, future studies could directly compare its effects on actual dialogue translation models, rather than solely verb tense translation models. This comparison would help quantify the direct improvements achieved through the incorporation of DAA. However, this approach would necessitate the use of a simplified DAA system or a substantial increase in data points, which could be costly.

One potential solution to address this issue is to enhance the efficiency of future investigations by developing methods for automatically predicting Dialogue Act Annotation, rather than relying solely on pre-annotated data. This approach would mitigate the challenges associated with having a small and biased database, as it would reduce the reliance on manual annotation.

One effective means of automating this process could involve the use of a simpler dialogue act annotation method, such as the one briefly explored in this research. Designing a machine based annotation system that can distinguishes questions from statements would be would facilitate the utilization of large datasets without significant cost implications.

Lastly, investigating the use and impact of alternative methods for annotating communicative functions could prove valuable. These alternative methods may potentially yield improvements over DAA when it comes to tense translation. Alternatively, researchers could explore entirely different features that exhibit a stronger correlation with tense usage compared to communicative function annotations.

By focusing on challenging cases, streamlining analysis processes, and exploring alternative annotation methods, future research endeavors hold the potential to unveil new discoveries. Such revelations could significantly enhance the translation of problematic tenses, such as the present perfect, or unearth new insights on semantic correlations of tense usage.

References

[1] Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research.

[2] Harbert, W. (2007). The Germanic Languages. Cambridge, UK: Cambridge University Press.

[3] Nicolescu, L., Tudorache, M. T. (2022). Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review. Electronics, 11(10), 1579. DOI: https://doi.org/10.3390/electronics11101579.

[4] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

[5] Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., Jurafsky, D. (2016). Adversarial Learning for Neural Dialogue Generation.

[6] Wang, L., Zhang, X., Tu, Z., Way, A., Liu, Q. (2016). Automatic Construction of Discourse Corpora for Dialogue Translation.

[7] Buwalda, I. (2020). Perfect use in dialogue contexts from Harry Potter and the Philosopher's Stone (Unpublished bachelor's thesis). Utrecht University. Retrieved from https://studenttheses.uu.nl/handle/20.500.12932/35158

[8] Karssies, V. M. (2020). Similarities in spontaneous speech and dialogue in Harry Potter and the Philosopher's Stone: How does the tense use in Harry Potter and the Philosopher's Stone compare to the tense use in the Switchboard corpus? (Unpublished bachelor's thesis). Utrecht University. Retrieved from https://studenttheses.uu.nl/handle/20.500.12932/36485

[9] de Swart, H. (2016). Perfect usage across languages. Questions and Answers in Linguistics, 3(2), 57-61.

[10] Van der Klis, M., Le Bruyn, B., De Swart, H. (2022). A multilingual corpus study of the competition between PAST and PERFECT in narrative discourse. Journal of Linguistics, 58(2), 423-457. DOI: https://doi.org/10.1017/S0022226721000244.

[11] Tellings, J., Fuchs, M., Van der Klis, M., Le Bruyn, B., De Swart, H. (2022). Perfect variations in dialogue: a parallel corpus approach. Proceedings of SALT 32, Tellings. DOI: https://doi.org/10.3765/salt.v1i0.5342.

[12] Wang, A., Cho, K. (2019). BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. DOI: https://doi.org/10.48550/arXiv.1902.04094

[13] Hans Reichenbach. *Elements of Symbolic Logic*. London: Dover Publications, 1966.

A Appendix

Source Code: https://github.com/JHunter101/thesis-2023