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?

Vera Marissa Karssies 5884128

Utrecht University

Bachelor Artificial Intelligence
Utrecht University
Supervisor: Henriette de Swart
Second corrector: Martijn van Ackooij
7,5 ECTS

Contents

1 Introduction

As the reliance on automated translations, such as Google Translate, and the reliance on automated customer support, via chatbots for example, grows it becomes more important to have a good grasp on how people use language.

A lot of research has been conducted to better understand the use of tenses. For the verbal tense Present Perfect no consensus of its definition has been reached among linguists. The research group Time in Translation looks at the different use of the Present Perfect across languages, aiming to find a fitting definition for this tense. For their research they have used different corpora in various languages. One of the corpora that has been used is *Harry Potter and the Philosopher's Stone* (which I will refer to as HP from now on). This book has been used because its dialogue seems to represent spontaneous speech. Another reason HP has been used for research is that it has been translated into numerous language making it possible to compare the tense use in these languages. Time in Translation found that the Present Perfect in HP only occurs in dialogue and not in narrative discourse, which is in line with theories about the Present Perfect. This insight made that they focused on investigating dialogue sentences in HP to study the use of the Present Perfect (Le Bruyn, Van der Klis & De Swart, 2019). Dialogue is the part where characters talk among each other. Conducting research into the Present Perfect using the dialogue from HP means that HP has been used as a proxy of a natural conversation, I will elaborate on this further in chapter 3. This assumption gives rise to the question of whether or not the tense use in HP resembles the tense use in spontaneous speech. And that is exactly what I will be focusing on in my research. I will investigate whether or not the dialogue use in HP is similar to spontaneous speech. To do so I will use the Switchboard (SB) corpus, a corpus consisting of 5-minute phone conversations. The results of the comparison between the HP corpus and SB corpus will be discussed in this paper.

The paper is structured as follows. In chapter 2 I will elaborate on why this subject is relevant within Artificial Intelligence. Thereafter I will discuss previous research in the verbal tense Present Perfect in chapter 3. Subsequently, I will use this information to discuss my research question and hypothesis in chapter 4. In chapter 5 I will describe the methodology used to conduct my research. After that I will present the results of my research in chapter 6, followed by an analysis of the data in chapter 7. Thereafter, I will elaborate on these results and what they indicate in the chapter 8. And finally, in chapter 9, I will draw up my conclusion, answer my research question and propose possible further research in the conclusion section.

Now that I have given a brief introduction of my bachelor thesis, I will describe the contribution of this thesis to the field of Artificial Intelligence.

2 Relevance for Artificial Intelligence

This paper is written as a bachelor thesis in Artificial Intelligence (AI). Therefore I would like to shed a light on the AI relevance and contribution to AI of this research in this section.

As noted before, a better understanding in tense use can help improve tools such as online translations and chatbots. This might be the first relevance to AI a person might think of in relation to this research, but there is more to it than that. There are different defenitions of Artificial Intelligence around, one proposed definition by Russel Norvig (2010) defines AI using four different categories. One of these categories is thinking humanly, to illustrate this category the Turing test is used. Alan Turing is considered to be the founding father of artificial intelligence. He proposed a test to see if a computer or machine is truly intelligent or not; the Turing test. The Turing test requires a person chatting with a computer and a person simultaneously. The computer passes the test if the person interacting with the computer is unable to tell the difference between a person or a computer (Copeland, 2000). If this is the case, the computer is considered to be intelligent. To pass such a test the computer will need a set of capabilities. These capabilities are: natural language processing in order to communicate, knowledge representation to store information, automated reasoning to form logical conclusions and machine learning to be able to detect patterns and adapt to circumstances (Russel Norvig, 2010). This research can contribute to natural language processing by improving language models and it can help machine learning take place in less amount of time. I will now elaborate further on how exactly this research can contribute to AI in this manner.

This research can contribute to passing the Turing test. This research is conducted to find out whether the dialogue act in Harry Potter is the same as naturally occurring dialogue (spontaneous speech). To analyze spontaneous speech a corpus containing spontaneous speech is needed. To acquire such a corpus spontaneous speech needs to be secured, by recording and transcribing the conversation. This can take up quite some time as this is usually done by hand. In some cases annotation is needed to further analyze the corpus, which again is mostly done by hand and can take up a lot of time. For this research the Harry Potter corpus will be used and will be compared to spontaneous speech. I will elaborate on this further in section 3, but it is also interesting to note that most corpora containing spontaneous speech are not translated into different languages. This makes a corpus such as HP extremely useful. It will save time and effort as no transcription is needed but in addition to that this corpus makes it possible to study the differences in spontaneous speech in across languages as the HP corpus has been translated. For the spontaneous speech the Switchboard corpus will be used. The Switchboard corpus has been used to create the SWBD-DAMSL annotation system, which I will explain more about in chapter 3. One of the aims of creating this annotating system was to create automatic utterance-type detectors in order to shorten the time needed to annotate text. This is useful for shortening the time needed to improve language models (LM).

If it turns out that the HP corpus and the SB corpus are similar, then nov-

els such as HP can be used instead of corpora consisting of spoken language. This will make it easier to construct a corpus as the data the corpus consists of would not have to be transcribed. This will make acquiring data to train utterance-type detectors or language models significantly faster. Speeding up this process makes it possible to increase the amount of data that is used and can decrease the time it takes to improve a language model. This way this research can contribute to creating better language models in less amount of time. Improving these language models can help us understand language better. Furthermore, they would enable algorithms to learn language faster by supplying them with more data, which will lead to them becoming more accurate. This improved understanding of language with the use of dialogue from books can in turn help a computer pass the Turing test as language models get better. This would be a significant contribution to the development of AI. This is one of the examples that illustrates how research in natural language can be a great contribution to AI.

Now that the contribution to AI of this research has been discussed I will elaborate on the theoretical background motivating and supporting my research.

3 Theoretical Background

In this section I will elaborate on the theoretical background of my research. Firstly I will discuss verbal tenses, secondly research done on the Present Perfect. Thirdly I will discuss research on the Present Perfect using Dialogue Act Annotation and lastly, I will discuss earlier comparisons between spontaneous speech and text.

3.1 Tenses

For my research I plan on looking at tenses in the English language. I will be looking in greater detail at the Simple Present, Simple Past and the Present Perfect, that is why I will be discussing these tenses in this section. Reichenbach (1947) introduced a system to describe how verbal tenses reflect the temporal structure of the real world. These temporal relations are determined with respect to the point of speech (S), which is the time an utterance is made. In order to be able to describe every verb tense two more points are needed. These points are the point of event (E) and the point of reference (R). The last point, R, was introduced by Reichenbach in order to be able to describe every verbal tense. Thus in this system there are three points in time, S, E and R. There are two types of relationships between these points, the points are either simultaneous or consecutive. These points can be used to describe verbal tenses as follows: for the Simple Present these three points; E, R, and S, coincide. This means that at the time of speech the event is happening and this is seen from the present point of reference. Reference time can be interpreted as the point in time from which something is considered. In the Simple Past the reference time and the event time are in the past, while the speech time is in the present. In the Present Perfect the event point is in the past, and the reference point and speech point are in the present. For each of the described tenses I have provided an example in sentence (1).

(1) a. Lisa walks to school — *Simple Present*
b. Lisa walked to school — *Simple Past*
c. Lisa has walked to school — *Present Perfect*

Point R was an important addition as describing the tenses of verbs was impossible with only points S and E according to Reichenbach (1947). The difference between the Simple Past and the Present Perfect in the Reichenbach system is the point R. In the Simple Past the point R comes before the point S, and in the Present Perfect the point R coincides with the point S. De Swart (2007) describes the Reichenbach system in order to investigate the Present Perfect across languages. While doing so de Swart gives an example which shows the importance of the point of reference which can be seen in sentence (2).

(2) a. Sarah left the party. — *Simple Past*
b. Sarah has left the party. — *Present Perfect*

Sentence (2a) and (2b) both describe a past event, but sentence (2b) retains the importance of S as the point of reference coincides with S. This shows the importance of the introduction of the point R in order to describe verbal tenses.

Now that the three most important tenses for my research have been discussed by means of the points S, R and E and it is clear what the differences are, I would like to discuss some of the research done in the Present Perfect.

3.2 Research on the Present Perfect

In this section I will discuss the research in the Present Perfect relevant for my research. As discussed in the previous paragraph, the verbal tense Present Perfect can be described in terms of points E, R, and S. This uncovers the temporal structure of the Present Perfect. However the semantic definition of the Present Perfect is unclear as the use of this tense differs across languages. For most verbal tenses an exact semantic definition can be found, but no consensus of the exact definition of the Present Perfect has been reached among linguists. The research group Time in Translation looks at the use of the Present Perfect across languages to unveil the semantic and pragmatic definition of the Present Perfect. They have used different corpora for quantitative research. One of these corpora was Harry Potter and the Philosopher's Stone, from which they used chapter 1 and chapter 17.

While investigating the HP corpus cross linguistic variations in the use of the Present Perfect were found. The research on the HP corpus also showed that the Present Perfect does not occur in narrative discourse, but only in dialogue (Van der Klis, Tellings De Swart, 2020), which is in line with the already existing literature about the Present Perfect. Figure 1 shows the distribution of tenses in narrative discourse and dialogue discourse for the English language from the HP corpus that was found with this research.

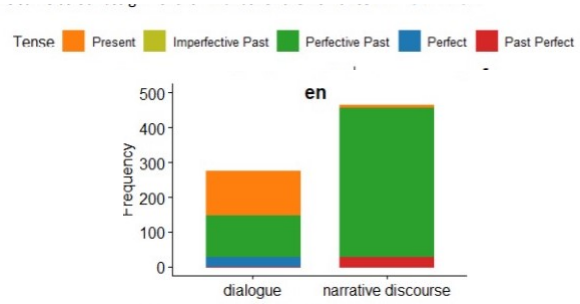

Figure 1: Tense use in HP

As is shown in figure 1 the most frequently used tenses in dialogue are the Simple Present, Simple Past, Present Perfect and lastly the Past Perfect. Because they concluded that the Present Perfect is only used in dialogue, they decided to focus on dialogue in HP for their research. This means that an assumption has been made that the dialogue in HP resembles naturalistic conversation.

This assumption was made in order to conduct research on the Present

Perfect. To conduct research on the Present Perfect conversations need to be studied, because as we just saw the Present Perfect only appears in dialogue. When investigating dialogue across languages the options are limited, as conversations that are transcribed usually are not translated into other languages. Although some corpora such as the Bible corpus, the Europarl corpus, or the OpenSubtitles corpus they are not necessarily suitable for the research into the Present Perfect (Van der Klis, Tellings De Swart, 2020). The Bible corpus is not suitable due to its religious nature, the Europarl corpus is not suitable because the direction of translation is not always clear and the Subtitle corpus is not suitable because it was not created by professional translators and translators could be limited due to the amount of characters that can appear on screen. Of course this is a simplified explanation of why these corpora were not selected for this research. I could elaborate on this further but I would rather discuss why the HP corpus does seem suitable. The reason that the HP corpus has been by Van der Klis, Tellings & De Swart (2020) selected is that it has been translated into various languages and is therefore an attractive corpus for conducting research. The dialogue is informal and the direction of translation is clear. In addition to that the HP corpus contains both narrative discourse and dialogue. Using the dialogue from HP gives rise to the question whether or not the dialogue in HP does in fact resemble spontaneous speech and is therefore a suitable corpus to investigate the Present Perfect.

The research done by (Van der Klis, Tellings De Swart, 2020) allowed them to confirm that the Perfect only occurs in the dialogue in HP. Thus the Present Perfect does not occur in narrative discourse, but only in actual speech or in text that is resembles spontaneous speech, such as dialogue in books or film scripts. The use of HP to investigate the Present Perfect therefore gave rise to a new question: namely if a novel such as HP can be a good proxy for naturalistic conversation. In order to be able to investigate this an example of spontaneous speech is needed. In the next section I will discuss a way to categorize spontaneous speech which can help compare spontaneous speech to text that resembles spontaneous speech.

3.3 Dialogue Act Annotation

In this section I will discuss the Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation system and other research on the Switchboard corpus.

The Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation system was created by Jurafsky, Shriberg and Biasca (1997) and is used to annotate the function of an utterance in dialogue. Shallow discourse consists of the act type of each utterance and sociolinguistic features such as the expected response certain utterances can give. Deeper conversational knowledge such as goals are not taken into account when looking at shallow discourse. The main goal of creating this annotation system was to use the information gained to improve language models (LM). To create this dialogue act annotation system the Switchboard corpus was used. The Switchboard corpus consists of 1115 5-minute spontaneous phone conversation, these phone conversations were be-

tween people who did not know each other and they were given a subject to talk about. Each utterance got assigned a tag to annotate the function of each utterance. In total 220 different tags were used to label the utterances, because some tags only appeared a few times they were grouped in 42 larger classes. The most frequent occurring tag in the Switchboard corpus is Statement-Non-Opinion, which is labelled as sd, an example can be seen in sentence (3).

(3) a. Me, I'm in the legal department. *sd*

The type sd is constructed as a normal sentence as can be seen above. Less structured utterances are also annotated. The second most occurring tag is Acknowledgement/Backchannel (b), this type of utterance does not necessarily usually lacks semantic content because and does not have a typical sentence structure, this can be seen in the example sentences in (4). Acknowledgement is used as a continuer, this is used when a speaker wants to acknowledge that he or she understands, is listening and wants the other to continue.

(4) a. Uh-huh. *b*
b. Yeah. *b*

The full set of 42 classes can be found in appendix A, including the label, an example, and the number of occurrences of this tag. These tags enable us to classify utterances. This in turn can be used to automatically learn to recognize utterances which can help improve language models. In addition to that this system is a good way to investigate differences in conversations.

Some further research on the Switchboard corpus and its utterances was done by Tellings, van der Klis, Le Bruyn and de Swart (2019), who investigated the tense use in the Switchboard corpus across different categories. For their research they only looked at the main labels of the utterance. The main label is the first letter of the tag assigned to an utterance, in the case of sd, this would be s which represents statements. The labels that were used were Statement (s), Agreement (a), Backchannel (b), Hedge (h), and Question (q). Figure 2 shows the tense distribution across these utterances in the Switchboard corpus.

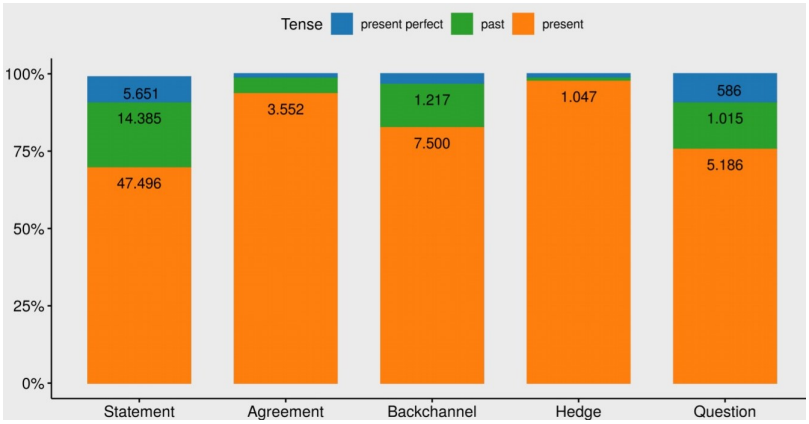

Figure 2: Distribution of tenses in each dialogue act in the SB corpus

As figure 2 shows the tense distribution differs for each label. For example you can see that in agreement and hedges almost only the present tense occurs, whereas the backchannel for example is more past oriented.

Dialogue act annotation is a way to categorize spontaneous speech, this system was developed using the Switchboard corpus. The tense distribution in the Switchboard corpus and the utterance groups have been discussed. Now that we have seen a corpus of spontaneous speech I will discuss earlier research that compares the text to spontaneous speech.

3.4 Comparing text to spontaneous speech

As mentioned before the use of the dialogue in HP as a proxy spontaneous conversation gives rise to the question whether or not this choice is justified. This could be investigated by comparing the HP corpus to spontaneous speech. In this section I will discuss earlier research done that compares spontaneous speech to text.

Levshina (2017) compared English subtitles from OpenSubtitles.org to written and spoken British and American English. Levshina did this using n-grams, n-grams are contiguous sequences of n number of words, in this case 1-grams and 3-grams, so sequences of one and three words were used for this research. She found that the subtitles mainly differed from spoken communication because they were less vague, they contain less sentence reformulations, and less narrative discourse than normal conversation. This probably happens because subtitles are more polished and meant for an overhearer of a conversation, but not somebody actually participating in the conversation. This also happens because the actors are working with prepared text and therefore need less sentence reformulations.

Based on the research by Levshina (2017), Van der Klis, Tellings & De Swart (2020) decided not to use this OpenSubtitles corpus, but as mentioned in section 3.2 they opted for the HP corpus. This was due to the differences between spoken speech and subtitles, but had to do with other shortcomings of the corpus.

Levshina (2017) describes that not all translations were done by professional translators. The translations might be based on other translations instead of the transcribed translations. Another problem for translators was that only a certain amount of subtitles would fit on the screen . Although the corpus was deemed unsuitable for the research in the Present Perfect, this corpus still gives insight in what differences can be expected between spontaneous conversations and written text.

Buwalda (2020) investigated the use of the Present Perfect in Harry Potter and the Philosophers stone by comparing this to the Switchboard corpus that was mentioned earlier. Although the research was unable to conclude whether or not the use of the Present Perfect was the same across both corpora, she did find that a smaller amount of agreement, backchannel and hedge was present in the Present Perfect in HP. This fits in with what Levshina (2017) found, HP is just like subtitles written before hand and therefore less sentence reformulations are needed.

Now that I have discussed the theoretical background leading up to this research, I am ready to formulate my research question. The research discussed in paragraphs 3.2, 3.3 and 3.4 allow me to form a hypothesis for my research. The research question and hypothesis can be found in the next chapter.

4 Research question & Hypothesis

In this section I will discuss my research question, my main hypothesis and my sub hypothesis.

As mentioned in section 3.2 the use of the Harry Potter corpus as a proxy for naturalistic language by Van der Klis, Tellings & De Swart (2020) gives rise to the question whether or not HP resembles naturalistic conversation. Based on this Buwalda (2020) investigated whether the use of the Present Perfect in HP was similar to the use of the Present Perfect in the SB corpus. This research was unable to conclude if this was the case and suggested to investigate all of the tenses in the corpora. This has led me to my research question which is:

- How does the tense use in Harry Potter and the Philosopher's Stone compare to the tense use in the Switchboard corpus?

This research will be done to justify the use of HP corpus in former research, since it is assumed to be similar to spontaneous speech. As mentioned before the Switchboard corpus is made up of phone conversations, and is therefore a corpus with spontaneous speech. Because this corpus consists of informal spontaneous speech it is a suitable corpus to compare to the HP corpus in order to discover if the HP resembles spontaneous speech.

Earlier research by Buwalda (2020) and the first sight of the HP corpus will allow me to form a hypothesis for this research question. Buwalda (2020) researched the distribution of the Present Perfect in the HP corpus and the SB corpus and was unable to find a difference. Therefore based on this research I do not expect to find a difference in the tense distribution across the corpora. When looking at the dialogue in HP this comes across as natural conversation. This is another reason I expect the tense use in HP to be similar to the tenses use in SB. The null hypothesis (H0), which is also the main hypothesis of this research, is formed based on this information:

H0: The distribution of tenses in Harry Potter and the Philosopher's Stone is the same as the tense distribution in the Switchboard corpus.

I expect this hypothesis to be true because at first glance the dialogue in HP seems like natural, spontaneous speech. This is also the reason that this corpus has been used before by the research group Time in Translation. In addition to that Buwalda (2020) did not find a significant difference in the use of the Present Perfect in the HP and the SB corpus, and I expect to find the same for all tenses.

In addition to my main hypothesis I was able to form a sub-hypothesis based on the literature discussed in section 3.4. Levshina (2017) found that in subtitles that less sentence reformulations were needed and the dialogue was less vague. This makes sense, as the script that actors are working from is pre-written and meant for an overhearer and not a person participating in the conversation. The the dialogue in HP is pre-written as well which is it would make sense if similar differences were found in this corpus. In fact Buwalda (2020) found a difference in the dialogue acts in HP and SB that is inline with this theory. She found that in the Present Perfect in the HP corpus fewer agreement, backchannel and

hedge acts were found than in the Present Perfect in the SB corpus. According to Jurafsky et all (1997) hedges are used to to diminish the certainty of what a speaker says or what the speaker answers to a question. Agreement refers to the degree of which a speaker accept the proposal or statement made by the other speaker. Backchannel, or acknowledgement (described in section 3.3), is used as a continuer. The fact that these types of dialogue acts occurred less in the pre-written HP corpus than in the spontaneous SB corpus is in line with Levshina (2017) as these utterances are mainly used to clear up conversations, which is not needed in polished text such as novels or subtitles. Based on this I formed my subhypthesis (S1).

S1: There will be a difference in the dialogue acts occurring in the corpora, I expect HP to have less backchanneling, hedges and agreements than the SB corpus.

I expect this subhypothesis to be true due to the different nature of the corpora. The SB corpus is spontaneously recorded and therefore people would need to explain and reformulate their sentences, HP however is pre-written and therefore no rephrasing is needed in this corpus.

I will research whether or not the tense distribution in the HP corpus and the SB corpus is similar. I expect to find a similar tense distribution in the HP corpus and the SB corpus and I expect to find some small differences in the dialogue acts in the corpora. In the next section I will discuss what steps I will have to take to find an answer to my research question.

5 Methodology

In this section I will elaborate on how I aim to find an answer to my research question:

- How does the tense use in Harry Potter and the Philosopher's Stone compare to the tense use in the Switchboard corpus?

As the research question already indicates two corpora will be used to find the answer to my research question. I will now elaborate on the data of both corpora.

As mentioned before the Switchboard corpus consists of 1115 5-minute spontaneous phone conversation, this corpus will be the corpus modelling spontaneous speech. This annotation system is used to label what type of utterance is presented. Because this corpus was used to create this system, the utterances in this corpus are already annotated. The project Time in Translation has already analyzed the tenses used in the HP corpus and the Switchboard corpus. Therefore I have access to the Switchboard corpus with the tenses and dialogue act for each utterance.

The HP corpus consists of dialogue from chapter 1, chapter 16, and chapter 17 from *Harry Potter and the Philosopher's Stone*. In order to enable myself to compare this corpus to the Switchboard corpus I will have to annotate the utterances occurring in the HP corpus. I will not look the full dialogue in *Harry Potter and the Philosopher's Stone*, but only this selection of chapters. This decision has been made because of the amount of data and the time it will take to annotate the utterances.

I will annotate the utterances from the HP corpus in an excel file, figure 3 shows how this a screenshot of this file. The first column shows the id that belongs to the utterance present, the second column shows the tense of the utterance. The third column, DAA, is where the I fill in the dialogue act of the utterance, if this dialogue act has a sub label this is filled in in the column next to it (sub). The fifth column, named first, saves the first letter of the DAA, which I will later use to analyze the main categories occurring in the corpus. The columns w1, w2, w3 and w4 show which verb is selected, the tense in the second column is written down for this verb. Next to these columns the column document is seen, 1.xml means that the utterance comes from the first chapter of HP. All sentences have an id which is shown under sentence id, the target id contains all the verbs from w1, w2, w3 and w4. And lastly the full sentence from the HP corpus can be found under full fragment.

	id	tense	DAA	sub	first	w1	w2	w3	w4	w5	document	sentence i	target ids	full fragment
2	50604	simple pas	sd		s	heard					1.xml	s10.1	heard	‘ The Potters , that 's right , that 's what I *heard* - ’
3	50605	simple pre	sd		s	's					1.xml	s10.1	's	‘ The Potters , that 's right , that *'s* what I heard - ’
4	50606	simple pre	aa		a	's					1.xml	s10.1	's	‘ The Potters , that *'s* right , that 's what I heard - ’
5	50468	imperative	sd		s	Don	't	be			1.xml	s15.4	Don 't be	On the contrary , his face split into a wide smile and he said in a s
6	50356	present pe	sd		s	has	gone				1.xml	s15.5	has gone	Rejoice , for You-Know-Who *has* *gone* at last !
7	50088	present pe	sd		s	have	been	behaving			1.xml	s22.1	have been	‘ And finally , bird-watchers everywhere have reported that the n
8	50089	present pe	sd		s	have	reported				1.xml	s22.1	have repo	‘ And finally , bird-watchers everywhere *have* *reported* that t
9	50097	present pe	sd		s	have	been				1.xml	s22.2	have been	Although owls normally hunt at night and are hardly ever seen in
10	50098	simple pre	sd		s	are	seen				1.xml	s22.2	are seen	Although owls normally hunt at night and *are* hardly ever *seer
11	50099	simple pre	sd		s	hunt					1.xml	s22.2	hunt	Although owls normally *hunt* at night and are hardly ever seen
12	50545	present pe	sd		s	have	changed				1.xml	s22.3	have chan	Experts are unable to explain why the owls *have* suddenly *cha

Figure 3: Excel file with Dialogue Act Annotation

When I am done annotating the HP corpus I will compare the distribution of tenses of the HP corpus and the Switchboard corpus, I will specifically look at the three most occurring tenses in dialogue; being the Simple Present, Simple Past and Present Perfect. In addition to that I will compare the distribution of utterances of the HP corpus and the Switchboard corpus. To compare the two corpora I will use the Chi-squared test, which I will explain more about in chapter 7. The data that has been gathered can be found in chapter 6 and an analysis of the data can be found in chapter 7.

6 Presenting the data

In this section I will present the data I have gathered from the HP corpus and the SB corpus. In order to better understand the data I first want to discuss the relevant dialogue acts from the Harry Potter corpus, followed by presenting the gathered data, then I would like to discuss the grouping of tags and finally I will discuss the inter-rater agreement.

6.1 Relevant utterances in HP corpus

For this set of data I have decided to group the data into main categories, this is the first letter that is used when annotating the utterances. Four categories were detected in the HP corpus, namely: Agreement, Backchanneling/Backwards looking, Questions and Statements. Each category consists of different tags. I will describe which tags were found in the HP corpus and provide an example. Category Agreement (a) consist of agreement (aa) and action directive (ad), a random example from the HP corpus can be found in respectively sentence (5a) and (5b). I will give an example of all the relevant dialogue acts in the same way. The category Backchanneling/Backwards looking (b) is made up of the types: sympathy (by) and back channeling in question (bh) form. The category Statements (s) is made up out of the tags statement-opinion (sv), statement non-opinion(sd). The category Questions (q) consists of the dialogue acts yes-no-question (qy), tag-question (qy∧ g), declarative yes-no-question (qy∧ d), wh-question (qw), declarative wh-question (qw∧ d) and rhetorical-question (qh).

(5) a. Yes , I quite agree. *aa*
b. 'Go to Madam Pomfrey,' Hermione suggested. *ad*

(6) a. 'I know...' he said heavily. *by*
b. See Professor Dumbledore? *bh*

(7) a. You 'd think they 'd be a bit more careful, but no - even the Muggles have noticed something 's going on. *sv*
b. I have one myself above my left knee which is a perfect map of the London Underground. *sd*

(8) a. Did you mention Hogwarts at all? *qy*
b. Howard, isn't it? *qy∧ g*
c. I suppose it was he who told you I'd be here, by the way? *qy ∧ d*
d. Who is it? *qw*
e. You know what everyone 's saying? *qw ∧ d*
f. Why do you think he wanted to referee your next match? *qh*

Now that the relevant dialogue acts have been discussed it is time to have a look at the data.

6.2 Data from the Switchboard corpus and the Harry Potter corpus

In total 967 utterances from the HP corpus have been annotated and will be compared to the Switchboard corpus, which consists of 216309 utterances. To

analyze the corpora the utterances have been grouped into main tags, as has been described in the previous section. The number of times each type occurs and in which verb tense this type occurs has been put in a cross table. Figure 4 shows the data from the Switchboard corpus and figure 5 shows the data from the HP corpus. In the table the labels are the main types of the utterance. For each type it is shown how many times it occurs in total and in how many times it occurs in a certain tense.

Labels	Gerund	Infinitive	Modal	None	Other	Participle	Simple Past	Simple Present	Present perfect	present perfect continuous	Total
"		1	2	5	3		3	11	1		26
%	61	136	298	10219	336	17	553	4417	128	3	16168
^	38	368	195	715	236	12	141	892	57		2654
a	30	351	251	8481	322	6	147	3471	59	2	13120
b	323	699	643	45165	1553	97	1221	7502	347	22	57572
c		33	41	4	12		1	16			107
f	102	237	69	1806	104	3	212	436	103	43	3115
h		9	25	62	46		8	1047	9		1206
n	9	21	60	4879	109	3	148	766	93	3	6091
o	1	17	16	662	5	1	7	108	4		821
q	34	75	262	1166	1033	25	1038	5322	619	32	9606
s	199	488	4152	6082	20631	118	14385	47489	6793	408	100745
t	2	17	10	56	13		30	91	6		225
x	28	90	68	3845	165	10	113	486	43	5	4853
Total	**827**	**2542**	**6092**	**83147**	**24568**	**292**	**18007**	**72054**	**8262**	**518**	**216309**

Figure 4: Table with the number of utterance per tense for the Switchboard corpus

Labels	future continuous	future in the past	future in the past continuous	future perfect in the past	imperative	infinitive	past continuous	past perfect	past perfect continuous	present continuous	present participle	present perfect	present perfect continuous	simple future	simple past	simple present	Total
a		1			74	1								1	1	11	89
b																6	6
q		5			1		2			8		13		8	36	80	153
s	1	16	1	5	18	1	11	9	2	30	1	47	7	68	197	305	712
Total	**1**	**22**	**1**	**5**	**93**	**2**	**13**	**9**	**2**	**38**	**1**	**60**	**7**	**77**	**234**	**402**	**967**

Figure 5: Table with the number of utterance per tense for the HP corpus

Figure 4 and 5 show that some utterances do occur in the Switchboard corpus, but not in the HP corpus. This can for example be seen because type c, which can be found under labels, occurs in the Switchboard table but not in the HP table. Type c refers to the tags commit and offer. Because groups of tags that are present in the SB corpus are not always present in the HP corpus it would be difficult to compare the full table as there would be a lot of blank spaces. Instead the relevant data will be compared using the chi-squared test, which I will elaborate on in the next chapter.

Because the content in every chapter in Harry Potter differs, I have decided to look at the difference across chapters. I suspect that there could be differences due to the nature of the chapters. Chapter 1 is the first chapter and introduces the reader to the story. Chapter 16 has a lot of action. And chapter 17 is the final chapter and looks back on the events in chapter 16. Because chapter 16 has a lot of action and chapter 17 looks back there could be differences in the

tense us in these chapters. In order to analyze this I have created figure 6, which shows a cross table of the distribution of each tense per chapter of the HP corpus.

chapters	future continuous	future in the past	future in the past continuous	future perfect in the past	imperative	infinitive	past continuous	past perfect	past perfect continuous	present continuous	present participle	present perfect	present perfect continuous	simple future	simple past	simple present	total
Chapter 1		10			2		1		1	11		22	6	12	38	70	173
Chapter 16	1	3		1	50	2	4	3		18	1	15		43	59	181	381
Chapter 17		9	1	4	41		8	6	1	9		23	1	22	137	151	413
total	**1**	**22**	**1**	**5**	**93**	**2**	**13**	**9**	**2**	**38**	**1**	**60**	**7**	**77**	**234**	**402**	**967**

Figure 6: Table with the verb tenses for each chapter

Figure 6 shows the frequency of verb tenses for each chapter of the HP corpus. This table also gives insight in how much data is present in each of the chapters. Chapter 16 and chapter 17 contain about the same amount of data, and chapter 1 only contains half the amount of data.

Now that the data that was gathered for the research has been presented I would like to elaborate a bit further on how this data set was constructed.

6.3 Grouping tags

To gather my data I annotated the dialogue acts in the HP corpus, as was mentioned in chapter 5. I did this using the coders manual by Jurafsky et all (1997). When annotating the Switchboard corpus Jurafsky et al. (1997) found that some utterances occurred less than 10 times, which is a very small number for a corpus this big. The authors decided to solve this by grouping them with other tags. Only one of the tags that was grouped in the Switchboard corpus occurred in the HP corpus, namely the explicit performative (fx) which was grouped with the statement-opinion (sv) tag.
The tag fx occurred seven times in the HP corpus, and only twice in the SB corpus. Sentence (9) shows examples of this type of sentence in the HP corpus.

(9) a. Shooting stars down in Kent - I'll bet that was Dedalus Diggle.
b. '... for the best-played game of chess Hogwarts has seen in many years, I award Gryffindor house fifty points.'
c. 'Second - to Miss Hermione Granger... for the use of cool logic in the face of fire, I award Gryffindor house fifty points.'
d. The room went deadly quiet. '... for pure nerve and outstanding courage, I award Gryffindor house sixty points.'
e. I therefore award ten points to Mr Neville Longbottom.

A lot of the occurrence are part of a speech given in HP. It makes sense that this would occur less in a phone conversation as those are very different from a speech. While this tag occurred relatively more in the HP corpus than in the SB corpus I decided to change the fx tag in the HP corpus to sv. I have done this to ensure that both corpora would be constructed in the a similar fashion. This will make it easier to compare the two corpora to each other.
Now that I have discussed some of the choices I made while gathering data I will say something about how reliable this data is based on inter-rater agreement.

6.4 Inter-rater agreement

As this mentioned in section 5 the dialogue acts had to be annotated. In this section I will shortly discuss how reliable these annotations are.

When annotating the HP corpus I asked M. van der Klis, who is affiliated with the project Time in Translation, to do the same. He annotated the first 105 utterances and I compared them to mine using Cohen's kappa. Cohen's kappa measures the inter-rater agreement, and is used to measure which part of the data correctly describes the variables measured. The kappa coefficient goes from zero (no agreement at all) to one (no differences, full agreement). According to Cohen the kappa coefficient can be interpreted as follows (McHugh, 2012): level of agreement with Cohen's kappa 0–0.20: no agreement, 0.21–0.39 minimal agreement, 0.40–0.59, weak agreement, 0.60-0.79 moderate agreement, 0.80-0.90 strong agreement and above 0.90 almost perfect agreement

When looking at the full tags, which are described in section 6.1, a kappa value of 0.69 was found, indicating a moderate inter-rater agreement.
When only comparing the main tags a kappa value of 0.88 was found which indicates a strong inter-rater agreement.

After discussing these differences the utterance got assigned the correct tags, and the things I needed to look out for were explained to me. I therefore believe that after this the inter-rater agreement went up as we discussed and adjusted the differences. As already mentioned the inter-rater agreement when looking at the main tags is strong, indicating that at least 64-81% of the data is reliable.

Now that the data has been presented, it is clear how this data set was constructed, and the inter-rater agreement indicating the reliability of the data has been discussed, it is time to analyze the data by comparing the two corpora to each other.

7 Analyzing the data

In this section I will analyze the data by comparing the HP corpus and SB corpus to each other. Firstly I will explain the significant test I will be using, namely the Chi-squared test. Subsequently I will analyze the tense distribution across the corpora. Followed by the analysis of the distribution of tenses per chapter. Next I will analyze the distribution of utterances across the corpora. And lastly I will zoom in further on the statement utterance and the question utterance.

7.1 Chi-squared test

The Chi-squared test (χ^2-test) is a statistical hypothesis test, which tests the independence of variables. The null hypothesis is that there is no association between the variables in a contingency table (Levshina, 2015). This is tested based on expected and observed frequencies of the variables in the table. The outcome of a χ^2-test always consists of a test statistic, which will be bigger when the difference is bigger, degrees of freedom which is bigger when the table that is being compared is bigger and a p-value, determining if the difference is significant. The p-value is what is the most important for this research, if the p-value is greater than 0.05 than the variables are independent of the event in the column.

Now that a little more is known about the statistical test that will be used I will analyze the data in order to compare the corpora to each other.

7.2 Tense distribution

To compare the tenses in both corpora I first put the number of occurrences in a contingency table for the three tenses that occur the most in both corpora, this can be seen in table 1. Then I calculated the ratio in which these tenses occur, which can be seen in table 2.

	HP	SB
Present Perfect	60	8262
Simple Past	234	18007
Simple Present	402	72054

Table 1: Frequency of tenses

	HP	SB
Present Perfect	0.086	0.084
Simple Past	0.336	0.183
Simple Present	0.578	0.733

Table 2: Ratio of tense occurrences

As can be seen in the table the ratio of the Simple Past is much higher in the HP corpus than in the SB corpus, and the other way around for the Simple Present. This shows us that there might be a difference in tense distribution across the corpora, especially when looking at these tenses. To be sure a statistical test is needed, the χ^2-test will be used.
The following outcome was found: $\chi - squared = 110.7, df = 2, p - value < 0.001$. This shows a p-value smaller than 0.05. This significant difference means that the tense use is dependent on the corpus, and thus is not the same in both corpora.

In figure 9 this difference is visualized using Pearson residuals. A Pearson residual is the difference between an observed frequency and an expected frequency, divided by the square root of expected value (Levshina, 2015). A negative residual indicates that the observed frequency is less than the expected frequency and a positive residual indicates that the observed frequency was more than the expected frequency. In the graph below this is illustrated as the negative residuals go below the x-axis and the positive residuals are above the x-axis. The residuals coloured blue indicate a significant higher residual and the residuals coloured red indicate a significant lower residual. The intensity of the colour show its relative importance, with a more intense colour being a bigger difference. In this graph the thickness of a bar represent the amount of data this bar represents. Thus a thick bar indicates that there was a lot of data, and a thinner bar indicate that there was less data. Because the SB corpus is bigger than the HP corpus the bars of the HP corpus are much thinner than the bars of the SB corpus.

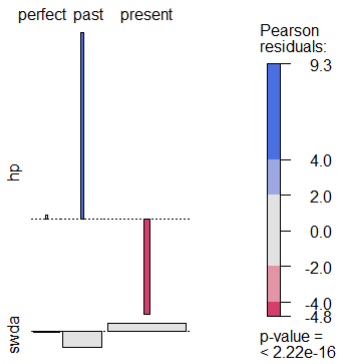

Figure 7: Tense use in HP and SB, plot of Pearson residuals

Figure 7 shows that the HP corpus contains significantly less Simple Present and significantly more Simple Past, no significant difference was found in the occurrence of the Present Perfect.

This section showed that the use of tenses depends on the corpus. It also explains how to read a graph with Pearson residuals, these residuals will continue to occur throughout this chapter in order to expose where the differences

in the corpora can be found if there is a significant difference present. In the next section I will analyze the tenses for the different chapters of the HP corpus.

7.3 Zooming in: Tenses per Chapter

The previous section proved that the HP corpus consist of more Simple Past and less Simple Present than the SB corpus. In this section I will zoom in on the use of tenses in the different chapters of the HP corpus. This way I aim to find out if this difference is due to a specific part of the corpus, or if this tense distribution is the same across the full HP corpus. I will investigate the tense distribution by comparing the tense use in each chapter to the SB corpus.

Table 3 shows the ratio in which the tense occur in the first chapter of HP and in the SB corpus.

	HP chapter 1	SB
Present Perfect	0.169	0.084
Simple Past	0.292	0.183
Simple Present	0.538	0.733

Table 3: Ratio of tense occurrences in HP chapter 1 and SB

This ratio looks similar to the ratio of the full corpus, which was shown in section 7.1. To further investigate this a statistical test is needed. A significant association of the tenses used in chapter 1 was found:
$\chi - squared = 26.338, df = 2, p - value < 0.001$

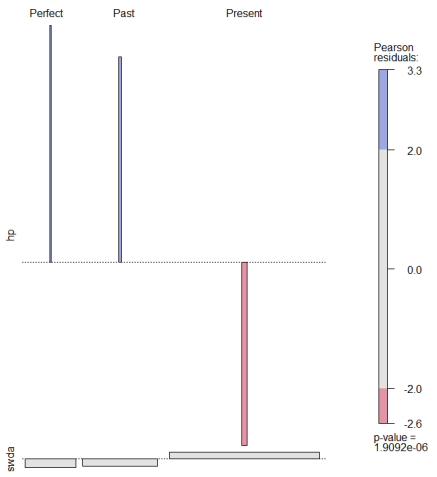

Figure 8: Occurrences of tenses in HP corpus chapter 1 and SB corpus, plot of Pearson residuals

Figure 8 shows that HP chapter 1 contains significantly less Simple Present and significantly more Simple Past and Present Perfect than the Switchboard corpus. Due to the difference in the amount of data of these two sets it was necessary to zoom in on the graph to preserve readability.

Next I will compare chapter 16 to the SB corpus. Table 4 shows the ratio of tense occurrences in HP for chapter 16 and the SB corpus.

	HP chapter 16	SB
Present Perfect	0.059	0.084
Simple Past	0.231	0.183
Simple Present	0.709	0.733

Table 4: Ratio of tense occurrences in HP chapter 16 and SB

The ratio in chapter 16 looks similar to the ratio found in the SB corpus. In order to further investigate this a statistical test is needed. No significant difference was found in the tense use in chapter 16 when compared to the SB corpus: $\chi - squared = 5.3369, df = 2, p - value = 0.069$. Meaning that there is no significant difference in the tense use for these corpora.

Latley I will compare the tense use in HP chapter 17 to the SB corpus. Table 5 shows the ratio in which the tense occur in the chapter 17 of HP and in the SB corpus.

	HP chapter 17	SB
Present Perfect	0.74	0.084
Simple Past	0.441	0.183
Simple Present	0.486	0.733

Table 5: Ratio of tense occurrences in HP chapter 1 and SB

This ratio looks similar to the ratio of the full corpus, which was shown in section 7.1. To further investigate this a statistical test is needed. A significant association of the tenses used in chapter 17 was found:
$\chi - squared = 137.91, df = 2, p - value < 0.001$. Figure 9 exposes where the differences in the tense use can be found.

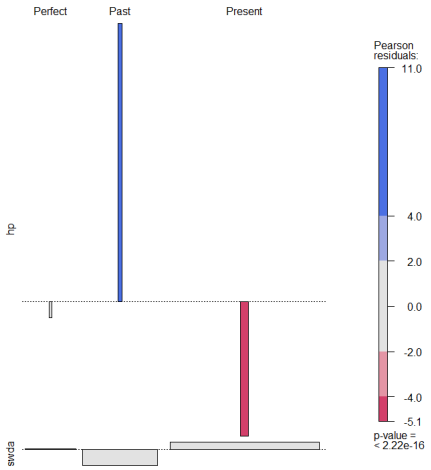

Figure 9: Occurrences of tenses in HP corpus chapter 17 and SB corpus, plot of Pearson residuals

Figure 9 shows that HP chapter 17 contains significantly less Simple Present and significantly more Simple Past than the Switchboard corpus. No significant difference was found in the occurrences of the Present Perfect. This is similar to the tense distribution of the full corpus.

Note that the fact that chapter 1 and chapter 17 have significantly more Simple Past and less Simple Present, and chapter 16 does not show these differences also indicates that chapter 16 will have more Simple Present and less Simple Past than chapter 1 and chapter 17.
Another thing that is important to keep in mind is that there are fewer dialogue acts in chapter 1. Chapter 16 and chapter 17 contain more than twice as much data as chapter 1, this can also be seen in the table in section 6.2.

Now that the tense use in HP has been compared to the tense use in the SB corpus, I will compare the utterance distribution across the corpora.

7.4 Utterance distribution

In this section I will analyze the distribution of utterances across the corpora.

In order to do so I put the number of occurrences of the four utterance categories found in HP in a cross table. Table 6 shows the number of occurrences of the dialogue acts and table 7 shows the ratio in which the utterance occur in the corpora.

	HP	SB
Statement	712	100745
Question	153	9606
Agreement	89	13120
Backchannel	6	57572

Table 6: Frequency of utterances

	HP	SB
Statement	0.742	0.556
Question	0.159	0.053
Agreement	0.093	0.072
Backchannel	0.006	0.318

Table 7: Ratio of utterances

These tables show that the distribution of utterances differs across the corpora. All the ratios show a difference, but the biggest difference can be seen in Backchannel, which barely occurs in the HP corpus but makes up a big part of the SB corpus. To further analyze the distribution of utterances across the corpora I have conducted the χ-squared test. The result of this test showed a significant association in the distribution of utterances: $\chi - squared = 558.81, df = 3, p - value < 0.001$. This means that the distribution of utterances differs across the corpora. Figure 10 shows where the differences in the corpus can be found using Pearson residuals.

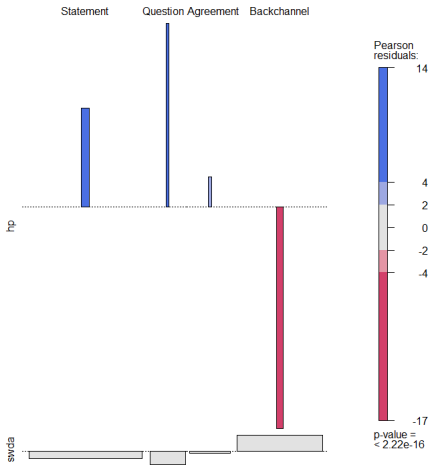

Figure 10: Utterance distribution in HP and SB, plot of Pearson residuals

As can be seen in figure 10 the utterances statement, question and agreement occur more than expected in the HP corpus, and there are fewer backchannels present in the HP corpus than in the SB corpus.

Because the frequency of backchannels differs so much the other utterances seem to be present a lot more. This is why in section 6.2 I argued I would only look at the dialogue acts present in the HP corpus and not take the blank spaces that would otherwise occur into account. This is why I will now compare the utterance types again but leave the backchannel type out.

To do so I have created a new table with the ratios in which these utterances occur, which can be seen in table 8.

	HP	SB
Statement	0.746	0.816
Question	0.160	0.078
Agreement	0.093	0.106

Table 8: Frequency of utterances statement, question and backchannel

As can be seen in table 8 the ratio of utterances in the HP corpus almost stayed the same but the ratio for the SB corpus changed. This makes sense because a large amount of data of the SB corpus is now left out, but only a small amount of data is left out of the HP corpus. To test whether or not the differences in the utterances types are different the χ-squared test was used. Again, a significant association in the distribution of utterances was found $\chi - squared = 89.432, df = 2, p - value < 0.001$. Meaning that the way the utterances are distributed is dependent on the corpus.

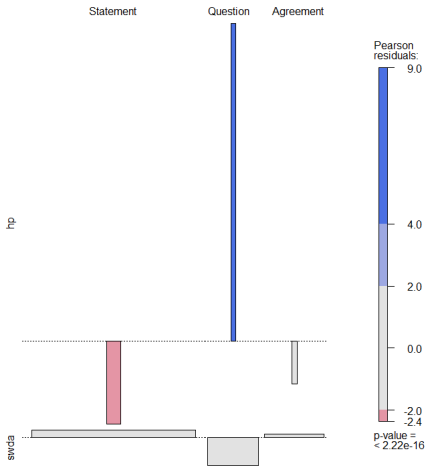

Figure 11: Utterance distribution in HP and SB, plot of Pearson residuals

Figure 11 shows that significantly less statements occurred in the HP corpus and significantly more questions occurred than in the SB corpus. There is no significant difference in the occurrence of the utterance agreement in this part of the data.

In this section the differences of the occurrences of utterances across the HP and SB corpus were discussed. In order to see if there are any other differences that can be exposed across the corpora I will zoom in on the two utterances that occur most frequently in the HP corpus, namely the statement and question utterance.

7.5 Zooming in: Statement utterance

In this section I will investigate the tense distribution of the statement utterance compared. This will be done using a significant test and if a plot with Pearson residuals.

Table 9 shows the frequency of statements occurring per tense for both corpora. Table 10 shows the ration in which this utterance occurs for each of the tenses.

	HP	SB
Perfect	47	6793
Past	197	14385
Present	305	47489

Table 9: Frequency of statement utterance per tense

	HP	SB
Perfect	0.086	0.099
Past	0.356	0.209
Present	0.556	0.692

Table 10: Ratio of statement per tense

A significant association of the tenses used in statements was found: $\chi - squared = 73.234, df = 2, p - value < 0.001$. Indicating that the tenses in the statement utterance differ across the corpora.

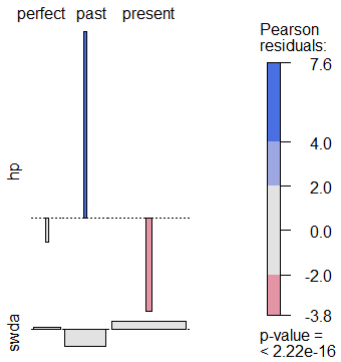

Figure 12: Occurrences of tenses for the statement dialogue act in HP and SB, plot of Pearson residuals

Figure 12 shows that the HP corpus for the statement dialogue acts contains significantly less Simple Present and significantly more Simple Past, no significant difference was found in the occurrence of the Present Perfect. The statement utterance is the most occurring utterance in the HP corpus. The differences between the tenses in the HP and SB corpus are similar to the differences in tenses in statements.

This section analyzed the tense use in the statement utterance. Tn the next section there will be a similar comparison, but this time for question utterance.

7.6 Zooming in: Question utterance

In this section the tense use of the question utterance in the HP corpus and the SB corpus will be compared. In order to compare these two the χ-squared test and a plot using Pearson residuals will be used.

Table 11 shows the frequency of questions occurring per tense for both corpora. Table 12 shows the ration in which this utterance occurs for each of the tenses.

	HP	SB
Perfect	13	619
Past	36	1038
Present	80	5322

Table 11: Frequency of question utterance per tense

	HP	SB
Perfect	0.101	0.089
Past	0.279	0.149
Present	0.620	0.763

Table 12: Ratio of question per tense

The ratio looks different for both corpora, in order to further investigate this a statistical test is needed. A significant association of the tenses used in questions was found: $\chi - squared = 17.829, df = 2, p - value < 0.001$

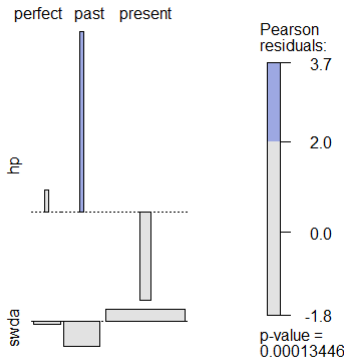

Figure 13: Occurrences of tenses for the question dialogue act in HP and SB, plot of Pearson residuals

Figure 14 shows that the HP corpus for the question dialogue acts contains significantly more Simple Past. No significant difference was found in the occurrence of the Present Perfect and the Simple Present.

Now that the data has been analyzed I will explain what findings mean and discuss where these results stand in relation to existing literature in chapter 8.

8 Discussion

In this section I will discuss what the results I presented in chapter 6 and 7 mean. Firstly I will discuss the occurrence of the utterances backchannel and hedge. Secondly I will discuss the distribution of tenses across both corpora. Thirdly I will discuss the distribution of utterance types. And lastly I will discuss the possible relationship between utterances and verbal tenses.

8.1 Backchanneling and hedging

An interesting difference between the two corpora is that some dialogue act types do not occur or barely occur in the HP corpus. Backchanneling for example is the smallest category in the HP corpus but the second biggest category in the SB corpus, as can be seen in the tables in section 6.2. This is probably due to the nature of the corpus. Backchannels are used in case of misunderstanding: think of requests for repetition or correcting something that was misspoken. Backchannels are also used as continuers. It makes sense that a polished text like HP will not have as many misunderstandings as a phone conversation and therefore will need less backchanneling than a phone conversation.

Hedges are used to to diminish the certainty of what a speaker says or what the speaker answers to a question. This dialogue act does not occur at all in the HP corpus. Again I think that this is due to the nature of the corpus, as the HP corpus comes from a book, so speakers do not have to be uncertain of what they are saying.

These findings are in line with Levshina (2017), who found that in subtitles less sentence reformulations are needed and the dialogue in subtitles is less vague than in a spontaneous conversation. These findings are partially in line with Buwalda (2020), who found that the Present Perfect in the HP corpus contains fewer Backchannels, Agreement and Hedges. When comparing the rest of the corpus Agreement does not occur less in the HP corpus than in the SB corpus, but the findings are in line with fewer Backchannels and Hedges occurring in the HP corpus.

Now we know how these three utterance are distributed across the corpora and how they fit in line with the research done by Levshina (2017) and Buwalda (2020). In the next section the tense distribution across the corpora will be discussed.

8.2 Distribution of tenses across both corpora

In general the HP corpus contained more simple past and less simple present than the Switchboard corpus. There was only one occurrence in which there was a significant difference in the Present Perfect. This difference was found in chapter 1, which is also the smallest chapter of the corpus. In all other comparisons there was no significant difference in the use of the Present Perfect.
These differences in tense distribution are shown in section 7.2 and supported by the sections 7.3, 7.5 and 7.6. The fact that the HP corpus contains more Simple Past than the SB corpus could be due to the selection of chapter in the HP

corpus. Now it will be important to look at what is events occur in the chapters as we look for explanations of the results we found in section 7. Chances are that the differences in tenses across the chapters is due to the nature of the chapter, I shortly want to summarize why certain tenses are found in a certain chapters. Chapter 1 is the start of the book, as an introduction a lot of 'looking back' is done, this makes sense as it draws us in the story and explains a bit of what is going on in the story. That is why it makes sense that this chapter is more past-oriented. Chapter 16 is full of action, and everything is described in that moment. That makes this chapter more present-oriented. Chapter 17 mainly looks back on the events occurring in chapter 16, and therefore is more past-oriented.
The Switchboard corpus on the other hand is happening in the moment, the speakers do not really have a lot to look back on. It therefore makes sense that no difference in tense use was found between the Switchboard corpus and chapter 16, because both of them are present-oriented and the conversations apply to the current moment.

It is interesting to note that no significant difference has been found in the use of the Present Perfect across the corpora. This is interesting because the HP corpus consists of British English and the SB corpus consists of American English, it is believed that in American English the Present Perfect will occur less than in British English. This was for example described by Hundt Smith (2009). The findings of my research are not inline with this theory, as this research shows no difference in the occurrence of the Present Perfect for American English and British English.

This finding is also interesting because this does not invalidate the use of the HP corpus for the Present Perfect by Van der Klis, Tellings & De Swart (2020) , as no difference has been found in the use of the Present Perfect.

8.3 Distribution of utterance types

Chapter 7 shows that the HP corpus contains more questions and less statements than the SB corpus. This is probably because questions keep the story more dynamic, this is inline with the research done by Levshina (2017). Another reason for this difference might be that when people have a conversation this conversation usually starts with a couple of questions and turns into people telling each other things in the form of statements. It could be the case that the dialogue in HP never reaches this level of communication and therefor has a higher occurrence of questions than the SB corpus.

8.4 Relation between tenses and utterances

Various reasons for the different tense distribution and utterance distribution of the corpora have just been discussed. In the last two sections the difference in the distribution of tenses and the difference in the distribution of utterances have been discussed as two separate things. There could be a relation between these two. In chapter 3 the tense distribution of the Switchboard corpus was shown for each utterance, which came from the research by Tellings, van der Klis, Le Bruyn and de Swart (2019). For ease I will show this distribution again

in figure 14.

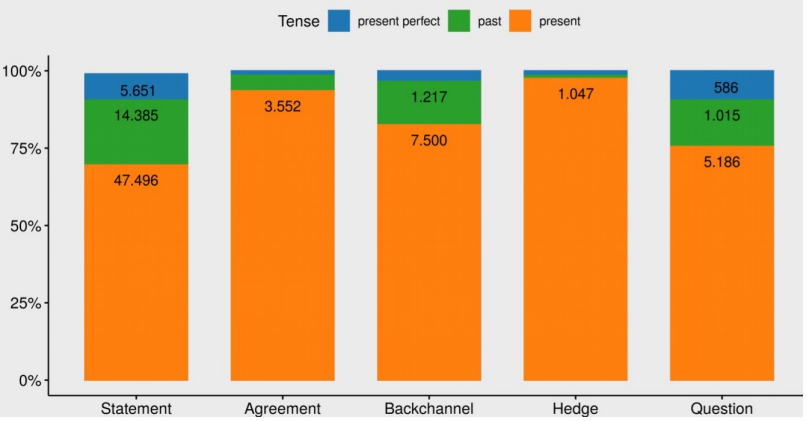

Figure 14: Distribution of tenses in each dialogue act in the SB corpus

The two utterances that contain the least Simple Present are statements and questions as can be seen in figure 14. These two utterances are the two most occurring utterances in the HP corpus. It could be the case that this is the reason the tense distribution of the corpora differs. However section 7.4 and 7.5 show that within these utterances the tense distribution also differs. Section 7.4 show that the statements contain more Simple Past and less Simple Present in HP than in SB. Section 7.5 showed that question utterance in HP contain more Simple Past than the question utterances in the SB corpus. These findings are in line with the difference in tense distribution across the full corpus. It seems like this finding rejects the idea of the differences in tenses usage being due to the difference in the distribution of tenses. However it is often recommended not to switch between the usage of tenses in narrative discourse unless an event calls for a change in tense use (Towson University, 2011). It could be the case that in dialogue the tense use is changed in a similar way. If fewer Present-oriented utterances occur than there are less reason to switch to the tense use, which could mean that more dialogue is in the Simple Past. This shows how the difference of tense distribution could be due to the difference in utterance distribution.

In this chapter the differences in the tense distribution and the utterance distribution have been discussed. Possible explanations for these differences, such as the selection of chapters, the difference between pre-written dialogue and actual dialogue and the possible relation between the differences has been discussed. Now that the data has been analyzed and discussed it is time to draw up a conclusion and answer my research question.

9 Conclusion

In this chapter I will summarize my research and answer my research question.

This research was done using two corpora, namely the Harry Potter corpus and the Switchboard corpus. I have compared the two corpora by looking at the verb tenses that have been used in the corpora and by looking at the dialogue acts that were used in both corpora.
The conclusions of this research is established by answering the research question, namely:

- How does the tense use in Harry Potter and the Philosopher's Stone compare to the tense use in the Switchboard corpus?

I expected to find that the tense distribution of the corpora would be the same and the dialogue acts in the corpora would have some differences. In this cha I will discuss the conclusion of this research.

Firstly I will discuss my sub hypothesis. Secondly I will discuss my main hypothesis, answer my research question and draw up further conclusions. And lastly I will discuss chances for future research.

9.1 Sub Hypothesis

I will now reflect on my sub-hypothesis. My sub-hypothesis was:
(S1) There will be a difference in the dialogue acts occurring in the corpora, I expect HP to have less backchanneling, hedges and agreements than the SB corpus.

As discussed in section 8.1 this hypothesis partially holds. Backchannels and hedges occurred in smaller amounts in the HP corpus than in the SB. However agreements occurred in a similar distribution as in the SB corpus. Thus this hypothesis partially holds. The difference in the utterance distribution are due to the nature of the corpus. This is in line with the research by Levshina (2017), who described that pre-written text is more polished than spontaneous occurring conversation. This finding is partially in line with Buwalda (2020), who found that less agreements, backchannels and hedges occurred in the Present Perfect in the HP corpus compared to the SB corpus.

The sub-hypothesis partially holds but still fits in with earlier research. In the next section I will discuss my main hypothesis and answer my research question.

9.2 Hypothesis & Research Question

My null hypothesis, which is also my main hypothesis was:
(H0) The distribution of tenses in Harry Potter and the Philosopher's Stone is the same as the tense distribution in the Switchboard corpus.

Based on the findings of my research I have to reject this hypothesis. As discussed in 8.2 and shown in chapter 7 the HP corpus contains more Simple Past and less Simple Present than the SB corpus. To strengthen this finding it might be useful to analyze the full Harry Potter corpus. Because as mentioned in section 8.2 the tense use differs across the chapters in HP, therefore this difference could be due to the selection of chapters from the Harry Potter corpus.

The findings of the this research have further implications. This research concludes that the SB corpus and the HP corpus differ in tense use and in utterance distribution. I will shortly discuss what these findings imply for AI and for research on the Present Perfect.

In chapter 2 the AI relevance of this research was discussed. If the HP corpus proved to be similar to naturalistic conversation than corpora such as HP could be used to speed up the process of algorithms learning conversational language. This way language models could be improved, which could contribute to passing the Turing test. With the findings of this research, showing a difference in the tense distribution and the utterance distribution across the corpora I must conclude that this corpus is not suitable to improve language models in the manner described in section 2, as it differs too much from spontaneous speech.

This research was done because of research on the Present Perfect that gave rise to the question if the dialogue in HP resembled spontaneous speech. As we just saw there are differences. But it is interesting that no difference has been found in the occurrence of the Present Perfect. This research therefore does show that the HP corpus differs from the spontaneous speech that occurs in the SB corpus, but does not give a reason why the HP corpus could not be used for research on the Present Perfect.

To ensure that my conclusion holds and to gain further insight in tenses and utterances I will propose options for further research in the next section.

9.3 Further research

In this research I found that the HP corpus contains more Simple Past and less Simple Present than the SB corpus. To strengthen this finding it might be useful to analyze the full Harry Potter corpus. Because as mentioned in section 8.2 the tense distribution across the chapters differs and the difference in tense distribution might therefore be due to the selection of chapters from the Harry Potter corpus.

Another topic that is interesting for future research is the tense distribution across utterances, and what the effect of missing utterances can have on the tense distribution. This is interesting because a the tense distribution of the HP differed from the HP corpus as well as the utterance distribution. As mentioned in section 8.4 the two most Past orientend were the most occuring in the HP corpus. This gives rise to the question whether or not the absence of some utterance can make a corpus more past oriented.

References

[1] Buwalda, I. (2020). Perfect use in dialogue contexts from Harry Potter and the Philosopher's Stone. *Unpublished*

[2] Copeland, B. J. (2000). The turing test. *Minds and Machines*, 10(4), 519-539.

[3] Hundt, M., & Smith, N. (2009). The present perfect in British and American English: Has there been any change, recently. *ICAME journal*, 33(1), 45-64.

[4] Jurafsky, D., Shriberg, L. & Biasca, D. (1997, August 1). *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13.* Retrieved from https://web.stanford.edu/jurafsky/ws97/manual.august1.html

[5] Klis, M. van der, Le Bruyn, B., & Swart, H. de (2020). *Temporal reference in discourse and dialogue.* Retrieved from https://time-in-translation.hum.uu.nl/publications/

[6] Levshina, N. (2015). Measuring associations between two categorical variables. *How to do linguistics with R: Data exploration and statistical analysis.* John Benjamins Publishing Company. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/uunl/detail.action?docID=4386605.

[7] Levshina, N. (2017). Online film subtitles as a corpus: an n-gram approach. *Corpora*, 12 (3), 311-338.

[8] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3), 276-282.

[9] Russell, S.J. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Pearson Education. Upper Saddle River, New Jersey

[10] Swart, H. de (2007). A cross-linguistic discourse analysis of the Perfect. *Journal of Pragmatics*, 39(12), 2273-2307.

[11] Tellings, J., van der Klis, M., Le Bruyn, B., & de Swart, H. (2019). Tense use in dialogue. Poster presented at SemDial 2019, London, England, 2019.

[12] Towson University. (2011). *Verb Tense Consistency.* Retrieved from https://webapps.towson.edu/ows/tenseconsistency.aspx

Appendix A

SWBD-DAMSL	SWBD	Example	Cnt	%
Statement-non-opinion	sd	*Me, I'm in the legal department.*	72,824	36%
Acknowledge (Backchannel)	b	*Uh-huh.*	37,096	19%
Statement-opinion	sv	*I think it's great*	25,197	13%
Agree/Accept	aa	*That's exactly it.*	10,820	5%
Abandoned or Turn-Exit	% -	*So, -*	10,569	5\%
Appreciation	ba	*I can imagine.*	4,633	2%
Yes-No-Question	qy	*Do you have to have any special training?*	4,624	2%
Non-verbal	x	*[Laughter], [Throat_clearing]*	3,548	2%
Yes answers	ny	*Yes.*	2,934	1%
Conventional-closing	fc	*Well, it's been nice talking to you.*	2,486	1%
Uninterpretable	%	*But, uh, yeah*	2,158	1\%
Wh-Question	qw	*Well, how old are you?*	1,911	1%
No answers	nn	*No.*	1,340	1%
Response Acknowledgement	bk	*Oh, okay.*	1,277	1%
Hedge	h	*I don't know if I'm making any sense or not.*	1,182	1%
Declarative Yes-No-Question	qy^d	*So you can afford to get a house?*	1,174	1%
Other	o,fo,bc,by,fw	*Well give me a break, you know.*	1,074	1%
Backchannel in question form	bh	*Is that right?*	1,019	1%
Quotation	^q	*You can't be pregnant and have cats*	934	.5%
Summarize/reformulate	bf	*Oh, you mean you switched schools for the kids.*	919	.5%
Affirmative non-yes answers	na,ny^e	*It is.*	836	.4%
Action-directive	ad	*Why don't you go first*	719	.4%
Collaborative Completion	^2	*Who aren't contributing.*	699	.4%
Repeat-phrase	b^m	*Oh, fajitas*	660	.3%
Open-Question	qo	*How about you?*	632	.3%
Rhetorical-Questions	qh	*Who would steal a newspaper?*	557	.2%
Hold before answer/agreement	^h	*I'm drawing a blank.*	540	.3%
Reject	ar	*Well, no*	338	.2%
Negative non-no answers	ng,nn^e	*Uh, not a whole lot.*	292	.1%
Signal-non-understanding	br	*Excuse me?*	288	.1%
Other answers	no	*I don't know*	279	.1%
Conventional-opening	fp	*How are you?*	220	.1%
Or-Clause	qrr	*or is it more of a company?*	207	.1%
Dispreferred answers	arp,nd	*Well, not so much that.*	205	.1%
3rd-party-talk	t3	*My goodness, Diane, get down from there.*	115	.1%
Offers, Options Commits	oo,cc,co	*I'll have to check that out*	109	.1%
Self-talk	t1	*What's the word I'm looking for*	102	.1%
Downplayer	bd	*That's all right.*	100	.1%
Maybe/Accept-part	aap/am	*Something like that*	98	<.1%
Tag-Question	^g	*Right?*	93	<.1%
Declarative Wh-Question	qw^d	*You are what kind of buff?*	80	<.1%
Apology	fa	*I'm sorry.*	76	<.1%
Thanking	ft	*Hey thanks a lot*	67	<.1%

Figure 15: The 42 final tags, including an example, the number of times they occurred and the percentage this annotation makes up of the total

Appendix B

Labels	future continuous	future in the past	future in the past continuous	future perfect in the past	imperative	infinitive	past continuous	past perfect	past perfect continuous	present continuous	present participle	present perfect	present perfect continuous	simple future	simple past	simple present	Total
a		1			74	1								1	1	11	89
b																6	6
f												1		1	1	4	7
q		5			1		2			8		13		8	36	80	153
s	1	16	1	5	18	1	11	9	2	30	1	46	7	67	196	301	712
Total	**1**	**22**	**1**	**5**	**93**	**2**	**13**	**9**	**2**	**38**	**1**	**60**	**7**	**77**	**234**	**402**	**967**

Figure 16: Table with the number of utterance per table for the HP corpus containing the original fx label.