UNIVERSITÄT
DES
SAARLANDES

Saarland University

Universität des Saarlandes

Fakultät 4 - Philosophische Fakultät

Fachrichtung 4.7 Allgemeine Linguistik

Computerlinguistik

Master's Thesis

submitted in fulfilment of the requirements for the degree of

"Master of Science" (M. Sc.)

On Universality of Prosody as a Turn-Taking Cue Across Languages

Submitted by:

Torsten Kai Jachmann

Born in: 10.04.1987 in Dudweiler

Matriculation Number: 2523139

Examiners and Thesis Advisors:

Dr. Maria Staudte

Dr. Chiara Gambi

SULZBACH, 02.09.2015

Declaration of Authorship

I, Torsten Kai Jachmann, declare that this thesis titled, 'On Universality of Prosody as a Turn-Taking Cue Across Languages' and the work presented in it are my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

SAARLAND UNIVERSITY

Abstract

Universität des Saarlandes

Fakultät 4 - Philosophische Fakultät

Fachrichtung 4.7 Allgemeine Linguistik

Computerlinguistik

"Master of Science" (M. Sc.)

On Universality of Prosody as a Turn-Taking Cue Across Languages

by Torsten Kai JACHMANN

In this thesis, I investigate the influence of prosodic information of three different languages (German, English and Japanese) on turn-end anticipation of German interlocutors. For that reason, I utilize a reaction time experiment in which participants listen to German, English, and Japanese sentences that were stripped of lexicosyntactic information and therefore only contain prosodic information. Contrary to previous studies, I am able to show that prosodic information is in fact helpful if not important for turn-taking management and that prosody is very language specific even across relatedness of languages.

Acknowledgements

At this point I want to express my gratitude towards all those who helped and supported me with and while writing this master's thesis.

First of all I want to express my gratitude towards Dr. Maria Staudte, who provided me with the possibility to write this thesis as well as guidance and support throughout the entire project and who initially raised and deepened my interest in and enjoyment of psycholinguistics. Further, I owe a dept of gratitude to Dr. Chiara Gambi, who's classes on turn-taking where the inspiration to this thesis and who agreed to be my thesis advisor. Both of you have been role models and ispirations over the years and I very much enjoyed working with you.

Aditionally, I want to thank Lisa Jachmann and Monika Nitze for participating in the recordings used to create the corpus for this thesis. Also, I would like to thank Hiroshi Aoyama, Mayumi Kawada and Claudia and Walter Verburg for translating the original turns into Japanese and English as well as Christine Ankener, Yumi Aoyama and Manisha Gandhi for voluntarily recording the turns in the three languages.

I also want to express my gratitude towards Hideki Yamaguchi, who deepened my love towards the Japanese language and who also aided with the Japanese translations of the turns. Thank you for supporting me over all those years and for being such a great teacher.

ご多忙にも関わらず、本研究に快く協力して頂いた青山大志、青山友美、河田真弓、古川実千子、山口秀樹の各氏に、この場を借りて感謝の意を表します。また、いつも応援し、支えてくださる五島馨にも、心から感謝いたします。本当にありがとうございました。

Last but not least, I want to thank all my friends all over the world and my family for supporting me morally and emotionally through even the hardest times. I wouldn't be who I am without you. Especially, I want to thank Sina Michels for helping out with trail runs and for listening to all my ideas and explanations. I know they sometimes came sudden, but your support was really helpful.

Speziell möchte ich meinen Eltern, Ilse und Rolf Jachmann danken, die mich bei jeder meiner Entscheidung unterstützen und mir immer eine Stütze sind und waren. Ich weiß, dass ich mich immer auf euch verlassen kann.

Contents

List of Figures

List of Tables

Abbreviations

AR	**A**rticulation **R**ate
SR	**S**peach **R**ate
SOV	**S**ubject **O**bject **V**erb
SVO	**S**ubject **V**erb **O**bject
TRP	**T**ransition **R**elevant **P**laces
TCU	**T**urn **C**onstruction **U**nit
PCA	**P**rincipal **C**omponents **A**nalysis

Chapter 1

Introduction

Turn-taking is an important feature of human conversations (Enfield and Levinson, 2006, Sacks et al., 1974). In such natural conversations, it is usually performed smoothly with only few gaps or overlaps between one turn of speaker A and the following turn of speaker B. Yet, up to now turn-taking is not very well understood in terms of the cues actually used to manage it (Magyari et al., 2014). Also, which combinations of available cues lead to the solid success that can be observed and which cues are even mandatory is still subject to research. Besides lexicosyntactic information (i.e., semantics and syntax) and visual information (i.e., body movements and gaze), another factor that is embedded in natural utterances is in suprasegmental information (i.e, tone and prosody). Studies conducted by De Ruiter et al. (2006) and Magyari and de Ruiter (2012) show that lexicosyntactic information is not only necessary to manage smooth turn-taking but might even be sufficient to do so.

However, the role of suprasegmental information cannot be cast aside. In natural human conversations, repairs, restarts, ungrammatical utterances and ellipses can often be found. Yet, such sentences seem not to affect turn-taking tremendously. Furthermore, lexicosyntactic information is not as informative in some languages as it appears to be in Dutch, the language that was examined by De Ruiter et al. (2006). Frequent cases in Japanese, for example, show that prosodic features are mandatory to distinguish turn endings from turn-internal utterances, as ambiguity is very common and even syntactic completion points are often disregarded for turn-taking if there are no so called terminal items that mark a sentence end without ambiguity (Tanaka, 2004). Syntactic completion points describe a point in speech, which completes a syntactic structure. An example for ambiguity in Japanese and its disambiguation by a terminal item can be seen in the following sentence: 凝視したあと帰ったよ。 *Gyoushi shita ato kaetta yo.* (*[Someone] went home after he stared at me.*) The sentence could already end after *Gyoushi shita* with the meaning *[Someone] stared at me.*. The *yo* as a terminal item unambiguously marks a sentence end. Without the *yo* it would be possible to continue the sentence for example with a noun like 人 (*hito*) meaning *person* which would render the sentence so far as a mere description of that person (*The person, who stared at me and went home afterwards.*) Also, even if lexicosyntactic information, especially syntactic completeness, was the most useful turn-yielding cue, it could still be only one important cue contributing to complex turn-taking cues as introduced by Duncan (1972). Duncan's findings for English, which were also extended and supported by Gravano and Hirschberg (2011), suggest an additive effect of different turn-taking cues, when they appear in combination. In other words, the more turn-taking cues are presented, the more likely it is that a smooth turn-taking takes place. Therefore, prosodic turn-taking cues might still contribute to smooth turn-taking.

Findings from Tanaka (2004) suggest that some prosodic cues that are found in Japanese partially correspond to those found by Duncan (1972) and Gravano and Hirschberg (2011). Further, a comparison of English and German intonation by Grabe (1998) reveals very similar contours in both languages in identical context. Considering these similarities across these three languages, the question arises whether prosodic contours utilized as turn-holding and respectively turn-yielding cues are universal across different languages or at least are more similar in more closely related languages. This would mean, that listeners should be able to anticipate turn-ends equally well no matter which of the three languages prosodies they are presented with. In this thesis, I present results from the empirical examination of this question.

1.1 Overview

Chapter 2 *Background* provides background information that is important for this thesis. This mainly includes differences and similarities between the three languages compared in this thesis. Following that, Chapter 3 *Related Works* focuses on previous studies on which this thesis is build or such studies that are closely related to the topic of this thesis. In Chapter 4 *Experimental Setup*, I provide information on the corpus that was used for the experiment and how it was created as well as information on the experiment itself. The analysis of the results of the experiment are then presented in Chapter 5 *Statistical Analysis*. This includes the approaches for outlier-removal and a description of the statistical models in addition to how they were created. Concluding this thesis, Chapter 6 *Discussion* contains the interpretation of the results of the analysis and a summary of the results of this thesis. Following the conclusion, the complete set of items used in the thesis is found in the Appendix followed by the Bibliography.

Chapter 2

Background

This chapter provides basic information and background knowledge needed to follow this thesis. This includes relevant differences and similarities of the three languages German, English and Japanese as well as a definition and description of Speech Rate (SR) and Articulation Rate (AR) as they are used throughout this thesis.

2.1 Languages

This section provides information on the three languages used in the experiment. I point out the similarities and differences between the three in terms of properties that are important for this thesis. This includes the language families in order to point out relatedness between the languages, as it is possible that languages that developed from the same family share more similarity also in terms of prosody than unrelated languages. Further, sentence structures of the three languages will be explained as the verb position might influence prosodic contours. Additionally, the accent type and underlying prosodic contours are explained as both have direct influence on the actual prosodic features of utterances in the languages. Also the presence or absence of noun and object in the three languages in natural verbal interaction is explained as the absence of either increases lexicosyntactic ambiguity which might lead to a more necessary and precise prosody in order to disambiguate the ambiguous lexicosyntactic information. Following a similar reasoning, additionally verb conjugation will be explained for the three languages. As adjectival modifications with their disambiguating and descriptive character might in cases follow special intonation patterns as they are usually utilized for precision and distinction, the position of the adjective might influence intonational patterns. Therefore, a description of the adjectives in each language was also included.

2.1.1 English

English is part of the Germanic language family. The grammar of modern English is resulting from a gradual change of a typical Indo-European dependent marking pattern with relatively free word order with little inflection and a rich inflectional morphology. The typical Germanic verb-second (V2) word order moved to an almost exclusive Subject Verb Object (SVO) word order (Van der Auwera and König, 1994). The intonational contour is partially driven by the stress accent nature of the language. A rule of English hereby states that a word can only have one stress and only vowels can be stressed (Chomsky and Halle, 1968). Further, intention can also alter the sentence intonation with a stress put on a word to make it more prominent. In English, subject and object are essential parts of a sentence and can rarely be omitted. Although subject omitting ellipses can be found in English (Nariyama, 2004), they are not considered standard sentence structure. The same holds for object omission. Yet, these cases seem to be restricted to answer ellipses, where redundant information that appeared in the question is elided, and stripping, which is limited to occurring in coordinate structures. Here, a word or phrase that was already present in the first sentence is omitted from the coordinated clause. The verb in English is conjugated. English hereby has forms for three singular persons and one plural person. In English, the adjective is used as a pre-modification, that means adjectives, or adjectival phrases appear in front of the noun they modify.

2.1.2 German

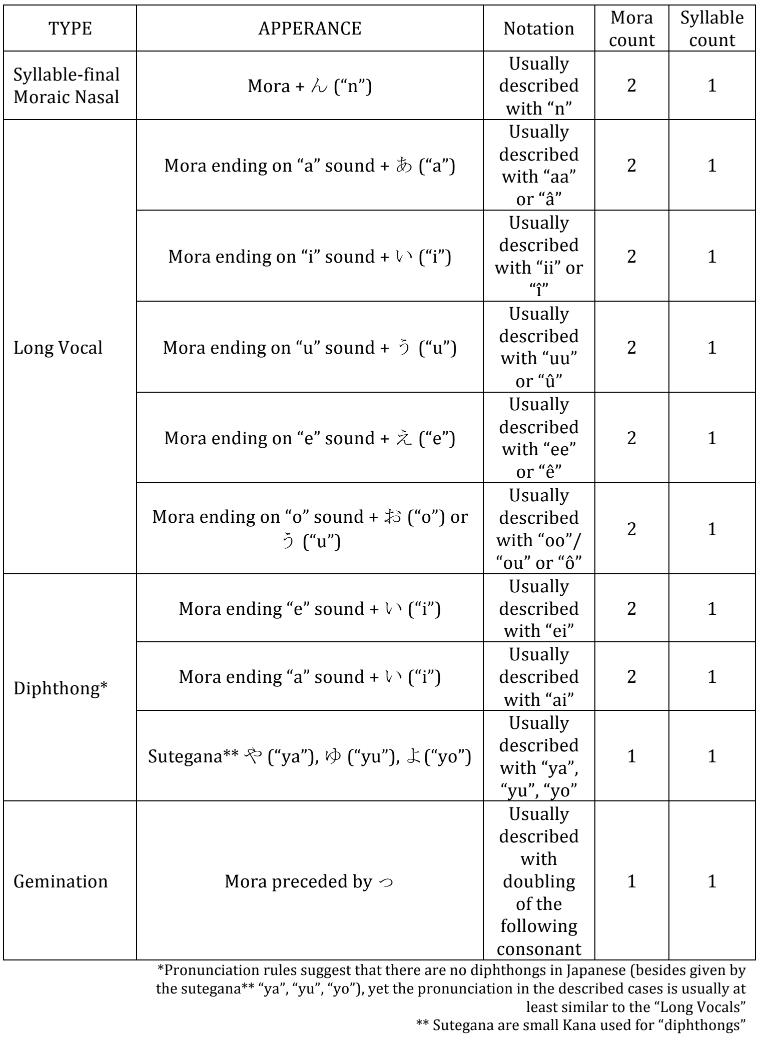
German as well is a Germanic language. The word order is more free than in English. For example in main clauses is a V2 verb order with SVO being the most common. Yet, an Object Verb Subject order is as well possible, as in the sentence *Den Mann sieht die Frau.*, which translates to *The woman sees the man.* This richness is provided by the grammatical cases that are visible in a word's inflection. German subordinate clauses on the other hand, are of the order Subject Object Verb (SOV). Also in other constructions, the meaning carrying verb is in the last position. For example in constructions including auxiliary verbs, the finite auxiliary verb is in the V2 position whereas the meaning carrying verb is put to the end of the sentence in its infinite form. The following sentence provides an example for this *Der Mann hat mir das Buch gegeben.*. The sentence translates into *The man has given me the book.* The meaning of the sentence, in this case which action has been performed on the book, is first revealed on the end of the sentence. Similar can be seen for modal verbs. Another phenomenon of German verb grammar led Bierwisch (1963) to believe that German originally was a SOV language. The phenomenon that he named as evidence are the separable verbs. Even in the main clause structure in German, a remnant of the separable verb is found in the very end of the sentence. For example the sentence *He bags the bananas.* can be translated as *Er tütet die Bananen ein.* The verb in its infinitive form is *eintüten* (*to bag*) and in fact appears in this form when auxiliary or modal verbs are included in the sentence, as in *Er hat die Bananen eingetütet.* (*He bagged the bananas*). Further, *tüten* alone is not a verb in German. Bierwisch (1963) therefore believed that the sentence final position is in fact the original position of the verb. In the scope of this thesis, German has a mixed word order of SVO and SOV sentence structure. The remaining properties that were additionally named for English are very similar to English properties. German as well has a stress accent that drives sentence intonation to some extent together with intention. Also subject and object omission follows similar rules as in English. In German verbs are also conjugated, yet, the grammar is richer. German has three persons each for singular and plural making it the richest of the three languages regarded in this thesis. Finally, adjectives in German are as well pre-modifying.

2.1.3 Japanese

Different from German and English, Japanese forms its own language family *Japonic.* It thereby is counted as an isolated language. The word order in Japanese is SOV, making it a strictly Verb-final language. Also Different from German and English, Japanese is usually not considered to be based on syllables, but rather on morae. The writing system in Japanese is closely related to this sound system, where one Kana[1] is usually describing one mora. One mora in Japanese consists of one or zero consonant sounds combined with

[1] A Kana describes one Japanese character of the basic writing systems Hiragana and Katakana.

exactly one vocal[2]. In order to calculate SR and AR, it was necessary to translate this mora-based system to a syllable-based system. Table 2.1 summarizes the changes made for this calculation.



TYPE	APPERANCE	Notation	Mora count	Syllable count
Syllable-final Moraic Nasal	Mora + ん ("n")	Usually described with "n"	2	1
Long Vocal	Mora ending on "a" sound + あ ("a")	Usually described with "aa" or "â"	2	1
	Mora ending on "i" sound + い ("i")	Usually described with "ii" or "î"	2	1
	Mora ending on "u" sound + う ("u")	Usually described with "uu" or "û"	2	1
	Mora ending on "e" sound + え ("e")	Usually described with "ee" or "ê"	2	1
	Mora ending on "o" sound + お ("o") or う ("u")	Usually described with "oo"/ "ou" or "ô"	2	1
Diphthong*	Mora ending "e" sound + い ("i")	Usually described with "ei"	2	1
	Mora ending "a" sound + い ("i")	Usually described with "ai"	2	1
	Sutegana** や ("ya"), ゆ ("yu"), よ ("yo")	Usually described with "ya", "yu", "yo"	1	1
Gemination	Mora preceded by っ	Usually described with doubling of the following consonant	1	1

*Pronunciation rules suggest that there are no diphthongs in Japanese (besides given by the sutegana** "ya", "yu", "yo"), yet the pronunciation in the described cases is usually at least similar to the "Long Vocals"

** Sutegana are small Kana used for "diphthongs"

Table 2.1: Transformations of morae to syllables done to calculate SR and AR.

Word accentuation in Japanese is strictly defined. Alternations of stress succession alone can change a word's meaning. For example, the word はし (*hashi*) can either mean "bridge" or "chopsticks". If the first mora は (*ha*) is low, followed by a accentuated し (*shi*), the word means bridge, whereas the reversed intonation pattern means chopsticks. This sometimes falsely leads to a classification of Japanese as a tonal language. Yet, the limited use of tone in Japanese rather classifies it as a pitch accent language[3] (Selkirk,

[2] An exception is formed by the syllable-final moraic nasal ん, which only consists of a consonant sound

[3] It is to be noted that the term *pitch accent language* might not be coherently defined (Hyman, 2006)

2009). Also the intonation pattern of a sentence in Japanese is following strict rules in polite standard Japanese. The first and second mora of a word must be of different level. If the first mora is high (H) the second mora has to be low (L) and vice versa. Thereby for the first two mora only two intonation patterns are possible: H-L and L-H. Further, the sentence intonation follows a downstep pattern. In other words, once the intonation stepped down, it can not rise again. Consequentially, this leads to different interpretation for high and low if appearing internally. High means that the level of the intonation is held, whereas low means a stepdown (Yamashita, 2004). Yet, these strict intonational patterns are not necessarily present in dialects. Many dialects in Japanese, such as the dialect spoken in the Kansai region, follow their own intonational rules (Yamashita, 2004). See Figure 2.1 for an example of an intonational contour of the same sentence in standard Japanese with strict intonational contour and the same sentence with Kansai dialect contour.

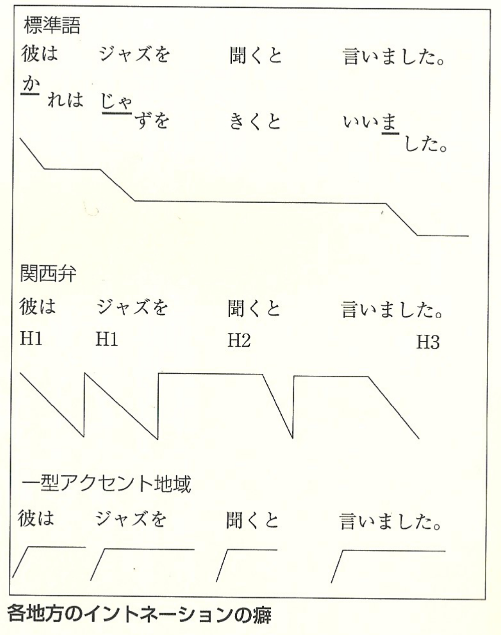

標準語
彼は ジャズを 聞くと 言いました。
かれは じゃずを きくと いいました。
関西弁
彼は ジャズを 聞くと 言いました。
H1 H1 H2 H3
一型アクセント地域
彼は ジャズを 聞くと 言いました。
各地方のイントネーションの癖

Figure 2.1: The upper contour shows the intonational structure of the sentence *Kare ha jazu wo kiku to iimashita.* (*He said he was listening to Jazz music.*) in standard Japanese with strict intonational contour. The contour below that shows the intonational contour of the same sentence in the Kansai dialect. The contour on the bottom shows the Type I accent region for each word. (in Yamashita (2004))

Additionally, the strictness is also reduced in standard non-formal Japanese (Venditti et al., 2008). In spoken Japanese, stress can be added to the intonation pattern to mark prominence and include intention similar to the way it is done in German and English. This difference to the standard formal Japanese intonation can also be seen in question intonation. In a standard formal question, the intonation would not rise towards the end of the question, which is usually a strong question marker in different languages. In an informal question however, a rising intonation towards the end is usually found. This is often even necessary to distinguish a question from a statement. The following example will explain this further. The sentence ケーキを食べた (*keeki wo tabeta*) could either mean *I ate cake.* or *Did you eat cake?* only dependent on the sentence final intonation. Here, a rising intonation towards to end indicates the question character of the sentence, whereas a falling intonation indicates a statement.

Another important point for Japanese is that the language use is highly context sensitive. If inferable from the context, the subject and even the object can be dropped from a sentence. This can be especially ambiguous, as Japanese verbs are not conjugated. For example, given the context sentence ケーキを食べますか (*keeki wo tabemasu ka*), which can be translated to *Are you eating your cake?*[4] the answer 食べます (*tabemasu*) mearly translating to *eat* is fully sufficient. In German and English the sentence would still need subject and object (*Ich esse ihn.* and *I eat it.* respectively). This increased number of ellipses compared to the other two languages is very characteristic for Japanese. For the creation of the corpus[5] used in the experiment to this thesis, I chose to keep initial subjects as the turns were presented to the participants without context[6]. If a subject was repeated in the same turn, it was omitted after its first use. This was done because an overuse of subjects would render the translations as unnatural and the meaning of a sentence could slightly shift, as a naming of the subject, even if *unnecessary*, would draw the focus on this subject. For example. ケーキを食べます (*keeki wo tabemasu*) could simply mean *I eat cake.* whereas 私はケーキを食べます (*watashi ha keeki wo tabemasu*) would put stress on 私 (*watashi*) *I*, leading to an interpretation like *If you ask about me, I eat cake.* As the two other languages, the use of adjectives in Japanese is pre-modifying.

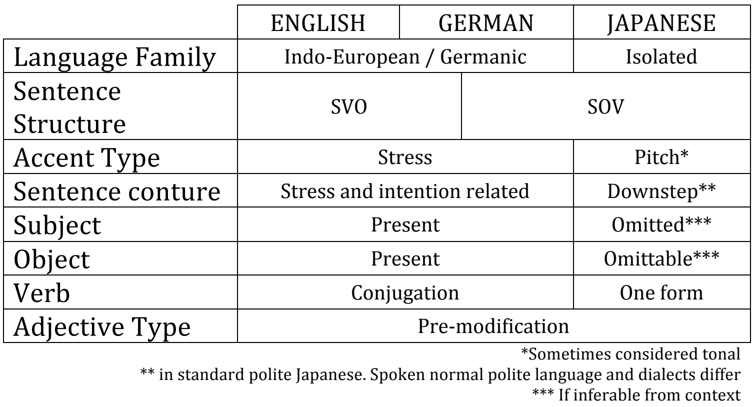
[4] Note that the question still is highly ambiguous. The lack of the subject in a natural way of asking the question still needs inference to who the person the question is referring to is. Therefore the question could also mean *Is he eating his cake?*, etc. Additionally the presence of the cake is also mandatory to receive this translation. An alternative translation could also be *Are you eating cake?*

[5] See Chapter 4, Section 4.1.1 "Corpus creation" for a complete description of the corpus.

[6] See Chapter 4, Section 4.2 "Experiment" for a full description of the experiment.

2.1.4 Comparison

This section summarizes the differences and similarities between the three languages used in this thesis. Table 2.2 shows a summary of the properties named in the preceding sections.



	ENGLISH	GERMAN	JAPANESE
Language Family	Indo-European / Germanic		Isolated
Sentence Structure	SVO	SOV	
Accent Type	Stress		Pitch*
Sentence conture	Stress and intention related		Downstep**
Subject	Present		Omitted***
Object	Present		Omittable***
Verb	Conjugation		One form
Adjective Type	Pre-modification		

*Sometimes considered tonal
** in standard polite Japanese. Spoken normal polite language and dialects differ
*** If inferable from context

Table 2.2: Properties of the three regarded languages.

For the language family, two groups form. German and English are both Germanic. The other group consists of the isolated language Japanese only. In terms of sentence structure, German with its mixed word order of SVO and SOV sentence structure serves as a bridge between the strictly SVO structured English and the strictly SOV structured Japanese. As all three languages use pre-modifying adjectives, intonation patterns due to adjectives are unlikely to differ across the three languages. For the remaining factors, again German and English are forming a group against Japanese. Yet, the intonational sentence contour shares some similarities in the non-formal use of Japanese. Given these properties, it is reasonable to assume that the two mainly overlapping languages German and English also express more similar overall prosodic contours than any of the two compared to Japanese. Yet, the ability of German to form both sentence structure types SVO and SOV might lead to a stronger closeness of prosodic contours to Japanese than between English and Japanese. In other words, it might be possible that the prosody of German SOV type sentences shows a closer similarity to Japanese prosody whereas the prosody of German SVO type sentence might show a closer similarity to English prosody.

2.2 Speech Rate and Articulation Rate

Speech Rate (SR) and Articulation Rate (AR) are both measurements of the fastness of an articulation as perceived by listeners. They are closely related, yet contain slight differences.

2.2.1 Speech Rate

SR is a measure of the number of speech units of a given type produced within a given amount of time. A common measurement is syllables per second, which is also used within this thesis. SR differs to some extent between speakers and also within a speaker depending on the emotional state (Arnfield et al., 1995). It also differs across languages (Roach, 1998). SR contains not only actual speech, but also pauses and hesitations within the time from the start of the speech till the actual end.

2.2.2 Articulation Rate

AR mostly describes the same information as also provided by SR but with a crucial difference. Different from SR, pauses are not included in the computation. Therefore, before the computation of AR, pauses within the speech have to be detected and their length determined. The complete length of all pauses within the speech is then subtracted from the total speech time before it is divided by the number of syllables within the speech to determine the actual AR.

Chapter 3

Related Work

Sacks (1995) names two important principles of conversation. One being that only one speaker talks at a time and the second one being that the space between the end of one speakers turn and the beginning of the following speaker's turn should be very short. In summary, these two principles mean that there should neither be gaps nor overlaps in a well-formed conversation. If in fact such overlaps or gaps do occur in a conversation, they are often perceived as an intentional speech act. For example, a longer overlap can mean that both speakers want to gain the right to speak and to some extent try to force the other speaker into the listener role. Conversely, a longer gap can often be interpreted as either denial of the speaker role by any participant or even as a meaningful signal to show that, for example, one does not know the answer to a preceding question. For the very reason that such a divergence from the optimal no gap/no overlap situation can lead to such interpretations, it is argued that smooth transitions are the norm in conversation. The fact that this is usually the case leads to the question how the next speaker is able to find the exact point in time to start his/her utterance. It is not possible for this speaker to simply wait for the preceding turn to end. The production of words and utterances takes at least 600 ms to reach a point where an utterance can be started, as shown in various experimental studies (Indefrey and Levelt, 2004, Jescheniak et al., 2003, Schnur et al., 2006). Therefore, in order to achieve a smooth turn-taking, there has to be a way to predict the upcoming end of a turn at least 600 ms before it actually comes up. Such predictions seem to be done using certain cues offered by the current speaker. Various possible cues have been found and researched, yet it is still an open question which of the cues or which combinations of them lead to the highest success rate (Magyari et al., 2014). Such cues could for example be prosody, eye-gaze in face-to-face conversations or lexicosyntactic information that enables listeners to find upcoming completion points, which might be used as Transition Relevant Places (TRPs) (Ford and Thompson, 1996). Such TRPs describe the end of a Turn Construction Unit (TCU) and thereby mark the point at which a speaker change can appear. If no speaker change appears at a TRP the same speaker can continue with another TCU again ending with a TRP.

3.1 How Smooth is Smooth Turn-taking?

Weilhammer and Rabold (2003) show that the understanding of smooth turn-taking has to be taken relatively. In their analysis of turn transitions of spontaneous face-to-face conversations in American English, German and Japanese, they found that slight pauses and overlaps are actually normally distributed. Even though such pauses and overlaps exist, turn transitions that include such can still be considered smooth as long as those gaps and overlaps are not perceptible. Even though it is not clear exactly how long a gap or overlap can be while still being imperceptible, Walker and Trimboli (1984) found that untrained transcribers show a threshold of about 200 ms for detecting between-speaker gaps. In Weilhammer and Rabold (2003) the average length of gaps was 380 ms and Wilson and Wilson (2005) found only 30% of the gaps between speakers' turns being shorter than 200 ms and 70% being shorter than 500 ms. Both findings still show a tendency for gaps to be shorter than the 600 ms needed for the articular system to be ready for an utterance, which shows that prediction of upcoming turn-ends is needed nonetheless.

3.2 Complex Turn-taking Cues

With the gaps between two consecutive speakers turns being longer than 200 ms Duncan's theory of turn-taking being based on behavioral cues close to the end of previous turns might still be viable. Duncan (1972) describes six behavioral cues. (1) Any phrase-final intonation other than a sustained, intermediate pitch level; (2) A drawl on the final syllable of a terminal clause; (3) The termination of any hand gesticulation; (4) A stereotyped expression like *you know*; (5) A drop in pitch and/or loudness in conjunction with such a stereotyped expression; (6) The completion of a grammatical clause.

An analysis of his data revealed a linear correlation between the number of these turn-taking cues and the number of turn-taking attempts. Utilizing more of the available signals seems to have an additive effect. However, it has to be taken into account that turn-taking is still successfully managed, even if interlocutors cannot see each other (e.g. in phone calls). Therefore, the list of turn-taking cues as described by Duncan (1972) does not necessarily have to be complete or fully mandatory in all possible scenarios. Additions for behavioral cues where for example made by Clark and Tree (2002). They proposed that repeats, repairs and prolonged syllables are cues that are intentionally utilized by speakers to express internal processes, such as preparing or rethinking an answer, and might serve as turn-holding cues. Wightman et al. (1992) further back this view with their finding that prosodic variation, such as intonation and segmental lengthening, seems to mark various segment boundaries in speech. Additionally, Shriberg (1994) found that hesitations are likely to appear preceding longer utterances and Watanabe et al. (2008) found that listeners see speaker hesitations further as predictors of upcoming words of high complexity. Even though Duncan's work has been criticized for the lack of formality and objectivity (Beattie, 1981, Cutler and Pearson, 1986), it was the first to introduce the existence of complex turn-yielding cues that consist of more than one turn-yielding signal. Thereby, his work formed a basis for various following research projects.

3.3 Intonation as Part of Complex Turn-taking Cues

Ford and Thompson (1996) examined the relation of grammatical completion defined by syntactic completion points and intonation in English. In their study intonation is encoded binary as either being final, independent of rising of falling intonation and non-final pitch contours. They identify a final intonation contour together with a syntactic completion as an important turn-yielding cue. However, they found a tendency of syntactic completion to be more prominent than intonation. Whereas 98.8% of the intonationally complete utterances are also syntactically complete, only 53.6% of the syntactically complete utterances are also intonationally complete in the boundaries of their interpretation of intonational completeness. This shows that it is highly unlikely to find cases in which intonational completeness can be found without syntactic completeness. Wennerstrom and Siegel (2003) extend Ford and Thompson's work with more precise definitions of final intonations. Using a predecessor of the ToBI transcription framework (Beckman and

Hirschberg, 1994, Pitrelli et al., 1994), they identify six final intonation patterns: high rise (H–H% in the ToBI system), low (L–L%), plateau (H–L%), low rise (L–H%), partial fall (also L–L%) and no boundary. Especially high rise and low are expressing strong turn-yielding cues with 67% and 40% respectively of their occurrences appearing together with speaker changes in Wennerstrom and Siegel (2003) interpretation. The remaining four patterns were more likely to appear in the context of turn holds. An analysis of the interaction between intonation and syntactic completion showed similar results to those found by Ford and Thompson (1996). Findings from De Ruiter et al. (2006) make an even stronger claim, with their suggestion that for a smooth turn-taking lexicosyntactic information alone might be sufficient but is at least mandatory. In their study on Dutch, they created five conditions for 108 turns by processing them with Praat. The first condition was a NATURAL version, which was the original recorded turn. A NO-PITCH version was created by flattening the pitch (F0) contour using PSOLA resynthesis with the mean pitch value of the original fragment, creating a pitch contour that was completely horizontal. A NO-WORDS version was created by applying a low-pass filter to the original fragment at 500 Hz. In this version lexicosyntactic information was made uninformative, but the original pitch contour was preserved and thereby still accessible. A NO-PITCH-NO-WORDS version was created with a combination of the two before named approaches. Thereby, neither lexicosyntactic information, nor pitch contour was intelligible, yet rhythm was contained. A NO-PITCH-NO-WORDS-NO-RHYTHM version, also referred to as NOISE version, merely contained constant noise with the same length and frequency spectrum as the original turn. This version served as a baseline, to test for any influence of short pauses and the amplitude-envelope information that was still contained in the NO-PITCH-NO-WORDS version. They found that listeners' accuracy in predicting an upcoming turn-end was not significantly impaired when the intonational contour was removed. However, when lexicosyntactic information was made uninformative, the existence of pitch was leading to a significantly higher accuracy in predicting upcoming turn ends compared to the NO-PITCH-NO-WORDS and NOISE conditions. Additionally, the absence of rhythm in the NOISE condition was leading to a significantly lower accuracy than any other condition (see Figure 3.1 for comparison). A study by Gambi et al. (2015) was able to reproduce those findings for German.

These findings, even though outlining the prominence of lexicosyntactic cues, still suggest that prosody, in terms of pitch as well as rhythm, is still used at least in situations where

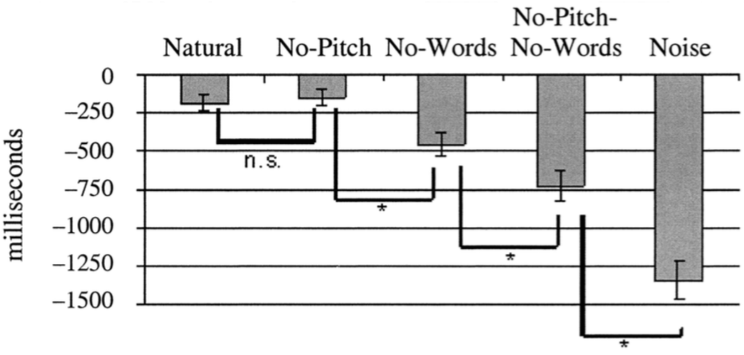


FIGURE 3.1: Average BIAS of responses per condition. * indicates statistical significance at the 0.05 level. (in De Ruiter et al. (2006))

lexicosyntactic information is uninformative.

3.3.1 Importance of Prosody for Turn-taking

Additionally, merely relying on lexicosyntactic information cannot account for uninterrupted turns in cases where one TCU is followed by another TCU of the same speaker or is extended beyond syntactic completion points. Couper-Kuhlen and Ono (2007) describe five types of TCU continuation found in English, German and Japanese. Especially, the Non-add-on continuation, which is mainly found in German and Japanese, seems to highly rely on intonation. The Non-add-on continuation describes a continuation after a TCU that is highly marked for syntactic closure, but shows no prosodic break in the transition from the TCU to the continuation. Syntactic closure is reached when an utterance could be interpreted as syntactically complete independent from intonation or pauses. An example for such a Non-add-on continuation is found in Auer (1996). „könn ma nomal zusamm sprechn morgn“ (we can talk about that again tomorrow). After „sprechn“ (talk) syntactic completion is actually reached, but the TCU is extended with „morgn“ (tomorrow), without any prosodic break. An example for Japanese is found in Couper-Kuhlen and Ono (2007). „kaku koto ga tanoshi n da yo yappari“ (Writing is fun, after all). Here as well, syntactic completion is already reached at „yo“ (an sentence end marker) but the TCU is extended by „yappari“ (after all) without any prosodic break. Additionally, even if rare, such continuations can also be found in English. „Cyd rang this evening Cyd Arnold“. Syntactic completion is reached after „evening“, but the TCU is extended by „Cyd Arnold“, without any prosodic break. This finding suggests that for each of the three languages I aim to compare syntactic completion alone cannot account for all TRPs of the languages to at least some extent and, further, that prosody seems to play a role in such cases. Gravano and Hirschberg (2011) further investigated different turn-taking cues in natural conversations. These cues contained a larger variety of acoustic, prosodic and lexicosyntactic cues compared to previous studies and also was conducted on a larger corpus to attain statistically robust results. They found seven cues that are more frequently appearing before smooth transitions compared to turn holds. These cues are (1) a falling or high-rising intonation at the end of a TCU, (2) a reduced lengthening of words at the end of TCUs, (3) a lower intensity level, (4) a lower pitch level, (5) a point of syntactical completion, (6) a higher value of the voice quality features jitter[1], shimmer[2] and NHR[3] and (7) an overall longer TCU duration. Out of these seven cues, five can be considered part of prosody alternation. All seven cues are predictors for upcoming turn transitions with a linear relationship. In other word, the more of these cues are present, the more likely it is for a smooth turn transition to occur. Even though these cues somewhat differ from those introduced by Duncan (1972), the findings support his account of complex turn-taking cues. In summary, all findings can to some extent be linked together. Duncan's account of complex behavioral turn-taking cues is not necessarily refuted by De Ruiter et al. (2006) findings. Even though lexicosyntactic cues have been shown to be very salient and helpful in turn-taking, they might just be the most helpful, but not sole cue used. This would explain turn-taking behavior in cases of continuations beyond syntactic completion as described by Couper-Kuhlen and Ono (2007). Further, the nature of the tasks in De Ruiter et al. (2006) and Gambi et al. (2015) has to be taken into account. Participants were not provided with a complete context or the intention of the utterances and therefore could not know if a turn was contextually complete or not. The lack of context does not allow for predictions about the content of the turn participants were listening to. The lack of information about the intention of the utterance might push the usefulness of prosodic cues further into the background as certain prosodic contours and speech behavior that are connected to intention are as well no longer predictable. Thereby, the mostly reliable lexicosyntactic information might be at an advantage.

[1] Fluctuations in pitch

[2] Fluctuations in amplitude

[3] Noise-to-Harmonics Ratio

Additionally, participants were only predicting turn ends and did not have to actually respond to the sentences verbally. This might be an explanation for the overall negative divergence relative to the actual turn end found by De Ruiter et al. (2006) and Gambi et al. (2015). The time needed for motor response to an auditory stimulus that was asked from the participants is very different from the time needed to prepare and produce an articulatory response as shown by Indefrey and Levelt (2004), Jescheniak et al. (2003) and Schnur et al. (2006). Motor responses are with an average of 360 ms (Ng and Chan, 2012) shorter than the 600 ms needed to react vocally. Therefore, if in fact early cues in the speakers' utterances trigger the preparation for turn-taking, the lower reaction time needed for motor responses might explain why the bias is overall more negative than 0 ms in the findings of De Ruiter et al. (2006) and Gambi et al. (2015) instead of slightly longer as predicted by Walker and Trimboli (1984), Weilhammer and Rabold (2003) and Wilson and Wilson (2005). Thereby, even a negative bias in experiments asking for motor responses is not necessarily incompatible with the natural distribution of inter-turn intervals.

3.3.2 Japanese Example for the Importance of Prosody

The importance of turn-taking cues besides those provided by lexicosyntactic information can be shown on the example of Japanese. In Japanese, lexicosyntactic information alone does not provide as much information as it does in English. The sentence structure in Japanese is Subject, Object, Verb (SOV). With the verb coming in the last position, an important part of information cannot be integrated until the end of a sentence. Therefore, predictions of possible upcoming sentence structures that usually can be inferred from verbs in languages in which the verb comes early in a sentence cannot be made. Also, considering that the lexicosyntactic cue comes very late, it might not be possible to use it efficiently to manage turn-taking (although there is some indication that predicitons in Japanese can be based on pre-verbial complements and case marking (Kamide et al. (2003), Exp. 3)). Further, interrogative and declarative sentence are usually of the very same form and only differ in the last word of a sentence. For example, the sentence "(watashi ha) keeki ga suki desu" (I love/like cake) is the declarative form. Just by adding the question particle "ka" to the end of the sentence it turns into an interrogative sentence. The same can be seen in a more informal way of saying the same sentence: "keeki ga suki da" vs. "keeki ga suki ka". In the declarative form the copula da is added without a question particle whereas in the interrogative form only the question particle is added. Such particles and copula amongst others (e.g. final suffixes, nominalizers, final particles) are therefore often referred to as utterance-final objects (Tanaka, 2004) as part of a turn-final grammatical design (Tanaka, 2000). Another factor is also displayed in the above-mentioned example. The subject of a sentence (in the example watashi ha – I am) is optional in Japanese sentences. It is only explicitly added to the sentence if it is not understandable from the context alone (Nariyama, 2003). In other words, as verb suffixes in Japanese do not reflect person or number, lexicosyntactic information alone could not provide sufficient information to understand the heard sentence. The final factor I want to mention that shows that lexicosyntactic information is not sufficient to detect an upcoming end of a turn is the form of relative clauses. The sentence "I thought you were already going home" in English shows in an early state of the sentence that a complement clause is coming up. In Japanese the same sentence would give that information only after the relative clause: "Mou kaeru to omoimashita". The particle "to" before the final verb "omoimashita" (thought) is the indicator for the preceding complement clause. If everything before the "to" was to be uttered alone, the semantic content would dramatically change into "I'm already going home". In the English sentence syntactic completion is only reached at the end of the whole utterance whereas the Japanese sentence consists of two syntactic completion points, one before "to" and one at the end of the utterance. Therefore, if only lexicosyntactic information was used for turn-taking management, the first completion point in the complement clause would be deemed misleading. Adding these findings together it is reasonable to say that the lexicosyntactic cue in English contains much more information on different levels than it does in Japanese. As the lexicosyntactic cue alone seems not to suffice to identify an upcoming end of a turn, Japanese speakers seem to have at least partially created a different approach, as described by Tanaka (2004). Syntactic completion points are not taken into account for turn-taking if they are not followed by at least one of the above-mentioned utterance-final objects. Yet, considering their only very late appearance in the turn in combination with their usually short form, they are unlikely to be used as reliable cues for smooth turn-taking management alone. Another point that has to be taken into consideration is that, although one might argue that especially the verb final structure of Japanese could be used as a strong cue to predict a turn ending, different aspects lead to a decrease in the reliability of verbs as a turn-taking cue. Besides the before mentioned aspect of word order which reveals complement clauses only after a syntactic completion point, the verb forms revealing sentence coordination as well only can be integrated very late. For example, the sentence “After I ate this cake, I’ll go home” reveals very early the upcoming sentence structure. The same sentence in Japanese would be “Kono keeki wo tabete kaerimasu”. Only when reaching the verb ending “te” of the verb “tabete” (eat), the sentence structure, especially the existence of a following subsentence, can be inferred. If the verb ending instead were, for example, “ta” (past tense marker), an utterance-final object, the sentence would simply mean “I ate cake”. In other words, whereas languages such as English usually already very early allow for expectations of upcoming sentence structures, Japanese reveals such structures only very late on the verb. Therefore, hearing the onset of a verb does not allow predicting an upcoming turn end for certain. Only the utterance-final objects could provide a reliable cue. Despite the lexicosyntactic cues being at least problematic for turn-taking management, Stivers et al. (2009) showed that turn-taking in Japanese is as precise as it is in English with a mean offset time for a consecutive turn of 7 ms and 236 ms respectively as displayed in Figures 3.2 and 3.3. Therefore, other cues have to be utilized in Japanese to achieve that smooth performance in turn-taking additional to the very late lexicosyntactic cues.

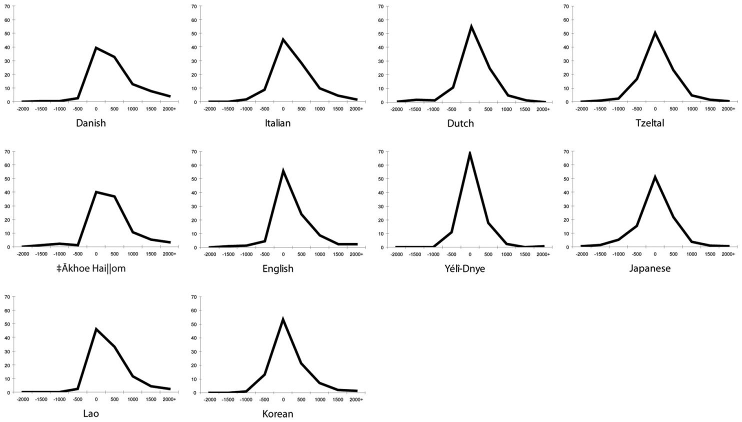


Figure 3.2: The distribution of turn transitions for each language examined by Stivers et al. (2009). All distributions are unimodal with the highest number of transitions occurring between 0 ms and 200 ms. The percentage of turn transitions is shown on the y axis, and milliseconds of turn offset are shown on the x axis. (in Stivers et al. (2009))

Tanaka (2004) found five types of prosodic alternation that appear as complex turn-yielding cues. (1) The lengthening of the final mora, (2) the lengthening of the penultimate mora, (3) a glottal stop at the end of an utterance, which is usually not present

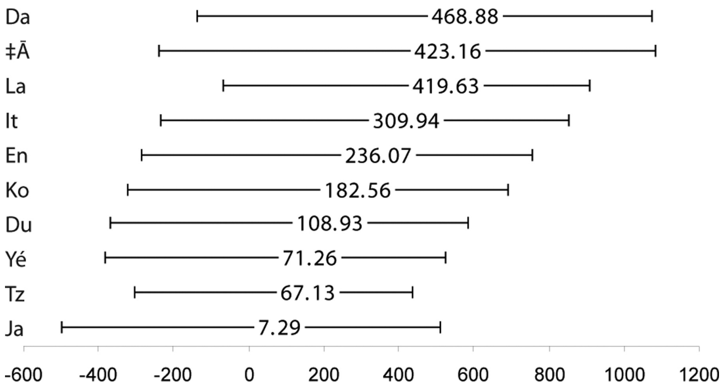
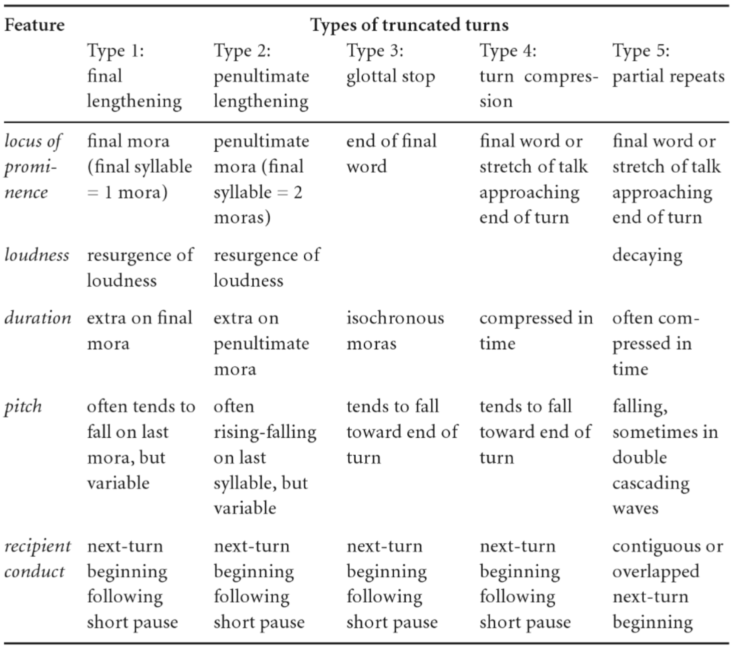


FIGURE 3.3: The mean time (in ms) of turn transitions for each language (±1 SD) for each language examined by Stivers et al. (2009). JA displays the results for Japanese and EN the results for English. (in Stivers et al. (2009))

in Japanese (4) turn compression and (5) partial repeat. The types 1, 3, 4 and 5 additionally show a falling pitch contour, whereas type 2 shows a rising-falling pitch contour. Furthermore, types 1 and 2 are accompanied by a resurging loudness and type 5 by a decaying loudness. Also the speed of the syllables production changes across the five types compared to utterances that do not contain any of the cues. A summary of the five complex turn-yielding cues can be found in Table 3.1.



Feature	**Types of truncated turns**				
	Type 1: final lengthening	Type 2: penultimate lengthening	Type 3: glottal stop	Type 4: turn compression	Type 5: partial repeats
locus of prominence	final mora (final syllable = 1 mora)	penultimate mora (final syllable = 2 moras)	end of final word	final word or stretch of talk approaching end of turn	final word or stretch of talk approaching end of turn
loudness	resurgence of loudness	resurgence of loudness			decaying
duration	extra on final mora	extra on penultimate mora	isochronous moras	compressed in time	often compressed in time
pitch	often tends to fall on last mora, but variable	often rising-falling on last syllable, but variable	tends to fall toward end of turn	tends to fall toward end of turn	falling, sometimes in double cascading waves
recipient conduct	next-turn beginning following short pause	next-turn beginning following short pause	next-turn beginning following short pause	next-turn beginning following short pause	contiguous or overlapped next-turn beginning

TABLE 3.1: Types of truncated turns: clusters of prosodic features and their receipt (in Tanaka (2004)).

3.3.3 Hints towards the Universality of Prosodic Cues

The findings of Tanaka (2004) for Japanese are partially corresponding to the cues named by Duncan (1972) and Gravano and Hirschberg (2011) for English, as for example the lengthening of the final syllable, or the drop in loudness. Further, Selting (1995) described similar prosodic patterns for German. A comparison of English and German intonation by Grabe (1998) revealed that in identical context speakers of the two languages produce very similar intonation patterns. Yet, there still are differences, as for example truncated pitch contours in German after falling movements when only little sonorant material is available, whereas in English compressed contours are found in the same context. Considering the similarities in pitch contours and other prosodic features, it seems reasonable to assume the possibility of universal prosodic cues that might contribute to complex turn-taking cues in these languages. Thereby, relatedness of languages might not be as important for prosodic cues. Yet, the differences that are found still call for an empirical investigation and might still reveal an effect of relatedness in a more fine grained level. If indeed the similarities of the prosodic contours of the three languages are strong enough to lead to universally usable prosodic turn-taking cues, listeners that are presented with any of the three languages prosodic contours should not treat them differently and achieve equal precision in turn-end anticipation for all three of them.

Chapter 4

Experimental Setup

In this chapter, I describe the creation of the stimuli and the setup of the experiment conducted for this thesis. It was created similar to DeRuiter et al. (2006) and Gambi et al. (2015) as a reaction time experiment in which participants had to anticipate the end of turns played to them. The main focus hereby lies on the three conditions for which lexicosyntactic information was made unintelligible and therefore only contained prosodic information of the three languages. If prosodic turn-taking cues overall are helpful and important for turn-taking management, the precision with which participants can predict upcoming turn-ends should be close to zero. Further, if prosodic cues are universal for all tested languages, the precision in which the participants can predict that the turn-ends[1] should not differ significantly from one another. If however a significant difference was to be found, a grouping of the languages would reveal the influence of relatedness of the languages on their prosodic contours. A grouping of German and English with a signifcant difference to Japanese would indicate the importance of language family, whereas a grouping of German and Japanese with a significant difference to English would hint towards the importance of sentence structure. If no grouping of German with either of the two languages was to be found, this would indicate that prosody is language specific and independent from any sort of relatedness. To answer these questions, the reaction time experiment contains four conditions. A full version of German turns are forming a baseline condition and provide the possibility to compare to previous works. The remaining three conditions are only containing prosodic information of turns in German, English and Japanese.

[1] Precision of turn-end anticipation is measured in bias as described in Section 4.2 *Experiment.*

4.1 Preparation

This sections provides information on the creation of the corpus used to extract the stimuli for the experiment and the pre-processing of these stimuli to create the experimental items as well as information on the creation of the lists and the experiment itself.

4.1.1 Corpus Creation

In order to conduct the experiment, I created a new corpus of 96 experimental turns extracted from recordings of a natural conversation. The recordings were obtained from two female native German speakers, who knew each other. They were paired up this way in order to achieve a fluent conversation. On arrival, they were handed a sheet with different topics to include in their conversation, beginning with short questions concerning their daily life, such as plans for the future, their first meeting, activities together, etc. After the short question section they were to perform a map task. In this task, both participants were handed a map containing various landmarks. One of the participants had a route drawn into the map that she was to explain to the other participant in order to retrieve a treasure hidden on the map only visible to the instructor. Additionally, six differences were included in the map, which were to be spotted by both participants. The maps handed to the participants can be found in the Appendix. The final topic on the instruction consisted of two topics for an open discussion. The participants were allowed to digress from the given topics as much as they wanted. The only constraint given to the participants was not to use dialect in their conversation. Both participants were payed with 15 Euros.

The recordings were cut and analyzed with Praat (Boersma and Weenink, 2013). For the experiment, I chose only turns that did not contain overlaps or pauses longer than 300 ms but for which turn-taking was successful in the conversation. Out of the 96 experimental turns extracted, about one third was labeled as Questions, one third as sentences of the structure SVO and one third as sentences of the structure SOV.

All turns that I used in the experiment were translated from German to English by two translators and from German to Japanese by two native speakers of Japanese with profound knowledge of German. The Japanese translation contained informal Japanese as the original conversation the translations are based on were between two familiar speakers. The translations were again checked for naturalness and up-to-dateness (i.e. whether a native speaker of approximately the same age as the original speakers in German would utter such a sentence or not.)

The sentences were then rerecorded in each language by female native speakers of similar age to the original speakers. I decided to also rerecord the German sentences for comparability. Each sentence was recorded at least twice and the better version in terms of similarity in intonation to the original version in German was chosen. For the first recording, the speakers were instructed to read out the sentence as if they were using it in an actual conversation without knowing the original turn they were reproducing. This was done to keep the recordings natural and avoid an "acting" character in the recordings. If however the intonation diverged to far from the original intonation, the speakers were given more precise instructions on the respective turns as for example information on where the stress should lie. The original turns were chosen in a way so that only few disfluencies were present. However, even if an original turn contained small disfluencies, they were not included in the re-recordings. This was done because it is nearly impossible to translate these disfluencies in the same position or respectively at the same word with the same character. Yet, repairs in the original turns were translated and kept in the re-recordings, as in the following example.

German (original) : Nee, ich glaub eher dass de Mark - dass Mark uns einladen würde.

English translation : Nope, I rather think that Mark - that Mark would invite us.

Japanese translation : Mushiro Hiro ga - Hiro ga watashitachi wo yonde kurerun janai.

Also filler sounds were translated and kept as in the following example.

German (original) : Also dann gehst du jetzt - ähm - rechtsrum um die Post. [...]

English translation : Well, then you go - ehm - right, around the post office. [...]

Japanese translation : Sore ja kore kara - ee - yuubinkyoku no tokoro wo migi ni magatte itte. [...]

The recordings were then again checked for naturalness by native speakers. Turns that were judged as unnatural in intonation were again rerecorded. All turns in all languages were then edited with Praat to obtain the versions needed to form the experimental conditions. This included amplitude normalization for all turns and a following low-pass filter at 500Hz was then applied to these turns to produce the no-words conditions. This step resulted in the creation of four conditions for each item. One condition contained the full version of the turn in German. The remaining three conditions contained the no-words version of the turn in German, English and Japanese. A full list of all items can be found in the Appendix.

4.1.2 List Creation and Counterbalancing

Four lists were created using a latin square within-subjects design. Thereby each participant was presented with every item only once but all conditions equally often and each condition for each turn was equally distributed across the four lists. As this would lead to an imbalance in the number of full turns and the number of no-word turns in each list (1:3), 48 filler items were included. These fillers were taken from the Lindenstraße-corpus and also rerecorded by the same native German speaker that also recorded the experimental turns for German. These turns as well were processed in Praat with amplitude normalization. The filler turns were only recorded in German and used only in their full version. Each list further contained ten practice items extracted from the newly recorded corpus. Half of the practice items were presented in the full German version and the other half in the no-words German version. Thereby, each list consisted of 96 experimental items and 48 fillers resulting in 144 items plus ten practice items. For the later analysis, additional information was added to the turns. To include AR as a factor, the syllables in each item and each language[2] were counted manually. An automatic Praat script to calculate SR and AR failed to cover all syllables, yet, the detection of pauses and their length was correct. Therefor, the automatically calculated pause number and length was used to calculate the AR[3]. Further additional information added to the turns were turn duration and number of pauses. A pause was marked as such when there was a turn-internal silence of 300 ms or longer.

[2] For a definition of a syllable in Japanese, a language that usually is based on mora rather than syllables, see subjection 2.1.3 "Japanese" in Chapter 2 "Background".

[3] For analysis and further description see 5.1.2 "Articulation Rate" in Chapter 5 "Statistical Analysis".

4.2 Experiment

I tested 40 participants (8 male, 32 female) in the age between 18 and 32. All participants were native speakers of German and received 5 Euros as compensation. The reaction time experiment was implemented and conducted with Experiment Builder[4].

On arrival, participants were seated in front of a computer and presented with the instruction screen. They were told, that they were listening to a conversation between three people in another room with only one person standing close enough to the door to be fully understood. The other two persons would stand further away so that their speech could only be heard as mumbling. This story was told to the participants to give them a natural reason for the presentation of the turns as well as an explanation for why the voices they heard differed. Their task was to anticipate the end of the current speakers turn and press a button when I believed the turn ended. Each item was played over earphones and presented with an empty screen. After the participant's button press, the turn was interrupted immediately to avoid a training effect leading to waiting for turn-end rather than anticipating it. An empty screen with silence was presented for one second after each item. Every 50 items, a pause was included that participants could end with another button press.

After the instructions, participants were presented with the ten practice items. If no questions remained, they could proceed to the main experiment with a button press. For each item, button presses were recorded. All participants listened to 144 turns with 96 experimental items.

The output file included bias estimated over the timing of the button press relative to the beginning of the turn minus the original turn length. Thereby, a negative bias indicates a button press before the actual turn end, whereas a positive bias indicates a button press after the actual turn end.

After the experiment, they were asked to note down their age and foreign languages they speak and give a self-assessment on their language skills according to the Common European Framework. All participants indicated English skills between the B1 and C2 level. Only two participants further added Japanese skills (A2 and B1 level).

[4] http://www.sr-research.com/eb.html

Chapter 5

Statistical Analysis

This chapter provides different statistical analyses performed on the collected data. All of the following analyses were performed on the data after outlier removal. Based on visualization, turns with a bias greater than 2500ms were treated as outliers as well as turns with a duration greater than 9000ms. The former eliminated eight responses (three responses to full German turns, three responses to German no-word turns and two responses to Japanese turns). Figure 5.1 shows the density plot of bias on the full data.

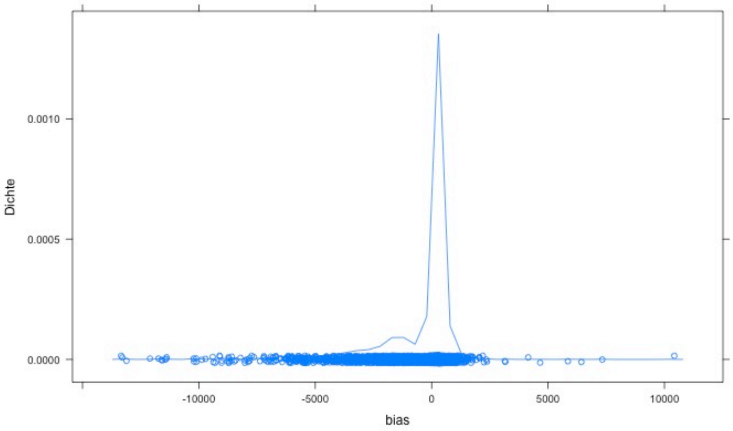


Figure 5.1: Density plot of bias on the full data set

The duration outlier removal eliminated nine full turns (seven Japanese turns and two English turns). Figure 5.2 shows the density plot of turn length on the full data set.

I conducted multiple statistical analyses on the collected data. In section 5.1 "Analysis of Speech Rate and Articulation Rate", I conduct analyses on SR and AR for the three

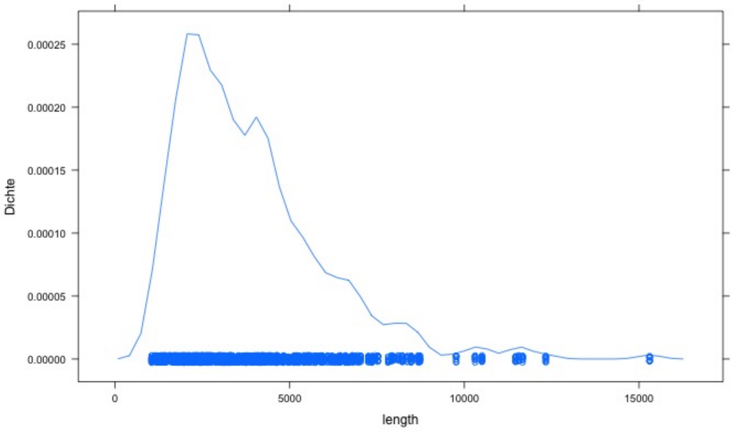


Figure 5.2: Density plot of turn length on the full data set

languages and justify the choice of AR as a factor for later analysis. Section 5.2.1 "Bias in full and no-word turns" shows the differences in bias for turns that offer access to lexicosyntactic information and turns that are reduced to prosody only. In section 5.2.2 "Bias in German and foreign turns" I show the analysis of the *no-words* condition of German and foreign language turns. Section 5.2.3 "Bias in English and Japanese turns" provides information about the difference in terms of bias found between the informativity of prosody for German listeners between the two foreign languages. Finally, section 5.4 "Bias by turn type" investigates whether there is a difference in terms of turn-end anticipation expressed in bias for the three sentence types (Questions, SVO, SOV) in the corpus.

5.1 Analysis of Speech Rate and Articulation Rate

Speech Rate (SR) and Articulation Rate (AR) are reflecting the fastness of an articulation as perceived by listeners. Therefore, a difference in either could have an influence on turn-taking behavior. In particular, because they could differ not only by language, as described in Section 2.2.1 *Speech Rate*, but also by speaker, they make a potential confound. In this Section, I compute both, SR and AR, in an automatic approach and a manual approach and show which of them is more likely to have an effect on listeners.

5.1.1 Analysis of Speech Rate

In a first step, a Praat script to automatically compute SR (De Jong and Wempe, 2009) was used. Pauses are ignored to receive Speech Rate by dividing turn duration by the number of syllables in each turn. The German full and no-words condition were treated as one as SR is not different for the two conditions. The results are summarized in Table 5.1. Figure 5.3 visualizes the SR per language in a box plot.

German	English	Japanese
4.762500	3.927083	4.485208

TABLE 5.1: Means of automatically computed SR in syllables per second for the three languages

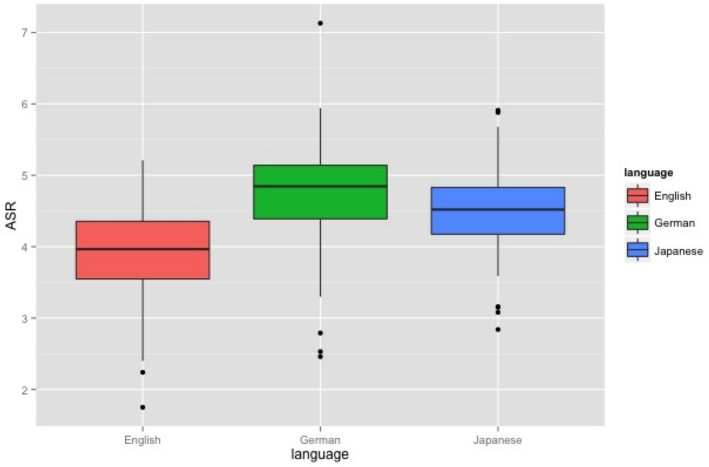


FIGURE 5.3: box plot of the automatically computed SR per language

However, the syllable count the output provides differs strongly from the actual syllable count conducted manually and would lead to the conclusion that SR differs significantly across the three languages as is shown by a repeated measures ANOVA, F(2,190) = 45.94, $p < 0.001$, $\eta_p 2 = 0.33$, and a follow up pairwise comparisons with Bonferroni adjusted p values, which suggests that German (M = 4.76, SD = 0.73) had a significantly higher SR than both, English (M = 3.93, SD = 0.67) and Japanese (M = 4.49, SD = 0.58), both $p < 0.01$. The results suggest further that Japanese had a significantly higher SR than English, $p < 0.001$. However, especially the results concerning German having a higher SR than Japanese contradicts previous findings (Braun and Oba, 2007, Pellegrino et al., 2011, 2004). The very different results of the automated syllable count compared to the manual syllable count and the contradiction with previous studies showed that the automatic approach is not reliable. Therefore, a re-analysis with the manually counted syllables was conducted. Table 5.2 summarizes the means of SR for the three languages.

German	English	Japanese
6.058662	4.691473	5.810232

Table 5.2: Mean SR in syllables per second for the three languages

Repeated measures ANOVA indicated that the different languages had significantly different Speech Rates, F(2,190) = 131.08, $p < 0.001$, $\eta_p 2 = 0.58$. Pairwise comparisons with Bonferroni adjusted p values revealed that English had a significantly lower SR (M = 4.69, SD = 0.75) compared to German (M = 6.06, SD = 0.69), $p < 0.001$, and Japanese (M = 5.81, SD = 0.7), $p < 0.001$. There was no significant difference between German and Japanese SR. Figure 5.4 shows a box plot of the SR per language.

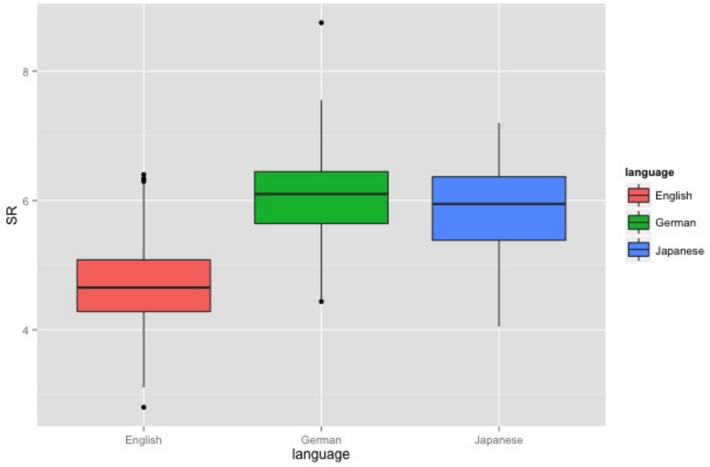


Figure 5.4: box plot of the SR per language

Given the participants perception of Japanese turns sounding the fastest (unknowingly those turns were Japanese) in addition to the lacking difference between German and Japanese with the tendency for German to have a higher SR indicates that SR is not necessarily reflecting perception. Therefore, Articulation Rate was investigated next.

5.1.2 Analysis of Articulation Rate

In order to calculate the AR, the total length of all pauses within a turn was subtracted from the full turn duration before it was divided by the number of syllables. This ensures that only spans of actual speech are taken into account. Even though the script used to obtain the automatically computed SR failed to detect all syllables in each turn, the pause detection in terms of both, length and number, have been very accurate. This was made sure of by manual inspection of the data. Therefore, the automatically computed number and length of pauses was used to obtain the AR. A pause was marked as such when there was a turn-internal silence of 300 ms or longer. For the automatic pause extraction the threshold for silence was set to -25 db. Table 5.3 summarizes the means of the AR for the three languages.

German	English	Japanese
6.079225	5.135669	6.610058

TABLE 5.3: Mean AR in syllables per second for the three languages

Repeated measures ANOVA indicated that the different languages had significantly different Articulation Rates, F(2,190) = 138, $p < 0.001$, $\eta_p 2 = 0.59$. Pairwise comparisons with Bonferroni adjusted p values revealed that English had a significantly lower AR (M = 5.14, SD = 0.75) compared to German (M = 6.08, SD = 0.7), $p < 0.001$, and Japanese (M = 6.61, SD = 0.74), $p < 0.001$. There was also a significant difference between German and Japanese AR, $p < 0.001$. Figure 5.5 shows a box plot of the AR per language.

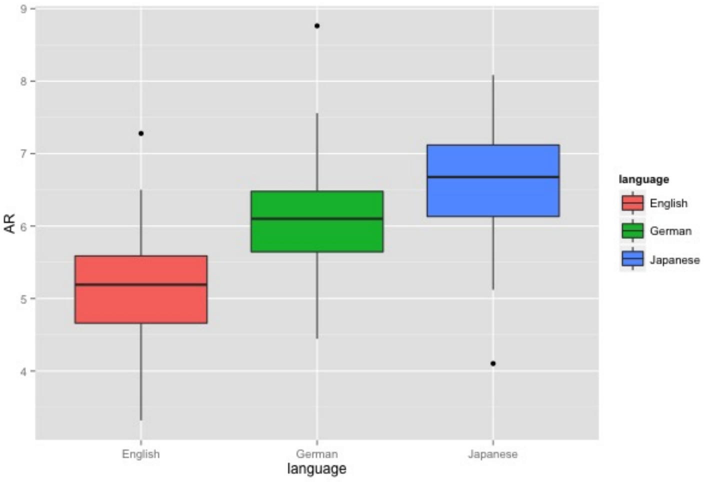


FIGURE 5.5: box plot of the AR per language

This finding reflects participants perception and are supported to some extent by Pellegrino et al. (2011, 2004), Braun and Oba (2007). It is to be noted though, that all three of their findings suggest a higher similarity in terms of SR and AR between English and German. This is most likely due to individual differences of the speakers in this experiment. It is likely that an average across different speakers for each language would result in an adjustment of the SR and AR per language. All previous findings suggest Japanese to posses the highest AR of the three languages calculated on multiple speakers per language. Therefore, AR was included as a factor for the later analysis rather than SR because of the higher similarity to previous findings in combination with the support by participants' perception. Also, the negative correlation between bias and AR (cor = -0.13, $p < 0.001$), which is visualized in Figure 5.6, qualified AR as a factor in the later analysis.

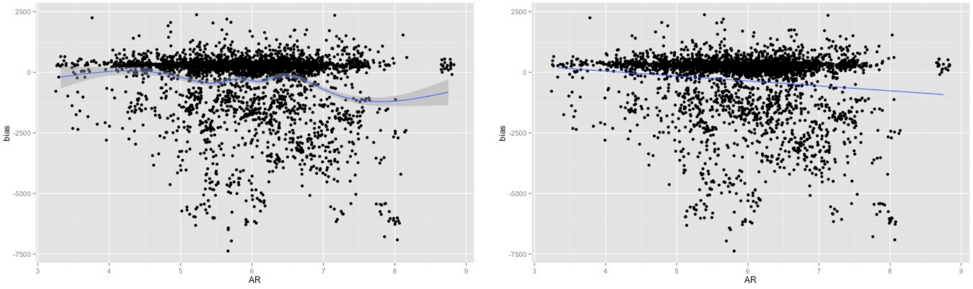


Figure 5.6: Correlation between bias and AR. The right side shows the linear fit.

In summary, a higher AR leads to a more negative bias.

5.2 Analysis of the Full Data Set

This analysis was conducted on the complete data set excluding the outliers as described in the beginning of this chapter. Table 5.4 summarizes the mean length and mean bias of all four conditions.

	German (full)	German (no-words)	English	Japanese
Mean length	3404.668	3400.953	3828.202	4159.023
Mean bias	-315.195	-56.524	-450.399	-609.713

Table 5.4: Means of length and bias in the full data set after outlier removal categorized by language[1]

These results further allow for a direct comparison to the findings of De Ruiter et al. (2006) and Gambi et al. (2015). All three findings, including those of this thesis, find a bias close to -300 ms in their most natural condition, e.g. the full version of the turn that was not alternated. This is summarized in Table 5.5. The difference between the findings of De Ruiter et al. (2006) and Gambi et al. (2015) and those in this thesis can possibly be explained by the language the three experiments were conducted on. It is well possible that Dutch, which was in the scope of De Ruiter et al. (2006) research, differs from German that was the scope in both Gambi et al. (2015) and this thesis.

Research	De Ruiter (2006)	Gambi (2015)	This thesis
Mean bias in the natural condition	-186	-315	-315

Table 5.5: Means of bias in the natural condition in De Ruiter et al. (2006), Gambi et al. (2015) and this thesis.

I used linear-mixed effects models with maximal random structure and defined 3 planned contrasts to test 1) whether having lexicosyntactic information on top of prosody increases the accuracy of turn-end anticipation. 2) whether German listeners are more accurate in turn-end anticipation when they are presented with German intonation compared to foreign intonation. 3) whether for German listeners there is a difference in accuracy of turn-end anticipation on English intonation and Japanese intonation. The three contrasts are orthogonal. Table 5.6 provides a summary of the contrasts.

Contrast	C1	C2	C3
Description	Full access to lexicosyntactic information vs. All no-words conditions	German no-words condition vs. Both foreign no-words conditions	English no-words condition vs. Japanese no-words condition

Table 5.6: Overview of the contrasts used in the analysis of the full data set.

Following Barr et al. (2013), I started with a maximum random structure. As factors I chose the three contrasts in order to rule out possible singular effects of an item or participant on either of the contrasts. Further, I included AR as described in section 5.1 "Speech Rate and Articulation Rate" and Duration due to the wide spread length of turns within and between conditions. As in previous studies, I found a negative correlation between bias and turn duration (cor = -0.57, $p < 0.001$) with longer turn duration leading to more negative bias. Figure 5.7 visualizes the correlation between bias and turn duration. The correlation also holds for all four conditions separated as visualized in 5.8.

Additionally, I included Pause as a factor. This is to rule out that the results are driven by short pauses that occur within a turn because they are interpreted as the end of the turn. A correlation test showed that the more pauses there are within a turn, the more

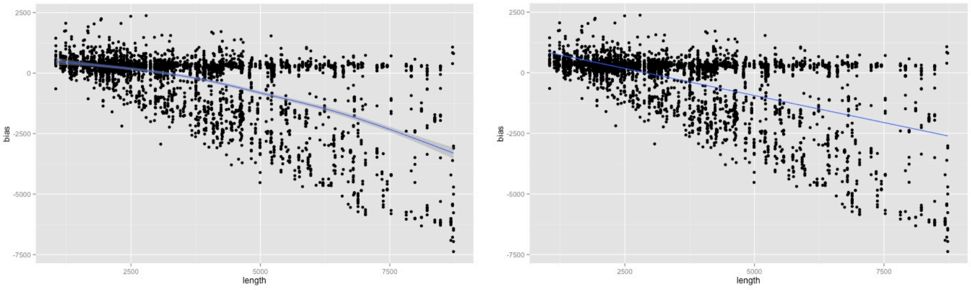


Figure 5.7: Correlation between bias and Duration coded as length. The right side shows the linear fit.

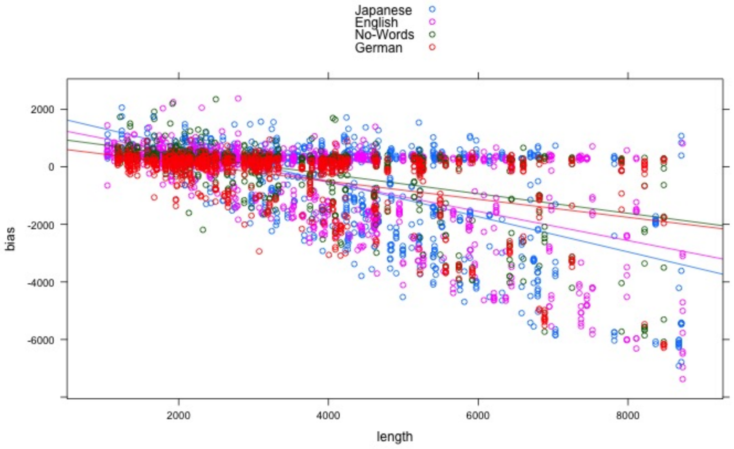


Figure 5.8: Correlation between bias and Duration (coded as length) per condition.

negative is the bias (cor = -0.36, p < 0.001). To keep the complexity reasonably low, pauses were simply coded as present or absent, which still preserved the correlation (cor = -0.29, p < 0.001). This coding also solves an additional issue that might arise with a high number of pauses. These are only possible in comparably longer turns. Thereby, a higher number of pauses could be confounded with turn duration. Yet, as about two thirds of the turns in the corpus were without pause, some long turns are free of pauses, which justifies this additional factor. Table 5.7 summarizes how the number of pauses is distributed within the corpus after outlier removal.

No. of pauses	0	1	2	3	4	5
No. of turns	236	99	32	5	1	2

Table 5.7: Distribution of pauses per turn in the corpus after outlier removal

The data was also residualized on Duration, AR, and Pauses to avoid colinearity between the factors.

The various factors included lead to a very complex model. As the model was not converging, the random structure had to be simplified. In order to decide how to simplify the random structure, I followed a parsimonious mixed models approach as described by Bates et al. (2015). In this approach, factors or interactions between factors are dropped according to the variance in the data they can account for. In order to see how many interactions in the random effect structure contribute to the explanation of the variance in the data, a Principal Components Analysis (PCA) of the random-effects variance-covariance estimates from the mixed-effects model has to be performed. The PCA orders the components of the random effect structure by their contribution to the explanation of the variance. According to Bates et al. (2015), the number of components that explain zero or close to zero variance should be dropped from the model. Yet, the PCA does not provide information on what components are liable for those values. Therefore, the model's random effects have to be consulted. Here, components with the lowest explanation for the variance are related to the components of the PCA with close to zero values. After a few iterations, this approach led to an elimination of AR and the second contrast from the random structure on items. The random structure on subjects remained unchanged. In summary the random structure on subject included the three contrasts, duration and AR, including all interactions between duration and AR with any of the contrasts as well as the intercept and Pause. The random structure on item included the first and third contrast together with duration, including the interactions of duration with both contrasts as well as the intercept and Pause. Figure 5.9 shows this model in R code.

```
residuals(res) ~ C1 + C2 + C3 +
(0+Pause | Subject) + (1+(C1+C2+C3)*Duration*AR || Subject) +
(0+Pause | item) + (1+(C1+C3)*Duration || item)
```

Figure 5.9: Model description in R

5.2.1 Bias in Full and No-word Turns

The contrast of full access to lexicosyntactic information and mere access to prosodic information improved model fit significantly ($\chi^2(1)$=7.77, p $<$ 0.01), indicating that the additional access to lexicosyntactic information leads to a more accurate anticipation of turn-end displayed by bias (B=191.62, SE=65.99, t=2.904). However, this result is obtained by comparing the full German turns to all of the three no-words conditions, which includes different languages. Also judging from the means of bias per condition, this does not necessarily hold for the comparison of the full German turns with only the no-word turns of German, which might yield more precise results for the question whether lexicosyntactic information is helpful on top of only prosodic information. Therefore, I further conducted an analysis of a subset of the data only consisting of the two German conditions. I again used linear-mixed effects models with maximal random structure and defined one planned contrast as described above. In order to keep the results comparable, the same random effect structure was used including the contrast, duration, AR and presence of pauses. Again the contrast of full access to lexicosyntactic information and mere access to prosodic information improved model fit significantly ($\chi^2(1)=5.21$, $p < 0.05$), yet the direction is different, indicating that the additional access to lexicosyntactic information leads to a more negative bias than mere access to prosodic information (B=-269.32, SE=54.36, t=-4.955).

5.2.2 Bias in German and Foreign Turns

The contrast of German prosody and foreign prosody also significantly improved model fit ($\chi^2(1)=7.77$, $p < 0.01$). This indicates that German listeners are more precise in turn-end anticipation if presented with their native intonation than they are with foreign intonation (B=-108.85, SE=45.78, t=-2.378).

5.2.3 Bias in English and Japanese Turns

The contrast of English and Japanese prosody did not improve model fit ($\chi^2(1)=2.15$, $p > 0.1$). This indicates that intonation different from the native intonation is not differentiated and leads to equal turn-taking behavior (B=98.74, SE=66.33, t=1.489).

5.3 Intonation or Prosody

It is arguable that the previous analysis mostly focuses on intonation rather than complete prosody as Pauses and AR are included as predictors. Yet, both of them are substantial parts of prosody by definition. Therefore, using them as predictors might to some extent eliminate their influence, as they are no longer considered as substantial part of the languages prosody but as overall factors more or less unrelated to the respective language they came from. For that reason, I additionally reanalyzed the data with a slightly different model. In this model, the data was only residualized on duration and, besides the contrast, duration also was the only additional factor in the random effect structure. Thereby, AR and pauses are no longer independent predictors but are encoded in the contrasts together with intonation to form the complete prosody. This allows to test the complete prosody with only the contrasts instead of teasing the three factors apart. Overall, condition significantly improved model fit ($\chi^2(3)$=21.41, p $<$ 0.001). However, the individual fixed effects of the contrasts differ. In perspective of full prosody, having full lexicosyntactic access over mere prosodic contour is not significantly different (B=130.02, SE=70.29, t=1.85). The other results however remained unchanged from the results of the previous analyses.

5.4 Bias by Turn Type

In this section I investigate whether there is an influence of turn type (Question, SVO and SOV) on bias. For this reason I introduced a further entry to my data called *TStype* (true sentence type). The value of this entry provides the real sentence structure of a turn. This means that all German turns kept their original sentence type label, but all Japanese turns that were not labeled as questions were labeled as SOV and all English turns that were not questions as SVO. This is done because Japanese does not posses any sentences of the structure SVO and English does not posses any sentences of the structure SOV. Table 5.8 summarizes the means of length and bias of the three sentence types of all conditions.

	QUE	SVO	SOV
Mean length	2822.978	3818.881	4289.110
Mean bias	20.593	-416.786	-603.351

Table 5.8: Means of length and bias categorized by turn-type

I used linear-mixed effects models with maximal random structure (Barr et al., 2013) and defined 2 planned contrasts to test whether question intonation leads to a more precise turn-end anticipation (i.e., a bias closer to zero) compared to other sentence structures and whether there was a difference in bias between SOV and SVO structures and report estimates and Wald t tests for fixed effects. Overall, sentence type did not improve model fit ($\chi^2(4)$=7.05, p = 0.13). However, a tendency of bias being closer to zero in questions compared to other sentence structures was found (B=220.89, SE=93.01, t=2.4). In addition, there was no differences between SVO and SOV sentence types (B=-94.15, SE=95.2, t=-0.989), indicating that German listeners can interpret declarative sentence intonation equally accurately independent from sentence structure.

As especially the proximity of bias to zero for questions is noticeable, I additionally grouped the bias and length of questions by language as summarized in Table 5.9.

	Questions			
	German (full)	German (no-words)	English	Japanese
Mean length	2601.767	2596.821	2894.148	3214.926
Mean bias	-34.862007	137.536	-3.177	-19.606

TABLE 5.9: Means of length and bias of questions categorized by language

Given the extreme data sparsity when narrowing the data set down to only the questions in the no-words condition, analysis performed on this small data set should be taken with caution. This also holds for the relative accuracy in turn-end anticipation displayed by the close to zero values of bias in all three languages. However, all four conditions exhibit a mean bias close to zero in question-type turns. A linear-mixed effects models with maximal random structure (Barr et al., 2013) using the same random effect structure as in 5.2 *Analysis of the full data set* in order to receive comparable results and 3 defined planned contrasts was used to test whether full lexicosyntactic access in question-type turns leads to a more precise turn-end anticipation (i.e., a bias closer to zero) compared to only the accessibility of prosodic information. Additionally the model tests for the same type of difference between German and foreign intonation as well as for differences between the English and Japanese question intonation. I report estimates and Wald t tests for fixed effects.

Overall, condition did improve model fit significantly ($\chi^2(3)$=16.045, p = 0.001). However, this is mainly driven by the contrast of full access to lexciosyntactic intformation and the remaining conditions. Only the languages in the no-word conditions did not improve model fit ($\chi^2(2)$=3.328, $p > 0.05$). The full access to lexicosyntactic information leads to a more positive bias compared to the conditions for which lexicosyntactic information was made unintelligible (B=-244.725, SE=65.67, t=-3.7). This reflects the findings of the analysis of the full data set. However, contrary to the analysis including all turn types, foreign intonation does not lead to a less precise turn-end anticipation as displayed by the missing significance in the comparison of German and foreign intonation (B=31.07, SE=52.1, t=0.6). Again, there was no differences between English and Japanese intonation (B=-161.7, SE=93.97, t=-1.72). This hints towards the possibility of resemblance of question intonation across the three tested languages.

Chapter 6

Discussion

This chapter provides an interpretation of the results of the statistical analysis and offer some suggestions for possible future work based on the findings of this thesis.

6.1 Interpretation

The results of the analysis show that prosody is very language specific. The significance of the contrast between German prosody and foreign prosody showed that German participants were significantly better in turn-end anticipation on their native prosody than on the foreign prosodies. This leads to the conclusion that prosody is very language specific. The lacking significance of the contrast between Japanese and English in terms of accuracy in turn-end anticipation further backs this interpretation. It shows that there is no difference between a presumably closer prosodic pattern in a related language such as the second Germanic language English in the experiment and a very distant language as the isolated language Japanese. If the prosody is not coming from the native language, utilizing it as a turn-taking cue seems to be hard to impossible. The very close to zero mean bias of the German no-words condition on the other hand shows, that prosody alone already provides a helpful cue to manage turn-taking. This finding contradicts the findings of De Ruiter et al. (2006). Yet, this experiment and the experiment mentioned before differ in one possibly important point. In this experiment, only turns in which turn-taking succeeded in the original conversation were used for the experimental items. Turns with overlaps and longer pauses were not used. In both experiments of De Ruiter et al. (2006) and Gambi et al. (2015) turns containing overlaps and turns containing pauses were used together with successful turns (in terms of turn-taking) to an equal amount.

6.1.1 Possible Differences between Turn Types

It is arguable to what extent lexicosyntactic and prosodic cues differ in case of turns that lead to smooth turn-taking and such turns that lead to inaccurate turn-taking behavior. Yet, it is reasonable to assume that differences are present, as there presumably is a reason for why the latter lead to inaccurate turn-taking in the first place. In case of occurring overlapping speech, this would lead to the assumption that a cue that is usually utilized to indicate a turn-end appears in a mid-turn position, which leads to a false interpretation of a turn-end by the interlocutor. On the other hand, in case of pauses in between two turns, cues that should indicate the end of a speakers turn are supposedly missing, which leads to a false interpretation of an ongoing turn by the interlocutor. The corpus recorded for this study also backs this assumption. As for example displayed in the following excerpt[1]

A: [...] oder ihr geht zu Mario, oder sowas. Und du gehst nicht...

B: Nee, die kommen halt zu uns.

In this excerpt, speaker B starts speaking when speaker A utters *oder sowas* (*or something.* In other situations, this short phrase also appeared mid-turn, but was correctly interpreted as such. In the excerpt, the intonation is falling on the name *Mario* which could have lead to the misinterpretation of a turn-end by speaker B.

If the assumption of misuse of cues in turns that do not lead to smooth turn-taking holds, it is reasonable to separate such turns from turns leading to smooth turn-taking in experimental designs as it was done in this experiment. In other words, if the misleading turns contain misused prosodic cues, it is hard to infer the usefulness of prosody altogether if all three types of turns are mixed.

[1] A: [...] or you're going to Mario's or something. And you're not going...
B: Nope, they are coming to our place.

6.1.2 The optimal Bias

The significant effect of the contrast describing the German turns containing lexicosyntactic information in contrast to the German turns containing prosody only, seems to suggest that prosody alone can even lead to better turn-end anticipation than a full access to lexicosyntactic information. However, this finding does not necessarily lead to the conclusion that the lack of lexicosyntactic information yields a more precise anticipation of turn-ends. Integrating the difference in time needed to prepare a vocal response (approximately 600 ms as described by Indefrey and Levelt (2004), Jescheniak et al. (2003) and Schnur et al. (2006)) compared to the time needed for a motor response (approximately 360 ms as described by Ng and Chan (2012)), a difference of about -240 ms to the turn-end should be expected. Considering this, the mean bias of turns with full access to lexicosyntactic information of -315 ms is closer to a possible optimal value than the only -56 ms for turns with access to prosodic information only. Following this line of reasoning, the latter would lead to slightly too late responses (approximately 200 ms) if the difference of the task in the experiment to actual conversations was regarded. This also aligns with the findings of De Ruiter et al. (2006) and Gambi et al. (2015) for their most natural condition. It is reasonable to assume that the bias found for the natural condition is most likely to mirror behavior in natural conversations.

It is still to be considered though that the missing context for the turns that participants react to in all named experiments might impair the results. If so, this could possibly explain the divergence from the findings of Weilhammer and Rabold (2003) and Wilson and Wilson (2005). Both studies show that even turn transitions that are perceived as smooth actually contain gaps of approximately 300 to 400 ms. Considering this in combination with Indefrey and Levelt (2004), Jescheniak et al. (2003), Schnur et al. (2006) and Ng and Chan (2012) findings, an experiment with motor responses on turns that in natural conversations would elicit a vocal response should show a mean bias of about 60 to 160 ms. The computation of this value is as follows. When the 600 ms needed for the preparation of an articulatory response are subtracted from the bias of 300 to 400 ms a bias of –300 to –200 ms is reached. When adding the 360 ms needed for an motor response to that value, a bias of 60 to 160 ms is obtained. However, the actual value of gaps between consecutive turns could also differ across languages as indicated by the findings of Stivers et al. (2009). They showed that the mean time of turn-transitions slightly differs across languages[2].

With the current state of knowledge it is still not possible to exactly tell which bias would be displaying an ideal turn-transition in experiments with similar tasks as the ones performed here and by De Ruiter et al. (2006) and Gambi et al. (2015). Yet, considering all findings of previous works and this thesis, it is likely that a bias displaying an optimal turn-transition in this kind of experimental setup lies in a range from -300 ms to 160 ms.

6.1.3 Noise and Meaningful Signals

Another important difference between this experiment and the one conducted by De Ruiter et al. (2006) lies in the compared conditions. In De Ruiter et al. (2006) only an improve in terms of precision on turn-end anticipation from mere noise compared to additional prosodic information was presented. In this experiment, I showed that two just as meaningful signals as presented by native intonation and foreign intonation also show a significant difference. This shows that listeners can not interpret the foreign prosody although it is just as rich in information as the native prosody. However, it is to be noted that I did not include a comparison to noise. It is possible that foreign prosody is still advantageous compared to a condition not containing any meaningful signal. Yet, if in fact prosody was not helpful for turn-end anticipation as proposed by De Ruiter et al. (2006), the differences in bias between native and foreign prosodies should not be existent.

6.1.4 Universality of Prosodic Turn-taking Cues

As mentioned above, I showed that prosody is very language specific and therefore can not be used universally for turn-taking management across languages. However, there appears to be an exception for question intonation. All three languages share a rising pitch contour in question-type turns. This is displayed in the Figures 6.1, 6.2 and 6.3. The analysis of only the question-type turns also shows this with missing significant differences in precision of turn-end anticipation for questions across the three languages. Even considering the data sparsity of only questions in the experiment, this at least indicates a possible universality of prosody in questions as a turn-taking management

[2] For comparison see Figure 3.3 in Chapter 3.3.2 *Japanese Example for the Importance of Prosody.*

cue across languages. However, again it is unclear what bias would display an ideal turn-transition. Therefore, question intonation with a mean bias of 20 ms could also display slightly to late responses. Yet, the value lies in the predicted range of an optimal bias. Additionally, with the bias mean of questions being below 300 ms, it is clear that an anticipation took place as motor responses need a preparation time of about 360 ms.

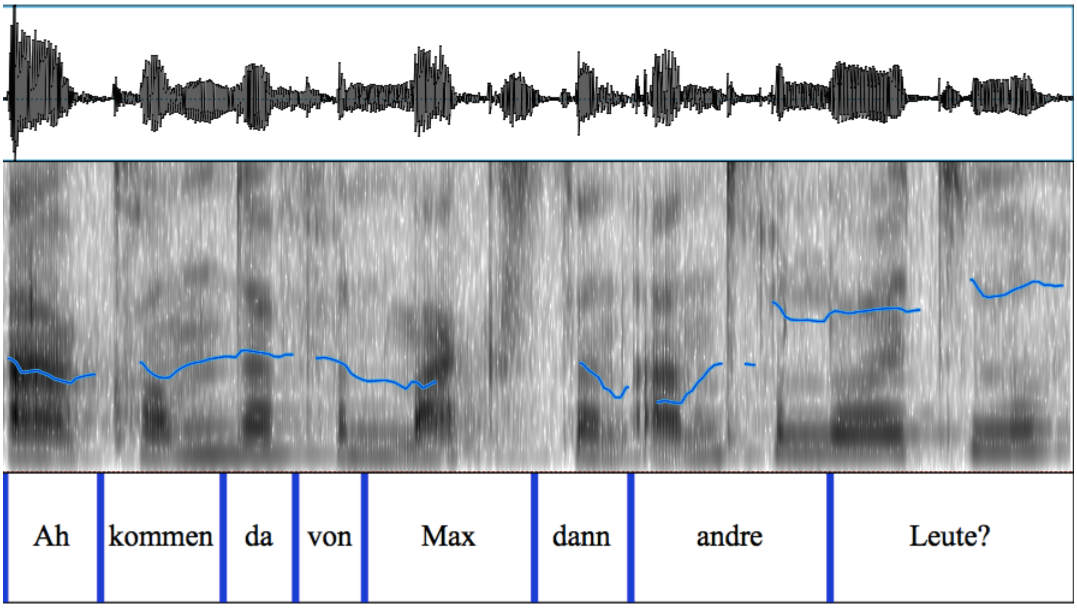


Figure 6.1: Example of a German question-type turn with pitch contour and transcription. The for question intonation typical rise lies on the penultimate word *andre* and remains high on the final word with a slight falling-rising movement.

6.2 Future Work

As the main question behind studies like this one and those conducted by De Ruiter et al. (2006) and Gambi et al. (2015) is what leads to the usually accurate and smooth turn-taking in natural conversations and how we manage turn-taking, it seems reasonable to conduct such research on turns that are in fact leading to the desired smooth behavior. Although this thesis already hints to the necessity of such an experimental setup, it lacks the comparison of turns leading to smooth turn transitions and such that do not. In order to further support the results yielded by this experiment, it therefore would be necessary to also analyze data that compares accuracy in turn-end anticipation between

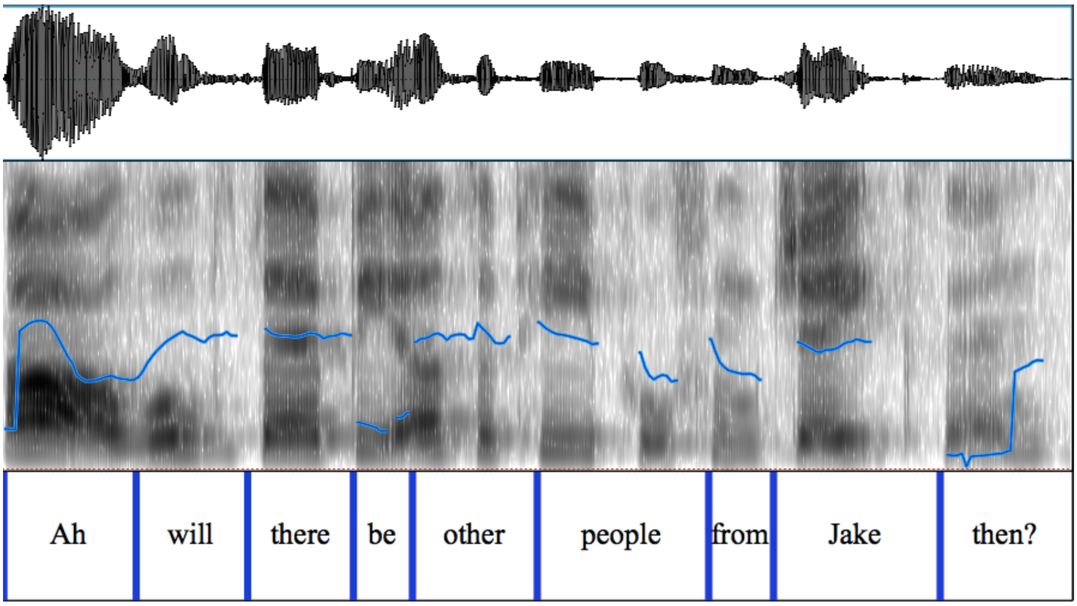


Figure 6.2: Example of a English question-type turn with pitch contour and transcription. The for question intonation typical rise lies on the final word *then* with a falling movement between the penultimate word *Jake* and the final word.

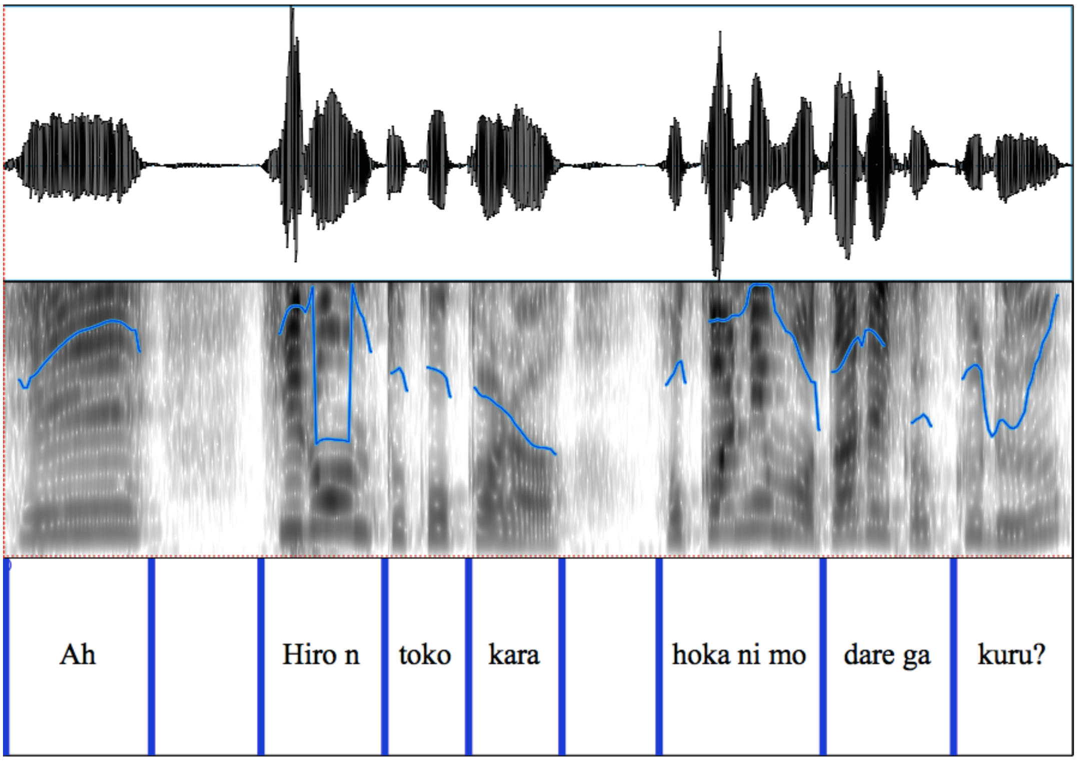


Figure 6.3: Example of a Japanese question-type turn with pitch contour and transcription. The for question intonation typical rise lies on the final word *kuru* with a falling-rising movement.

turns leading to smooth turn-taking behavior and the other two types of turns with flawed turn-transitions. This could be easily done by conducting a similar experiment as the ones of De Ruiter et al. (2006) and Gambi et al. (2015) and in this thesis, but with a high and equal number of turns leading to smooth turn-taking, turns leading to pauses and turns leading to overlaps. To support the assumptions made in this thesis, it would be necessary to see significant differences between these three conditions. By testing full turns, containing all lexicosyntactic information in addition to the prosodic information, such an experiment could show an overall difference between the turn types, whereas an experiment comparing the no-words conditions could further show whether these differences are also reflected by the prosodic information alone.

As shown before, based on the results of this thesis it seems unreasonable to deem prosodic information unimportant for turn-taking management. Therefore, following the experimental setup proposed before, it would also be reasonable to rebuild the experiments of De Ruiter et al. (2006) and Gambi et al. (2015) using only such turns that lead to smooth turn-taking behavior in order to receive a deeper insight into the usefulness of prosodic information for turn-taking management.

Additionally, it would be interesting to conduct the same experiment that was done for this thesis also in the other two languages as a baseline. Such a study could yield information about the restrictiveness of prosodic cues across the different languages. For example, it might be possible that the cues of one language are hard to interpret for interlocutors of a different languages but not the other way around. Such information could lead to a more thorough insight into the richness of prosody used as turn-transition relevant cues across different languages.

For all of the above named possible future works it would also be beneficial to find out what bias in a reaction time experiment with motor responses would display an ideal and smooth turn-transition. Given a certain range in which this optimal bias could lie, current works can only make statements about the differences between certain cues but can not tell which cue is more beneficial.

6.3 Summary

In this thesis, I was able to show that prosody is very language specific and that relatedness of languages seems to be of little help. Therefore, even though some prosodic patterns seem to be similar across the three languages tested in this thesis as shown by the comparison of Selting (1995), Duncan (1972), Gravano and Hirschberg (2011) and Tanaka (2004), there are apparently more fine grained differences in their prosodic contours. Furthermore I showed that prosody is relevant for turn-taking management, contrary to the results of previous studies. This also shows that the understanding of turn-taking behavior is yet open to many questions and that the experimental setup and the choice of turn types for experiments might be crucial to future studies.

Appendix A

Map Task

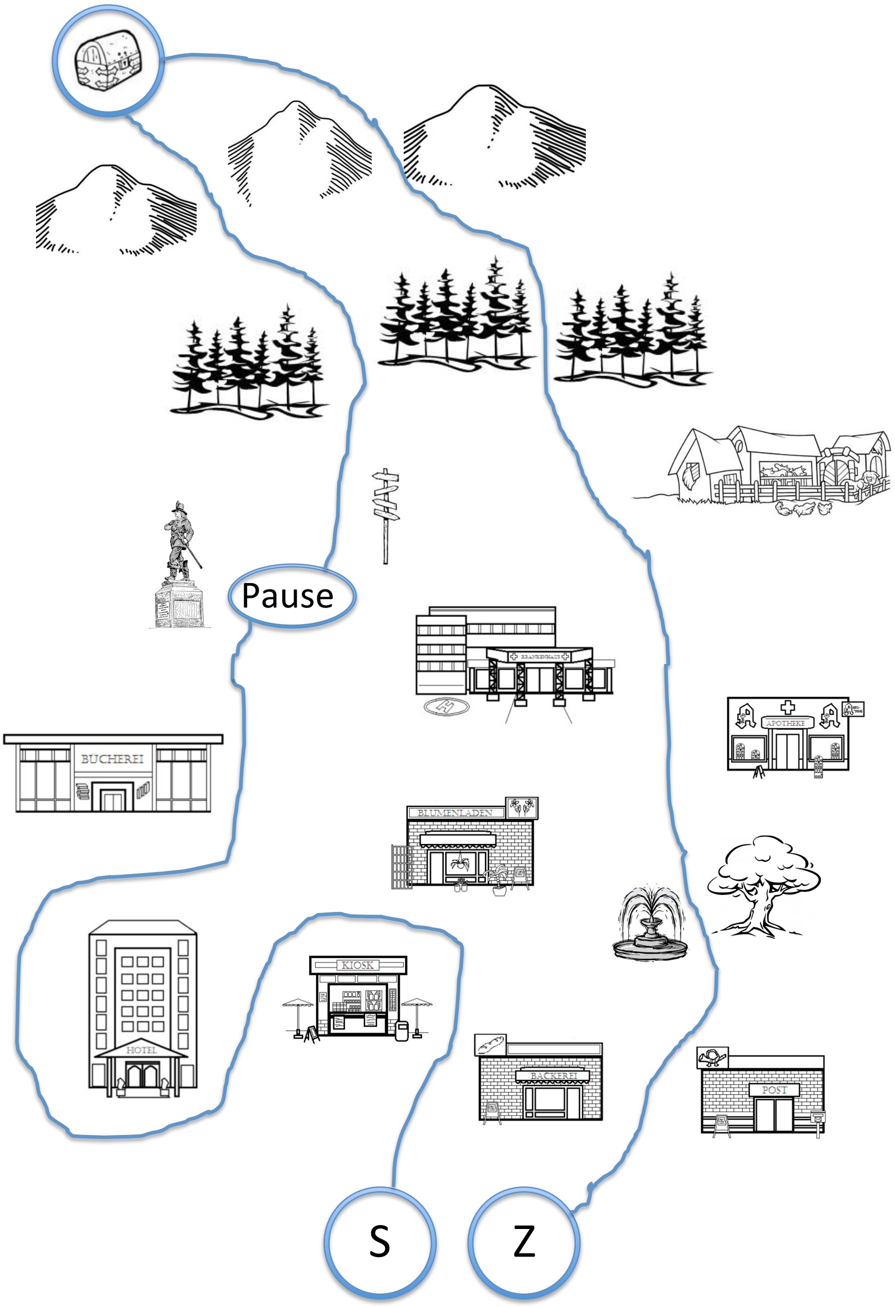


Figure A.1: The map handed to the instructor containing the path that was to be explained and six differences from the searchers map.

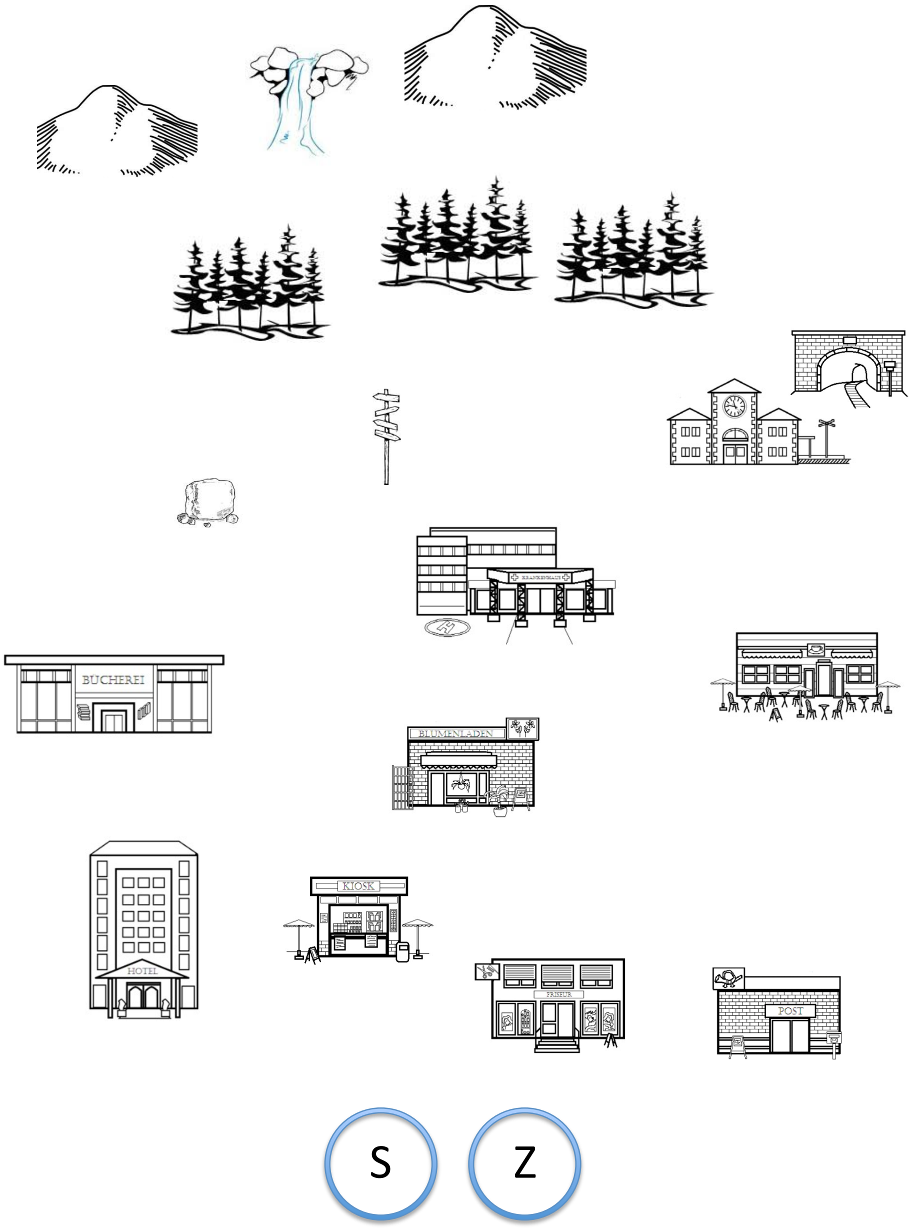
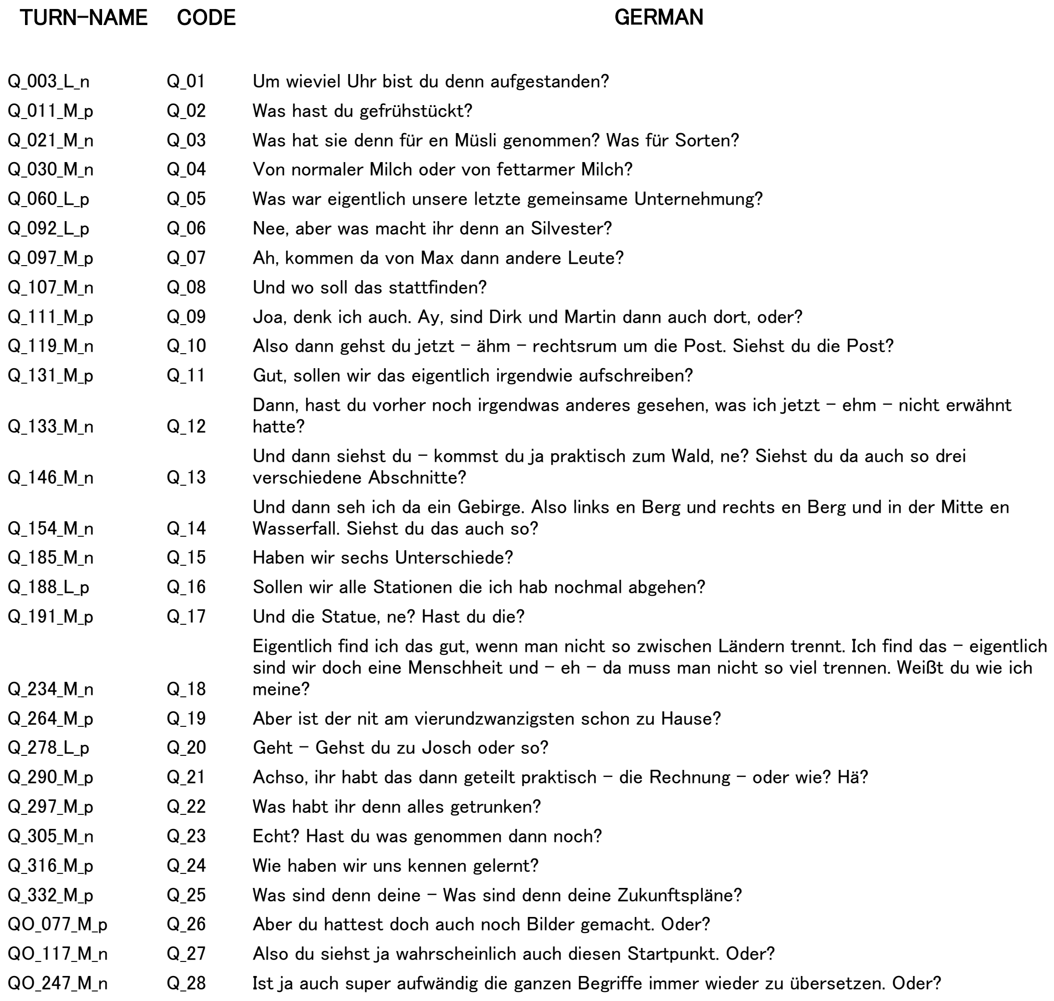


Figure A.2: The map handed to the searcher with six differences from the instructors map but without the path.

Appendix B

Corpus



TURN-NAME	CODE	GERMAN
Q_003_L_n	Q_01	Um wieviel Uhr bist du denn aufgestanden?
Q_011_M_p	Q_02	Was hast du gefrühstückt?
Q_021_M_n	Q_03	Was hat sie denn für en Müsli genommen? Was für Sorten?
Q_030_M_n	Q_04	Von normaler Milch oder von fettarmer Milch?
Q_060_L_p	Q_05	Was war eigentlich unsere letzte gemeinsame Unternehmung?
Q_092_L_p	Q_06	Nee, aber was macht ihr denn an Silvester?
Q_097_M_p	Q_07	Ah, kommen da von Max dann andere Leute?
Q_107_M_n	Q_08	Und wo soll das stattfinden?
Q_111_M_p	Q_09	Joa, denk ich auch. Ay, sind Dirk und Martin dann auch dort, oder?
Q_119_M_n	Q_10	Also dann gehst du jetzt – ähm – rechtsrum um die Post. Siehst du die Post?
Q_131_M_p	Q_11	Gut, sollen wir das eigentlich irgendwie aufschreiben?
Q_133_M_n	Q_12	Dann, hast du vorher noch irgendwas anderes gesehen, was ich jetzt – ehm – nicht erwähnt hatte?
Q_146_M_n	Q_13	Und dann siehst du – kommst du ja praktisch zum Wald, ne? Siehst du da auch so drei verschiedene Abschnitte?
Q_154_M_n	Q_14	Und dann seh ich da ein Gebirge. Also links en Berg und rechts en Berg und in der Mitte en Wasserfall. Siehst du das auch so?
Q_185_M_n	Q_15	Haben wir sechs Unterschiede?
Q_188_L_p	Q_16	Sollen wir alle Stationen die ich hab nochmal abgehen?
Q_191_M_p	Q_17	Und die Statue, ne? Hast du die?
Q_234_M_n	Q_18	Eigentlich find ich das gut, wenn man nicht so zwischen Ländern trennt. Ich find das – eigentlich sind wir doch eine Menschheit und – eh – da muss man nicht so viel trennen. Weißt du wie ich meine?
Q_264_M_p	Q_19	Aber ist der nit am vierundzwanzigsten schon zu Hause?
Q_278_L_p	Q_20	Geht – Gehst du zu Josch oder so?
Q_290_M_p	Q_21	Achso, ihr habt das dann geteilt praktisch – die Rechnung – oder wie? Hä?
Q_297_M_p	Q_22	Was habt ihr denn alles getrunken?
Q_305_M_n	Q_23	Echt? Hast du was genommen dann noch?
Q_316_M_p	Q_24	Wie haben wir uns kennen gelernt?
Q_332_M_p	Q_25	Was sind denn deine – Was sind denn deine Zukunftspläne?
QO_077_M_p	Q_26	Aber du hattest doch auch noch Bilder gemacht. Oder?
QO_117_M_n	Q_27	Also du siehst ja wahrscheinlich auch diesen Startpunkt. Oder?
QO_247_M_n	Q_28	Ist ja auch super aufwändig die ganzen Begriffe immer wieder zu übersetzen. Oder?

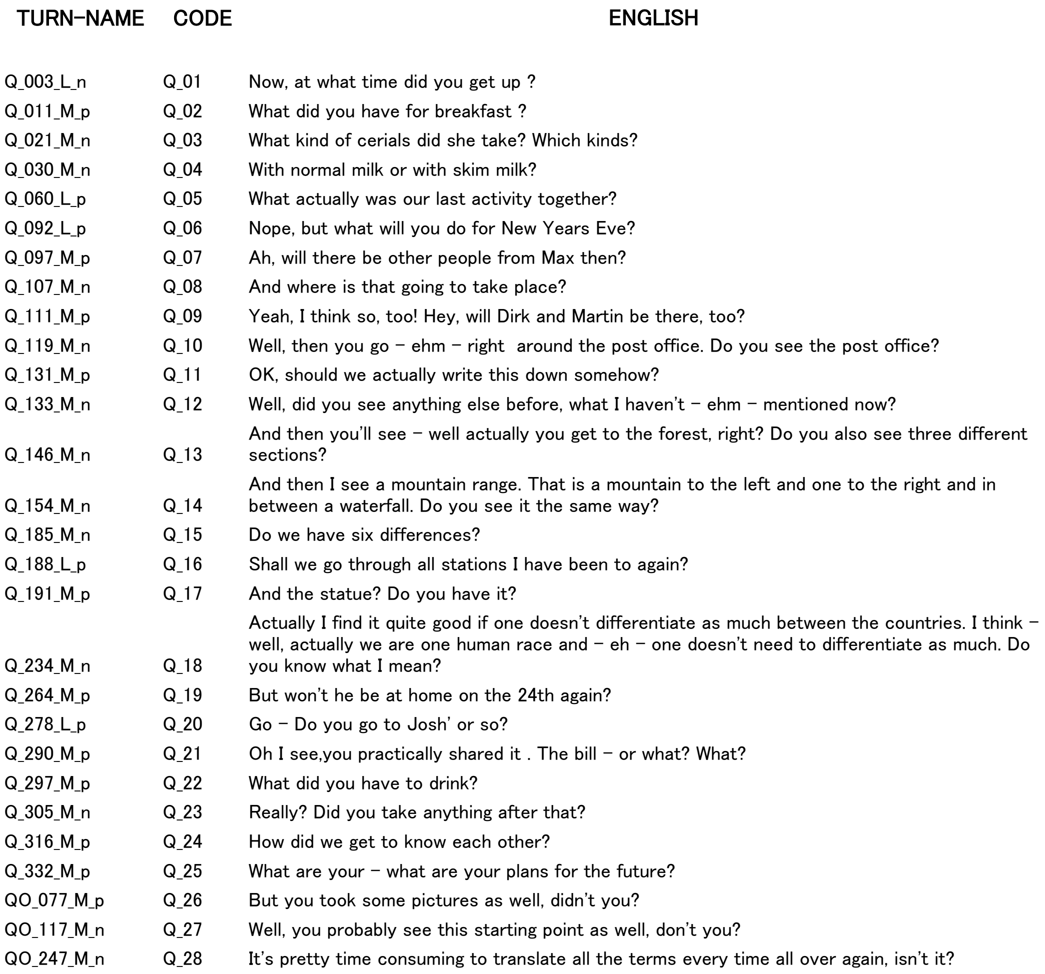
Table B.1: German experimental items labeled as Questions.

TURN-NAME	CODE	GERMAN
SOV_01_L_n	SOV_01	Ähm – Wie war deine erste Stunde nach dem Aufstehen? Du hast ja heute dein Referat gehabt.
SOV_054_L_n	SOV_02	Aber das ist sau geil das Ding. Ich probier das auf jeden Fall aus, wenn wir nochmal bei Chris sind.
SOV_057_L_n	SOV_03	Meine Mutter hat gesagt sie machen das in der Gymnastik und sie hat immer voll schiss, dass sie voll mit dem Gesicht auf den Boden knallt.
SOV_088_L_n	SOV_04	Okay. Also brauch er's vielleicht gar nicht mehr zurück kriegen, weil er's schon hat.
SOV_089_M_p	SOV_05	Er hat mir dann auch wirklich Leid getan. Er hat total gekotzt.
SOV_105_M_n	SOV_06	Nee, ich glaub eher dass de Mark – dass Mark uns einladen würde.
SOV_132_L_p	SOV_07	Nee, ich denke wir unterhalten uns ja und dann haben die gehört, dass wir den Unterschied gefunden haben.
SOV_148_M_p	SOV_08	Und dort könnte man wenn man wollte jetzt ne Pause einlegen. Also steht bei mir jetzt halt da.
SOV_152_M_n	SOV_09	Also ja, jetzt nach links, genau. Da ist ja links die Baumgruppe und die gehst du praktisch rum.
SOV_160_L_n	SOV_10	Ah, okay. Dann hab ich den Schatz gefunden.
SOV_163_M_n	SOV_11	Da gehst du drüber und dann zwischen Wasserfall und drittem Berg nochmal runter.
SOV_203_L_p	SOV_12	Gut. Dann lass mal diskutieren. Das wird das schwerste werden.
SOV_209_L_n	SOV_13	Irgendwie. Hier steht ja, dass Frankreich einen festgesetzen Anteil von französischer Musik im Radio hat.
SOV_213_L_n	SOV_14	Es hat sich ja eigentlich auch erst so wieder entwickelt – so seit den 2000er Jahren – als dass es jetzt wieder deutsche Musik gibt.
SOV_224_M_p	SOV_15	Ja, nee, stimmt. Hast du mich gut überzeugt.
SOV_249_M_p	SOV_16	Ja ja, es wird alles englisch publiziert eigentlich fast alles so.
SOV_259_L_n	SOV_17	Und dann am ersten Weihnachtsfeiertag ist bei uns immer so – jetzt nicht so ganz fest was geplant und am zweiten sind wir immer bei meiner Großmama Gans essen.
SOV_283_M_n	SOV_18	Aber wir waren da mal an Ostern und da hats ziemlich gut geschmeckt, deswegen gehen wir da jetzt noch einmal hin.
SOV_288_M_p	SOV_19	Da hätte da er doch mehr Trinkgeld geben können einfach.
SOV_296_L_n	SOV_20	Aber ich war echt – eh – voll überrascht, dass es trotzdem noch so teuer war. Das hätte ich nie gedacht.
SOV_303_M_n	SOV_21	Das würde mir direkt den Rachen weg fetzen.
SOV_308_L_p	SOV_22	Nee, bis jetzt am Montag glaub ich. Also es war so en Paket das ich halt bekommen hab als ich beim Arzt war und dann ging das so zweieinhalb wochen oder so durch.
SOV_312_M_n	SOV_23	Also ich würd das auch nicht hinkriegen jetzt – über Weihnachten oder so keinen Glühwein zu trinken.
SOV_319_L_p	SOV_24	Das musst du jetzt mal erzählen.
SOV_322_L_n	SOV_25	Du kannst ja mehrere Zukunftspläne erzählen. Ich weiß zwar schon, dass du jetzt mal deine Masterarbeit schreibst, aber du kannst mir das ja erzählen.
SOV_324_M_n	SOV_26	Ich hab auch noch eine mündliche Prüfung, die so total ätzend ist.
SOV_330_M_n	SOV_27	Oh Mann, das ist so peinlich, dass das sich nachher Leute anhören
SOV_340_L_p	SOV_28	Nee, ich bin eigentlich echt froh. Ich hab nicht mehr so viel Lust hier zu sein.
SOV_347_L_n	SOV_29	Wie soll ich das bei dir machen? Du brauchst keinen Arschtritt, weil du machst eh immer alles was du dir vornimmst.
SVO_81_L_n	SOV_30	Ja, das wo du den Pokal in der Hand hast. Das ist eigentlich witzig. Ich hab nur nicht verstanden, warum du den Pokal in der Hand hast.
SVO_112_L_n	SOV_31	Ja, eigentlich alle so. Deshalb war ich auch echt schon verwundert, aber ich hatte auch nicht so viel Zeit. Ich muss dir nachher noch was anderes erzählen. Das passt jetzt hier nicht so gut rein.
SVO_122_M_p	SOV_32	Oder du gehst dann in Richtung Hotel nach Links, aber dann unten am Hotel vorbei.
SOV_16_L_n	SOV_33	Ja, dürfen wir jetzt eigentlich voll abschweifen? Weil beim Wichteln hab ich ein Geschenk von Hannah bekommen.
SOV_348_M_end	SOV_34	Apropos, gestern hab ich mein Zimmer gereinigt, ne? Ich hab mein Zimmer dermaßen ordentlich gereinigt, wie ich es noch nie gemacht hab.

Table B.2: German experimental items labeled as SOV.

TURN-NAME	CODE	GERMAN
SVO_41_L_p	SVO_01	Ja, wahrscheinlich. Aber Mandelmilch finde ich jetzt leckerer als Sojamilch.
SVO_47_L_n	SVO_02	Aber das war jetzt nicht so lustig für mich.
SVO_78_L_n	SVO_03	Aber nur von den Spielern eigentlich. Also, von uns gibt es keine Bilder.
SVO_80_L_n	SVO_04	Achso, doch. Von dir gibt' s ein Bild.
SVO_83_L_n	SVO_05	Ja, du warst auch Teil des Teams und Gewinner.
SVO_115_L_p	SVO_06	Genau, wollte ich auch gerade sagen. Das macht bestimmt Spaß.
SVO_130_L_p	SVO_07	Also bei mir ist das Krankenhaus und dann ist da in der Mitte die Apotheke rechts vom Krankenhaus. Da ist keine Kirche.
SVO_135_M_p	SVO_08	Achso die Bank, ja genau. Aber die sehe ich auch.
SVO_140_L_p	SVO_09	Nein, bei mir ist da eine Statue und ein Bauernhof.
SVO_144_L_p	SVO_10	Genau. Und in der Mitte ist ne Statue bei mir.
SVO_165_M_n	SVO_11	Das nächste was ich jetzt sehe ist der Bauernhof den siehst du ja auch links und rechts der Bahnhof.
SVO_172_M_p	SVO_12	Und dann ist das nächste was ich sehe ein Park. Also so nen Baum und so nen Springbrunnen.
SVO_197_L_p	SVO_13	Dann war das – ähm – ein Trick.
SVO_218_L_p	SVO_14	Vielleicht hört da auch keiner mehr französisches Radio, sondern alle hören online – was weiß ich – Radio USA oder so.
SVO_221_M_p	SVO_15	Ich eigentlich auch noch so im Auto oder so. Aber ansonsten hör ich immer meine Musik vom Handy.
SVO_223_L_p	SVO_16	Ich glaube das Radio – und auch gerade für lokale Nachrichten – richtig gut ist. Und deshalb wahrscheinlich – also ich glaube nicht an den Tod des Radios.
SVO_225_L_p	SVO_17	Also ich denke dass es wichtig – also dass es andere Möglichkeiten gibt, wenn man jetzt seine Sprache schützen will als diese Radiozensur, so zu sagen.
SVO_236_L_p	SVO_18	Ja, ist halt schwierig weil man die anderen nicht versteht, wenn man sie nicht gelernt hat. Aber das ist ja eigentlich auch nicht so das Problem.
SVO_265_L_n	SVO_19	Am vierundzwanzigsten ist er zu Hause. Dann wäre er am ersten und am zweiten gibt es halt bei seiner Schweter was. Gehen wir zu seiner Schwester.
SVO_269_L_p	SVO_20	Ja, also das fände ich eigenltich mal ganz schön. Aber jetzt geht er am zweiten wahrscheinlich auf eine LAN–Party bei Max.
SVO_284_L_n	SVO_21	Ja, also bei uns war das eigentlich auch ganz gut. Also ich hatte halt so ein Fleisch, was ich nicht so gut fand. Aber das war ja nicht denen ihre Schuld. Also so war schon okay. Ist bestimmt ganz nett.
SVO_315_L_p	SVO_22	Ja, so schlimm wird's auch nicht sein. Meine Mama hat ja gesagt in Maßen.
SVO_320_M_p	SVO_23	Eh, nee. Kann ich leider nicht erzählen. Habe keine.
SVO_346_M_p	SVO_24	Dann nicke ich einfach immer und unterstütze dich.
SVO_129_L_p	SVO_25	Ah, okay. Bei mir ist hier die Apotheke. Das wäre dann wohl ein Unterschied.
ADV_08_L_p	SVO_26	Und ich habe aber trotzdem noch gefrühstückt. Sehr lecker.
ADV_20_L_n	SVO_27	Aber ich fands voll cool, also war es dann echt super.
ADV_294_M_n	SVO_28	Okay, aber dann bringt das jetzt auch nicht wirklich viel.
ADV_309_L_n	SVO_29	Jetzt bin ich mal gespannt. Aber gestern hatte ich auch schon keine genommen und da war eigentlich alles okay.
ADV_344_L_n	SVO_30	Nee, das hilft auch nicht so viel. Da werde ich aggressiv.
ADV_84_M_p	SVO_31	Ich würde jetzt ja sagen ″Ich hab lautstark unterstützt″ hab ich aber ja nicht.
SVO_143_M_p	SVO_32	Genau. Das ist dann wahrscheinlich bei mir der Fels.
ADV_24_L_p	SVO_33	Nee, Mandelmilch. Sau geil. Die ist aber schon leer.
ADV_L_n	SVO_34	Ja, also ich finds sehr fein. Aber nur – Also kalt ist besser als warm.

Table B.3: German experimental items labeled as SVO.



TURN-NAME	CODE	ENGLISH
Q_003_L_n	Q_01	Now, at what time did you get up ?
Q_011_M_p	Q_02	What did you have for breakfast ?
Q_021_M_n	Q_03	What kind of cerials did she take? Which kinds?
Q_030_M_n	Q_04	With normal milk or with skim milk?
Q_060_L_p	Q_05	What actually was our last activity together?
Q_092_L_p	Q_06	Nope, but what will you do for New Years Eve?
Q_097_M_p	Q_07	Ah, will there be other people from Max then?
Q_107_M_n	Q_08	And where is that going to take place?
Q_111_M_p	Q_09	Yeah, I think so, too! Hey, will Dirk and Martin be there, too?
Q_119_M_n	Q_10	Well, then you go – ehm – right around the post office. Do you see the post office?
Q_131_M_p	Q_11	OK, should we actually write this down somehow?
Q_133_M_n	Q_12	Well, did you see anything else before, what I haven't – ehm – mentioned now?
Q_146_M_n	Q_13	And then you'll see – well actually you get to the forest, right? Do you also see three different sections?
Q_154_M_n	Q_14	And then I see a mountain range. That is a mountain to the left and one to the right and in between a waterfall. Do you see it the same way?
Q_185_M_n	Q_15	Do we have six differences?
Q_188_L_p	Q_16	Shall we go through all stations I have been to again?
Q_191_M_p	Q_17	And the statue? Do you have it?
Q_234_M_n	Q_18	Actually I find it quite good if one doesn't differentiate as much between the countries. I think – well, actually we are one human race and – eh – one doesn't need to differentiate as much. Do you know what I mean?
Q_264_M_p	Q_19	But won't he be at home on the 24th again?
Q_278_L_p	Q_20	Go – Do you go to Josh' or so?
Q_290_M_p	Q_21	Oh I see,you practically shared it . The bill – or what? What?
Q_297_M_p	Q_22	What did you have to drink?
Q_305_M_n	Q_23	Really? Did you take anything after that?
Q_316_M_p	Q_24	How did we get to know each other?
Q_332_M_p	Q_25	What are your – what are your plans for the future?
QO_077_M_p	Q_26	But you took some pictures as well, didn't you?
QO_117_M_n	Q_27	Well, you probably see this starting point as well, don't you?
QO_247_M_n	Q_28	It's pretty time consuming to translate all the terms every time all over again, isn't it?

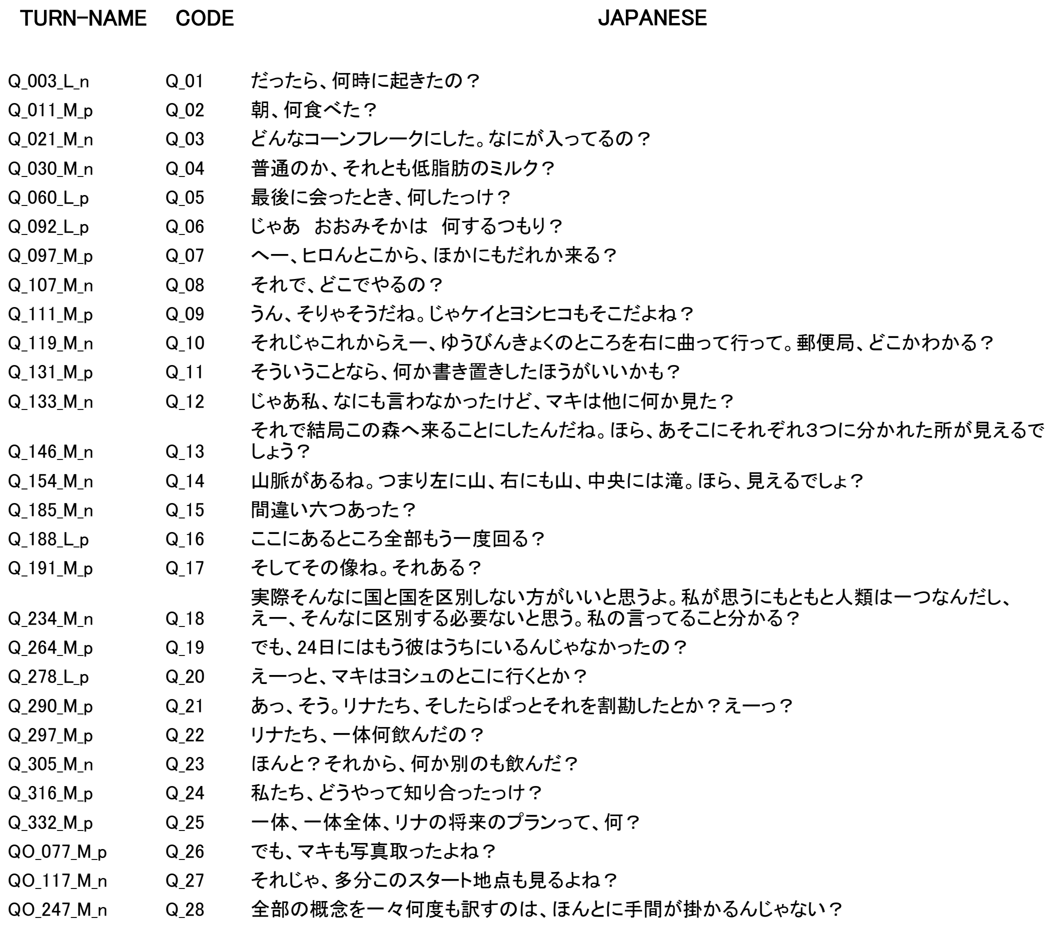
Table B.4: English translations labeled as Questions.

TURN-NAME	CODE	ENGLISH
SOV_01_L_n	SOV_01	Ehm – how was the first hour after you got up? You gave your talk today didn't you?
SOV_054_L_n	SOV_02	But that is fucking good that thing. I'll try it out for sure, when we're at Chris' again.
SOV_057_L_n	SOV_03	My mom told me that they do that in gymnastics and she's always scared stiff that she'd bang her head on the floor, face down.
SOV_088_L_n	SOV_04	OK! So perhaps he doesn't need to get it back as he's already got it.
SOV_089_M_p	SOV_05	I was really sorry for him then. He threw up like hell.
SOV_105_M_n	SOV_06	Nope, I rather think that Mark – that Mark would invite us.
SOV_132_L_p	SOV_07	Nope, I think we talk and then they've heard that we found the difference.
SOV_148_M_p	SOV_08	And there one could take a break, if one wanted to. Well that's written on mine anyway.
SOV_152_M_n	SOV_09	OK then, now to the left, exactly. There is the group of trees on the left and you practically walk around it.
SOV_160_L_n	SOV_10	Ah, ok. Then I have found the treasure.
SOV_163_M_n	SOV_11	There you cross and then between the waterfall and the third mountain down again.
SOV_203_L_p	SOV_12	OK. Then let's discuss it. That'll be the most difficult part.
SOV_209_L_n	SOV_13	Somehow. Here it is written that France has specified percentage of French music on the radio.
SOV_213_L_n	SOV_14	It has actually only developed since, well, the year 2000 – that there is German music again.
SOV_224_M_p	SOV_15	Yes, no, right. You've convinced me.
SOV_249_M_p	SOV_16	Yes, yes everything is published in English, actually almost everything.
SOV_259_L_n	SOV_17	And the on Christmas day it is always like that at our place – not everything planned and on Boxing Day we always go to grandma's place and eat goose.
SOV_283_M_n	SOV_18	But we were there once at Easter and food was pretty good, that's why we're going there again.
SOV_288_M_p	SOV_19	He could have simply given a higher tip.
SOV_296_L_n	SOV_20	But I really was – eh – surprised that it was still that expensive. I would never have thought so.
SOV_303_M_n	SOV_21	That would directly burn my throat.
SOV_308_L_p	SOV_22	No, until Monday I think. Well, it was a package I got when I was at the doctor's and then it went on for two weeks or so.
SOV_312_M_n	SOV_23	Well,I wouldn't manage either now – no hot wine over Christmas or so.
SOV_319_L_p	SOV_24	You'll have to tell that now.
SOV_322_L_n	SOV_25	You can talk about more than one plan for your future. I already know that you're writing your Master thesis, but you can tell me about it.
SOV_324_M_n	SOV_26	I also have an oral exam that sucks.
SOV_330_M_n	SOV_27	Oh man, it's so embarrassing that people will listen to this later on.
SOV_340_L_p	SOV_28	No, I'm actually quite glad. I'm not really keen on staying here.
SOV_347_L_n	SOV_29	What shall I do in your case? You don't need a kick in the ass, as you do everything anyway.
SVO_81_L_n	SOV_30	Yes, the one where you hold the cup. That is actually quite funny. I just didn't understand why you're holding the cup in your hand.
SVO_112_L_n	SOV_31	Yes, actually everyone. That's why I was really baffled, but I didn't have that much time either. I have to tell you something else later on. It doesn't fit here right now.
SVO_122_M_p	SOV_32	Or you go to the left in the direction of the hotel, but then at the bottom you pass the hotel.
SOV_16_L_n	SOV_33	Yes are we allowed to digress that much? Because I got a present from Hannah when we did a secret Santa.
SOV_348_M_end	SOV_34	By the way, yesterday I cleaned my room, right? I cleaned my room more scrupulously than ever before.

Table B.5: English translations labeled as SOV.

TURN-NAME	CODE	ENGLISH
SVO_41_L_p	SVO_01	Yes probably. But I find almond milk better than soy milk
SVO_47_L_n	SVO_02	But that wasn't that funny for me.
SVO_78_L_n	SVO_03	But actually only of the players. Well, there are no photos of us.
SVO_80_L_n	SVO_04	Oh well yes. There is a photo of you.
SVO_83_L_n	SVO_05	Yes, you were also part of the team and the winner.
SVO_115_L_p	SVO_06	Exactly, I just wanted to say that. This will certainly be fun.
SVO_130_L_p	SVO_07	Well, on mine there is the hospital and then there is a pharmacy in the middle, to the right of the hospital. There is no church.
SVO_135_M_p	SVO_08	Oh well the bank, yeah exactly. But I see that, too
SVO_140_L_p	SVO_09	No, on mine there is a statue and a farm.
SVO_144_L_p	SVO_10	Exactly. And in between there is a statue on mine.
SVO_165_M_n	SVO_11	The next thing I see is the farm, you can see that, too on the left and the station to the right.
SVO_172_M_p	SVO_12	And then the next thing I see is a park. That is a tree and a fountain.
SVO_197_L_p	SVO_13	Then that was – ehm – a trick.
SVO_218_L_p	SVO_14	Perhaps no one listens to the French radio anymore, but everybody listens online to – what do I know – Radio USA or the like.
SVO_221_M_p	SVO_15	Well, me too, in the car or so. But normally I always listen to music from my mobile.
SVO_223_L_p	SVO_16	I think that the radio is really good – in particular for local news. And that is why – well I don't believe in the death of the radio.
SVO_225_L_p	SVO_17	Well I think that it is important – well that there are other possibilities to protect your own language than via this radio censure, so to speak.
SVO_236_L_p	SVO_18	Yes, that is difficult,as you don't understand the others if you didn't learn them. But that actually isn't the problem, right?
SVO_265_L_n	SVO_19	He'll be at home on the 24th. Then he'd be on the first – and on the second there is something at his sister's... Are we going to his sister's.
SVO_269_L_p	SVO_20	Yes, I'd actually find that quite nice. But now he's probably going to a LAN-party at Max' on the second.
SVO_284_L_n	SVO_21	Well actually it was quite allright for us, too. Well, I had some meat I didn't like that much. But that wasn't their fault. Well, it was quite ok. It will certainly be nice.
SVO_315_L_p	SVO_22	Yes, it won't be that bad. My mom said in moderation.
SVO_320_M_p	SVO_23	Eh, no, I can't tell you that, sorry. I haven't got one.
SVO_346_M_p	SVO_24	Then I'll just nod all the time and support you.
SVO_129_L_p	SVO_25	Ah, ok, on mine there is the pharmacy here. That would probably be a difference.
ADV_08_L_p	SVO_26	And I still had breakfast. Yummy.
ADV_20_L_n	SVO_27	But I found it really cool, well it was superb then.
ADV_294_M_n	SVO_28	OK, but then that doesn't really help much.
ADV_309_L_n	SVO_29	Now I'm really curious. But yesterday I didn't take any either and actually everything was ok.
ADV_344_L_n	SVO_30	No, that won't help that much. I'll get aggressive.
ADV_84_M_p	SVO_31	I'd say now "I supported loudly" but I didn't do it.
SVO_143_M_p	SVO_32	Exactly. That is probably the rock on mine.
ADV_24_L_p	SVO_33	No, almond milk. Awesome. But it's already gone.
ADV_L_n	SVO_34	Yes, well, I find it very nice. But only – well it's better cold than warm.

Table B.6: English translations labeled as SVO.



TURN-NAME	CODE	JAPANESE
Q_003_L_n	Q_01	だったら、何時に起きたの？
Q_011_M_p	Q_02	朝、何食べた？
Q_021_M_n	Q_03	どんなコーンフレークにした。なにが入ってるの？
Q_030_M_n	Q_04	普通のか、それとも低脂肪のミルク？
Q_060_L_p	Q_05	最後に会ったとき、何したっけ？
Q_092_L_p	Q_06	じゃあ　おおみそかは　何するつもり？
Q_097_M_p	Q_07	へー、ヒロんとこから、ほかにもだれか来る？
Q_107_M_n	Q_08	それで、どこでやるの？
Q_111_M_p	Q_09	うん、そりゃそうだね。じゃケイとヨシヒコもそこだよね？
Q_119_M_n	Q_10	それじゃこれからえー、ゆうびんきょくのところを右に曲って行って。郵便局、どこかわかる？
Q_131_M_p	Q_11	そういうことなら、何か書き置きしたほうがいいかも？
Q_133_M_n	Q_12	じゃあ私、なにも言わなかったけど、マキは他に何か見た？
Q_146_M_n	Q_13	それで結局この森へ来ることにしたんだね。ほら、あそこにそれぞれ3つに分かれた所が見えるでしょう？
Q_154_M_n	Q_14	山脈があるね。つまり左に山、右にも山、中央には滝。ほら、見えるでしょ？
Q_185_M_n	Q_15	間違い六つあった？
Q_188_L_p	Q_16	ここにあるところ全部もう一度回る？
Q_191_M_p	Q_17	そしてその像ね。それある？
Q_234_M_n	Q_18	実際そんなに国と国を区別しない方がいいと思うよ。私が思うにもともと人類は一つなんだし、えー、そんなに区別する必要ないと思う。私の言ってること分かる？
Q_264_M_p	Q_19	でも、24日にはもう彼はうちにいるんじゃなかったの？
Q_278_L_p	Q_20	えーっと、マキはヨシュのとこに行くとか？
Q_290_M_p	Q_21	あっ、そう。リナたち、そしたらぱっとそれを割勘したとか？えーっ？
Q_297_M_p	Q_22	リナたち、一体何飲んだの？
Q_305_M_n	Q_23	ほんと？それから、何か別のも飲んだ？
Q_316_M_p	Q_24	私たち、どうやって知り合ったっけ？
Q_332_M_p	Q_25	一体、一体全体、リナの将来のプランって、何？
QO_077_M_p	Q_26	でも、マキも写真取ったよね？
QO_117_M_n	Q_27	それじゃ、多分このスタート地点も見るよね？
QO_247_M_n	Q_28	全部の概念を一々何度も訳すのは、ほんとに手間が掛かるんじゃない？

Table B.7: Japanese translations labeled as Questions.

TURN-NAME	CODE	JAPANESE
SOV_01_L_n	SOV_01	えーっと、今日起きてすぐ気分はどうだった？ほら、今日プレゼンがあるから、、、。
SOV_054_L_n	SOV_02	それにしてもこれすごいね！またケイの家に来た時、絶対やってみるよ。
SOV_057_L_n	SOV_03	うちのおかあさんは体操するときそれを練習してるらしいけど、いつ鼻ぺちゃになっちゃうか、ひやひやしてるって、、。
SOV_088_L_n	SOV_04	オッケー、じゃあもうあるんだったら返してもらわなくてもいいね。
SOV_089_M_p	SOV_05	ほんと可哀想だった。吐きまくっていたからね。
SOV_105_M_n	SOV_06	むしろヒロが、ヒロが私たちをよんでくれるんじゃない。
SOV_132_L_p	SOV_07	いや、私たちが話をしてて、だからあの人らは違いが見つかったって知ったんでしょ。
SOV_148_M_p	SOV_08	そこで家休憩をとれるよ。私のとこにはそう書いてあるね。、
SOV_152_M_n	SOV_09	そうそにこを左に！うん。そこの木がいっぱい生えているところをぐるっと回るの。
SOV_160_L_n	SOV_10	そうわかった。じゃその宝物は見つけたね。
SOV_163_M_n	SOV_11	そこを登って滝と3番目の山の間を下りていくんだよ。
SOV_203_L_p	SOV_12	よしじゃあデイスカッションしよ。これがいちばん難しいかも。
SOV_209_L_n	SOV_13	なんとなく、、、フランスはラジオでのフランス音楽の割合を定めている。
SOV_213_L_n	SOV_14	２０００年来ずっとそうしてきたように、また発展してきたからこそ今のドイツ音楽があるのだ。
SOV_224_M_p	SOV_15	うーん、そうか、よくわかった。
SOV_249_M_p	SOV_16	そうだね、研究結果はすべて英語で発表されているもんね、だいたい全部。
SOV_259_L_n	SOV_17	それからうちらの家じゃ、クリスマスの一日目は毎年これと言って決まっていないんだ。うーん、二日目はいつもおばあちゃんのうちでがガチョウのこんがり焼いたのを食べるんだ。
SOV_283_M_n	SOV_18	でも復活祭の時一回あそこに行ってけっこうおいしかったよ。だから今度もう一回行こうよ。
SOV_288_M_p	SOV_19	その時、彼ただもっとチップをあげれたのに。
SOV_296_L_n	SOV_20	でもそれそれでも超高かったから、えーほんとびっくりしちゃった。思いもよらなかった。
SOV_303_M_n	SOV_21	それだったらのどがもろにやけるね。
SOV_308_L_p	SOV_22	いや、次の月曜までだと思うよ。そうねそれはこんな箱で私が医者のところに行った時、ちょうどもらったものでそれから、二週間半ぐらいかなそれでずーっときたよ。
SOV_312_M_n	SOV_23	まあクリスマスにとかグリューワインを飲まないで過ごすなんて、今は私もできっこないかな。
SOV_319_L_p	SOV_24	それはリナに今ちょっと話してもらわなきゃ。
SOV_322_L_n	SOV_25	マキも、いつくもの将来のプランを話せると思うよ。でも、私ももう今マキが修士論文を書いてるって知ってるし、まずは一応それから話していいんじゃない。
SOV_324_M_n	SOV_26	それに口頭試験もあるんだけど、それがまたほんとやらしいのよ。
SOV_330_M_n	SOV_27	ああもう嫌。それを人に後から聞かれるなんて恥ずかしい。
SOV_340_L_p	SOV_28	いやその方がほんとよかった。だって私もうそんなにここにいる気ないし。
SOV_347_L_n	SOV_29	私はマキにどうすべきだって言うのよ。マキはやろうと思ったこと、むしろ全部いつも自分でやるじゃない。だからあと押しなんていらないじゃん。
SVO_81_L_n	SOV_30	ええ、リナがそのトロフィーを手に持っているとこ。ほんとにそれ笑えるよね。ただ、何でリナがそのトロフィーを手に持っているのか分からないけど。
SVO_112_L_n	SOV_31	そう元々みんなそうだもん。だから、私もほんと驚いた。でも、その時は時間もあまりなくて。後で、違う話をちょっとしてあげるね。でも、今ここではあんまり都合がよくないわ。
SVO_122_M_p	SOV_32	それかここをホテルの方向に左に行く。でも、ホテルの下を通り過ぎたところね。
SOV_16_L_n	SOV_33	ねえ、私たち、これでバッくれちゃってもいい？だって、プレゼント交換会で、私、ナナコのプレゼントもらったもん。
SOV_348_M_end	SOV_34	ところで昨日、私自分の部屋をきれいにしたの。それもね今までにないくらい、きちんときれいにしたんだから。

Table B.8: Japanese translations labeled as SOV.

TURN-NAME	CODE	JAPANESE
SVO_41_L_p	SVO_01	うんたぶん。でもアーモンドミルクの方が豆乳より美味しいかな。
SVO_47_L_n	SVO_02	笑えないんですけど。
SVO_78_L_n	SVO_03	だけどあるのはゲーム出場者の写真だけで私たちのはないの。
SVO_80_L_n	SVO_04	ほら、やっぱり。リナの写真あるじゃん。
SVO_83_L_n	SVO_05	そう。リナもチームの一員で勝者だったしね。
SVO_115_L_p	SVO_06	それが言いたかったんだよ。絶対面白いよ。
SVO_130_L_p	SVO_07	病院があってその右側に薬局はあるけど、教会はないよ。
SVO_135_M_p	SVO_08	あーそういえば、銀行なら私にも見えてるよ。
SVO_140_L_p	SVO_09	ううん、ここから見えるのは銅像と農場だよ。
SVO_144_L_p	SVO_10	そうだね。で、まんなかには銅像がある。
SVO_165_M_n	SVO_11	その次に見えるのがマキの左側にある農場で右側には駅があるでしょ。
SVO_172_M_p	SVO_12	それから公園も見える。木と噴水も見えるね。
SVO_197_L_p	SVO_13	それならうーん、だまされたかな。
SVO_218_L_p	SVO_14	そこじゃおそらくフランスのラジオ放送じゃなくてネットで聞いている、ほら、USAラジオとか、、、。
SVO_221_M_p	SVO_15	わたしも結局、車で聞くよ、それか携帯で音楽を聴いたりとか。
SVO_223_L_p	SVO_16	ラジオは、質のいいローカルニュースもあるしいいと思う。ラジオ放送がなくなるとは思わない。
SVO_225_L_p	SVO_17	ようするにわたしが言いたいのは言語を保護するためにラジオ放送の規制だけでなく他にもいろいろな方法があるということなの。。
SVO_236_L_p	SVO_18	そうね、それを習っていなかったら他人をよく理解していないんだし、難しいよね。でもそのことは本来そんな一番の問題ってわけではないよ。
SVO_265_L_n	SVO_19	２４日には彼はうちにいるよ。そして、一日目と 二日目に彼のお姉さんの家になにか、、、一緒にお姉さんのところに行く。
SVO_269_L_p	SVO_20	うん、それは私もいいと思うけど。でもこの二日には彼たぶんヒロシのとこのLANパーティーに行くと思うよ。
SVO_284_L_n	SVO_21	ええ、私たちの時はまあ悪くなかったわよ。ただその時は私にはあまりおいしくない肉が出てきちゃって。でもそれはあの人たちが悪かった訳じゃないから。まあ、あれはあれでオッケーだったわよ。きっととってもいいと思うよ。
SVO_315_L_p	SVO_22	ええ、そんなにひどくなることはないでしょう。ママが程ほどにって言ってたし。
SVO_320_M_p	SVO_23	んー、いや残念だけど言えない。持ってないから。
SVO_346_M_p	SVO_24	それじゃ、ただずーっとうなづきながらマキのことを応援しとくね。
SVO_129_L_p	SVO_25	ああ、分かった。私のとこはここに薬局があって、それがつまり違ってるとこね。
ADV_08_L_p	SVO_26	そして、私はそれでも朝ごはんを食べた。とってもおいしかったよ。
ADV_20_L_n	SVO_27	でも、私はほんとかっこよかったと思うわ。うん、ほんとよかった。
ADV_294_M_n	SVO_28	分かった。それじゃ、それは今からじゃあんまり役立たないわね。
ADV_309_L_n	SVO_29	これからどうなるか。でも、昨日はもう何も飲まなかったし、もうほとんどオーケーだったよ。
ADV_344_L_n	SVO_30	いや、これもあんまり役立たないね。もうぶん殴りたくなるわ。
ADV_84_M_p	SVO_31	それじゃ、言うけどさ、「声を張り上げて、応援した。」ってわけでもないよ。
SVO_143_M_p	SVO_32	全くその通り！これが恐らく私の岩ね。
ADV_24_L_p	SVO_33	ううん、アーモンドミルク。ほんとマジうまい。でももう飲んじゃった。
ADV_L_n	SVO_34	ええそうね、私はとっても舌触りがいいと思う。ただ私は温かいのより冷たい方がいいと思う。

Table B.9: Japanese translations labeled as SVO.

CODE	TURN
F_01	Und dann hat er sie die Treppe runter geschubst, ne?
F_02	Achso. Die kenne ich auch gar nicht.
F_03	Ja obwohl, kann gut sein. Der hat nämlich die Rede gehalten, praktisch.
F_04	Achso, achso, ja, das kann sein.
F_05	Er hat sie doch die Treppe runtergeworfen.
F_06	Ah, siehst du das war bei mir. Was war denn bei dir?
F_07	Ja genau, das im Kuhstall. Wie er so scheiße gesungen hat.
F_08	Nee, das hab ich nicht gesehen.
F_09	Oh nee. Aha! Das ist der erste Unterschied.
F_10	Nee, die Szene nicht. Da wird nur von gesprochen.
F_11	Oh, das kommt mir aber bekannt vor.
F_12	Oh. Nee, die kenn ich nicht. Das hab ich nicht.
F_13	Wusstest du das denn eigentlich mit dem Typen und so?
F_14	Ja, was war sonst noch so bei dir?
F_15	Achso. Nee, so nicht.
F_16	Ja ja genau. Es geht wohl auch eigentlich um ihre Tochter, die nicht mehr bei ihr wohnen will. Oder irgendwie sowas.
F_17	Ich habe leider ihren Text nicht verstanden.
F_18	Achso. Das hab ich nicht gesehen.
F_19	Aber ich sag mal – Ich denk, das ist ja Phonetik und nicht Psychologie.
F_20	Nee, das hab ich auch gesehen. Im Kuhstall da, ne?
F_21	Achso, das hattest du gar nicht?
F_22	Achso, nee, mir ist das – bei mir ist das noch vorher mit der Musik.
F_23	Nee, ich erinner mich jetzt irgendwie auch nicht mehr dran.
F_24	Ja, die Folge kannte ich vorher schon.
F_25	Das habe ich ganz vergessen.
F_26	Die habe ich nicht gesehen.
F_27	Achso, die Szene hab ich ja auch gesehen. Ja.
F_28	Da hat wohl jemand vorher schon gewohnt und der will ausziehen.
F_29	Nee, das hab ich nicht.
F_30	Nee, also das habe ich nicht wahrgenommen.
F_31	Ja und sie hat es ihm freiwillig – also sie hat es ihm nur aus ihrem schlechten Gewissen heraus gesagt.
F_32	Ich hab garnicht mitgekriegt, dass sie sich danach haben scheiden lassen. Der Typ ist niewieder in einer Folge aufgetaucht, oder?
F_33	Ist sie jetzt eigentlich unfruchtbar? Kann sie nie wieder Kinder kriegen, oder?
F_34	Wie kam das denn überhaupt noch raus, dass das Kind von ihm ist?
F_35	Nee, er meinte doch auch, dass sie sich auch ganz einfach aus dem Staub machen würde, oder so.
F_36	Ich weiß garnicht. Er hat auch noch eine Tochter, ne? Beathe ist seine Tochter, oder?
F_37	Am schlimmsten, ja. Wie heißt das Mädchen noch mit den langen blonden Haaren? Lisa.
F_38	Aber dieser Konflikt irgendwie mit – wie heißt sie denn noch?
F_39	Ja, im Schweinestall die Szene habe ich dann wieder gesehen.
F_40	Ah, nö. Hab ich auch gar nichts. Also ich hab wohl nur diese beiden Szenen zu dem Thema gesehen.
F_41	Ja, wenn die das jetzt schon so häufig gespielt haben, oder so, kann ja auch die Tonqualität leiden.
F_42	Ich glaub nicht, wir müssen jetzt nur rausfinden, wo da jetzt die Unterschiede sind.
F_43	Er – Sie sagt ihm wohl anscheinend, dass das Kind nicht von ihm ist.
F_44	Aber du hast das dann wieder wo nachher dann ihr Vater kommt und ihn total betrunken im Stall findet.
F_45	Nee, also ich hab glaub ich jetzt alles durch.
F_46	Und sie ist natürlich total eifersüchtig.
F_47	Genau die hab ich auch. Ja, die hatte ich auch, die Szene.
F_48	Also hast du keine weiterfolgende Szene davon.
TR_01	Deshalb war das zwar nicht ganz so stressig, aber auch nicht die schönste Stunde nach dem Aufstehen, die man sich vorstellen kann.
TR_02	Also ich wollte eigentlich mit Mark einfach reden.
TR_03	Dann wär es vielleicht wirklich so wie du gesagt hast, dass es dann halt so zwanghaft wird und dann sagt man halt ″Ach Gott, jetzt kann ich's nicht mehr hören.″
TR_04	Deutsch stirbt jetzt nicht aus nur weil wir ein Paar englische Wörter benutzen oder Strukturen übernehmen.
TR_05	Weil man's ja eh auf Englisch übersetzt als Deutscher, dann kann man's auch in seiner Sprache formulieren.
TR_06	Ich hab ich auch gewundert, weil ich dachte eigentlich, dass Mark euch dann einläd. Aber ihr wart bis jetzt noch nicht eingeladen.
TR_07	Aber irgendwie gibt' s da ja immer diese zwei Ligen irgendwie. Meine Oma, die immer sagt ″Können die nicht Deutsch reden? Versteh gar nix und so.″ Und die Jugend, die halt selber ständig irgendwie englische Wörter benutzt.
TR_08	Weil auch jemand aus Indien leicht hier herkommen kann und irgendwas studieren kann ohne dass er gar nix versteht, sondern er kann sich eigentlich ziemlich gut zurecht finden.
TR_09	Und dann hab ich mir mein Referat durchgelesen. Ungefähr zwei Stunden lang, damit ich überhaupt weiß was ich da sage.
TR_10	Ja und ob das halt die Sprache dann schützt ist dann auch wieder dahingestellt. Weil ich weiß nicht ob jetzt Frankreich nur deshalb seine Sprache besser schützt, weil es im Radio französische Lieder spielt.

Table B.10: Filler items followed by practice items.

Bibliography

Arnfield, S., Roach, P., Setter, J., Greasley, P., and Horton, D. (1995). Emotional stress and speech tempo variation. In *Speech under Stress*.

Auer, P. (1996). On the prosody and syntax of turn-continuations. *Prosody in conversation*, pages 57–100.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Beattie, G. W. (1981). Sequential temporal patterns of speech and gaze in dialogue. *Nonverbal Communication, Interaction, and Gesture*, pages 297–320.

Beckman, M. E. and Hirschberg, J. (1994). The tobi annotation conventions. *Ohio State University*.

Bierwisch, M. (1963). *Grammatik des deutschen Verbs*. Akademie Verlag.

Boersma, P. and Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. version 5.4.09, retrieved 2015/06/02 from http://www.praat.org/.

Braun, A. and Oba, R. (2007). Speaking tempo in emotional speech–a cross-cultural study using dubbed speech. In *Proceedings 16th International Conference on Phonetic Sciences, Saarbrücken, Germany*, pages 77–82. Citeseer.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. ERIC.

Clark, H. H. and Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Couper-Kuhlen, E. and Ono, T. (2007). Practices in english, german and japanese. *Pragmatics*, 17:4–513.

Cutler, A. and Pearson, M. (1986). On the analysis of prosodic turn-taking cues. *Intonation in discourse*, pages 139–156.

De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390.

De Ruiter, J. P., Mitterer, H., and Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, pages 515–535.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

Enfield, N. J. and Levinson, S. C. (2006). Roots of human sociality. *New York: Berg.*

Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184.

Gambi, C., Jachmann, T., Staudte, M., et al. (2015). The role of prosody and gaze in turn-end anticipation.

Grabe, E. (1998). Pitch accent realization in english and german. *Journal of Phonetics*, 26(2):129–143.

Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, 23(02):225–257.

Indefrey, P. and Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1):101–144.

Jescheniak, J. D., Schriefers, H., and Hantsch, A. (2003). Utterance format effects phonological priming in the picture-word task: Implications for models of phonological encoding in speech production. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2):441.

Kamide, Y., Altmann, G. T., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, 49(1):133–156.

Magyari, L., Bastiaansen, M. C., de Ruiter, J. P., and Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of cognitive neuroscience*, 26(11):2530–2539.

Magyari, L. and de Ruiter, J. P. (2012). Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3.

Nariyama, S. (2003). *Ellipsis and reference tracking in Japanese*, volume 66. John Benjamins Publishing.

Nariyama, S. (2004). Subject ellipsis in english. *Journal of pragmatics*, 36(2):237–264.

Ng, A. W. and Chan, A. H. (2012). Finger response times to visual, auditory and tactile modality stimuli. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 2, pages 1449–1454.

Pellegrino, F., Coupé, C., and Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, 87(3):539–558.

Pellegrino, F., Farinas, J., and Rouas, J. (2004). Automatic estimation of speaking rate in multilingual spontaneous speech. In *Speech Prosody 2004, International Conference*.

Pitrelli, J. F., Beckman, M. E., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the tobi framework. In *ICSLP*.

Roach, P. (1998). *Some languages are spoken more quickly than others*. na.

Sacks, H. (1995). Lectures on conversation: volumes i & ii, ed. by g. jefferson; with an introduction by ea schegloff.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735.

Schnur, T. T., Costa, A., and Caramazza, A. (2006). Planning at the phonological level during sentence production. *Journal of Psycholinguistic Research*, 35(2):189–213.

Selkirk, E. (2009). On clause and intonational phrase in japanese: The syntactic grounding of prosodic constituent structure. *Gengo Kenkyu*, 136:35–73.

Selting, M. (1995). *Prosodie im Gespräch: Aspekte einer interaktionalen Phonologie der Konversation*, volume 329. Walter de Gruyter.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* PhD thesis, University of California at Berkeley.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Tanaka, H. (2000). *Turn-taking in Japanese conversation: A study in grammar and interaction*, volume 56. John Benjamins Publishing.

Tanaka, H. (2004). Prosody for marking transition-relevance places in japanese conversation. *Sound patterns in interaction: Cross-linguistic studies from conversation*, 62:63.

Van der Auwera, J. and König, E. (1994). *The Germanic Languages.* Taylor & Francis.

Venditti, J. J., Maekawa, K., and Beckman, M. E. (2008). Prominence marking in the japanese intonation system. *Handbook of Japanese linguistics*, pages 456–512.

Walker, M. B. and Trimboli, C. (1984). The role of nonverbal signals in co-ordinating speaking turns. *Journal of Language and Social Psychology*, 3(4):257–272.

Watanabe, M., Hirose, K., Den, Y., and Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, 50(2):81–94.

Weilhammer, K. and Rabold, S. (2003). Durational aspects in turn taking. *International Congresses of Phonetic Sciences*, pages 823–828.

Wennerstrom, A. and Siegel, A. F. (2003). Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707–1717.

Wilson, M. and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic bulletin & review*, 12(6):957–968.

Yamashita, Y. (2004). *Kansaiben Kogi.* Kodansha.