Computational models of disfluencies : fillers and discourse markers in spoken language understanding

Tanvi Dinkar

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAT001

Thèse de doctorat

Computational Models of Disfluencies:
Fillers and Discourse Markers in Spoken Language Understanding

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Telecom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 26 Janvier 2022, par

TANVI DINKAR

Composition du Jury :

Liesbeth Degand Université Catholique de Louvain, Belgium	Rapportrice
Frédéric Béchet Aix Marseille Université	Rapporteur
Justine Cassell Carnegie Mellon University, USA	Examinatrice
Martine Adda Decker CNRS-LIMSI, France	Président du Jury
Chloé Clavel Telecom Paris, Palaiseau, France	Directrice de thèse
Catherine Pelachaud Sorbonne Université, France	Co-Directrice de thèse
Ioana Vasilescu CNRS-LIMSI, France	Co-Directrice de thèse
Mohamed Chetouani Sorbonne Université, France	Invité

This thesis is dedicated to the girl I used to be. The one that listened to her bullies, internalised the worst thoughts about herself, and narrowly escaped becoming a high school drop out. You do not need to be scared to fail, or doubtful about your ability to learn. This thesis is proof of that.

Acknowledgements

This work would not have been possible without the help of many. I would firstly like to thank my supervisor, Dr. Chloé Clavel and co-supervisors Dr. Ioana Vasilescu and Dr. Catherine Pelachaud for their guidance throughout my PhD. I would also like to thank my jury members Dr. Liesbeth Degand, Dr. Frédéric Béchet, Dr. Justine Cassell, Dr. Martine Adda Decker and Dr. Mohamed Chetouani for their insightful comments and feedback. I would like to thank my collaborators Utku Norman, Dr. Pierre Colombo, Dr. Barbara Bruno, Dr. Matthieu Labeau and Dr. Beatrice Biancardi. I would like to thank my unofficial collaborator, Dr. Alessandra Cervone. Thank you to the people in my lab; Lucien Maman, Dr. Giovanna Varni and Dr. Léo Hemamou. Thank you to Dr. Manuel Bied, Dr. Sera Buyukgoz, Jauwairia Nasir, Gülseren Norman, Maha El Garf, Silvia Tulli and Natalia Calvo, I'm so happy that I got to meet you through this scholarship. In no particular order, thank you Dr. Jean Zagdoun, Lijo Thomas, and Dr. Asma Atamna. Thank you to Dr. Tanya Machado for helping me stay centred during the last exhausting stretch of the PhD. Thank you to my extended family, Vinay Krupakaran, Vikhyat Kaushal and Alessandra Cervone, who witnessed my chaotic journey to achieve the PhD, who always picked up the phone when I needed them. I'm so blessed that I have a family that supported this journey – especially during these isolating pandemic years. Thank you to Tejas Dinkar, Dr. Jananee Muralidharan, and little Tara Dinkar. A big hug and thank you to Rama Chandrashekar and Anita Arun. Lastly, thank you Dr. Dinkar Sitaram and Swarna Dinkar, for being such loving parents.

Abstract

People rarely speak in the same manner that they write – they are generally disfluent. Disfluencies can be defined as interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously. Despite being a natural characteristic of spontaneous speech, and the rich linguistic literature that discusses their informativeness, they are often removed as noise in post-processing from the output transcripts of speech recognisers. So far, their consideration in a Spoken Language Understanding (SLU) context has been rarely explored. The aim of this thesis is to develop computational models of disfluencies in SLU, focusing on tasks related to social interaction. To do so, we take inspiration from psycholinguistic models of disfluencies, which focus on the role that disfluencies play in the production (by the speaker) and perception (of the listener) of speech. Specifically, when we use the term "computational models of disfluencies", we mean to develop methodologies that automatically process disfluencies to empirically observe 1) their impact on the production and perception of speech, and 2) how they interact with the primary signal (the lexical, or what was said in essence). To do so, we focus on two discourse contexts; monologues and task-oriented dialogues.

Our results mainly contribute to tasks in SLU related to social interaction, but also research that could be relevant to Spoken Dialogue Systems. When studying monologues, we use a combination of traditional and neural models to study the representations and impact of disfluencies on SLU performance. Additionally, we develop methodologies to study disfluencies as a cue for incoming information in the flow of the discourse. In studying task-oriented dialogues, we focus on developing computational models to study the roles of disfluencies in the listener-speaker dynamic. We specifically study disfluencies in the context of verbal alignment; i.e. the alignment of the interlocutors' lexical expressions, and the role of disfluencies in behavioural alignment; a new alignment context that we propose to mean when instructions given by one interlocutor are followed with an action by another interlocutor. We also consider how these disfluencies in local alignment contexts can be associated

with discourse level phenomena; such as success in the task. We consider this thesis one of the many first steps that could be undertaken to further research disfluencies in SLU contexts.

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing
SDS	**S**poken **D**ialogue **S**ystem
SLU	**S**poken **L**anguage **U**nderstanding
SOTA	**S**tate **O**f **T**he **A**rt
HRI	**H**uman **R**obot **I**nteraction
TTS	**T**ext **T**o **S**peech
POM	**P**ersuasive **O**pinion **M**ining dataset
LIWC	**L**inguistic **I**nquiry **W**ord **C**ount
SWBD	**SW**itch**B**oar**D** corpus
NSE	**N**on **S**entence **E**lements
MS-State	**M**ississippi **S**tate transcripts
FOK	**F**eeling **O**f **K**nowing
FOAK	**F**eeling **O**f **A**nother's **K**nowing

Contents

List of Figures

List of Tables

1 | Introduction

1.1 Motivations

In his famous lecture series on Artificial Intelligence, the late Prof. Patrick Winston stated that some 50,000 years ago, what started to separate humans as a species from the rest of the animal kingdom is the ability to start describing things, in a way that is intimately connected with language[1]. However, when the medium of language meets with the goal of communication, the process is not always straightforward. As Bernard Werber states; "Between what I think, what I want to say, what I believe I say, what I say, what you want to hear, what you believe to hear, what you hear, what you want to understand, what you think you understand, what you understand ... They are ten possibilities that we might have some problem communicating. But let's try anyway ...".

And try we do, as we often go beyond only the words used, in an effort to understand one another. As discussed in Corley and Stewart (2008), between intentional signals (e.g. a gesture deliberately used to point at an object, ...) and *un*intentional signals (e.g. slips of the tongue, ...) of communication, lies *disfluencies*; which can be defined as interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously (Fraundorf, Arnold, and Langlois, 2018). With the increasing popularity of voice assistant technologies, there is a growing need to design systems that can comprehend the different signals of speech communication. However, when taking stock of a machine's capability to understand language, there is an emphasis on the learning of *forms* (such as in a language modelling (LM), where the task is string prediction), but not on *meaning*; or the relationship *between* linguistic form and communicative intent (Bender and Koller, 2020). This issue is now widely acknowledged, for e.g. the UnImplicit workshop state as motivation "... an important question that remains open is whether such methods are actually capable of modeling

[1] Adopting the perspective from Chomsky.

how linguistic meaning is shaped and influenced by context, or if they simply learn superficial patterns that reflect only explicitly stated aspects of meaning . . . ". Thus Bender and Koller (2020) point out that this distinction between form and meaning should not be lost when considering the bigger picture of progress in the field. This is imperative to recognise in present-day, as it could easily be overlooked given the achievement of impressive benchmarks that rival human performance (e.g. in machine translation).

From a research perspective, this especially becomes challenging when considering how to computationally study spontaneous speech phenomena such as disfluencies, which often have *implicit* and *contextual* meanings, (such as when a speaker says "uh", "hmm", "er" and so on). The aim of this thesis is to computationally study the representations of disfluencies in Spoken Language Understanding (SLU)[2] (with a focus on the fillers and the discourse marker "oh"), inspired by psycholinguistic perspectives on speech. Psycholinguistic perspectives focus on the role that disfluencies play in the *production* and *comprehension* of speech. The main will be on computationally studying the role that these disfluencies have on broad tasks in SLU, relevant to social communication. Thus, this thesis is motivated by the following observations:

People rarely speak in the same manner with which they write. As Bailey and Ferreira (2003) state, "The processes involved in speaking and in writing differ substantially from each other, and so the products of the two systems are not the same". The following is an example of a transcription taken from a corps of conversational speech:

> **A:** For a while there I, I, I, uh, subscribed to New York Times, a-, actually a couple of newspapers because, uh, you know, my fiance, well, she was unemployed for a while . . .
> **B:** Uh-huh.
> **A:** . . . so she, you know, really needed to look at the, the, want-, help wanted ads.

The above transcript is not *fluent*, and if the same style was prolonged in written text, it may not be deemed "readable" by a reader. As shown, when speaking, people tend to repeat themselves ("I, I, I"), interrupt each other ("Uh-huh"), rapidly shift the focus of the topic in a conversation (from newspaper subscriptions to unemployment), and are in general, *disfluent*. One of the departures from written text is the presence of disfluencies. Disfluencies

[2]SLU is not one defined application, it combines speech processing and natural language processing (NLP) by leveraging technologies from machine learning (ML)and artificial intelligence (AI) (Tur and De Mori, 2011).

are frequent in spoken language, as spoken language is rarely fluent; with estimates that natural human-human conversations comprise of $\approx 5 - 10\%$ of disfluencies (Shriberg, 1994). **Thus disfluencies are ubiquitous to spontaneous speech.**

Despite this, there are varying attitudes in the treatment of disfluencies depending on the field of study. Generally, the field of psycholinguistics (dealing with communicative and cognitive aspects of language) focuses on the role disfluencies in the production and comprehension process of speech, with many works to show their importance in the communicative process (such as in Levelt (1983), Bailey and Ferreira (2007), Corley, MacGregor, and Donaldson (2007), Corley and Stewart (2008), Brennan and Williams (1995), Smith and Clark (1993), Arnold et al. (2004), and Barr and Seyfeddinipur (2010) ...). For e.g. they inform us about the linguistic structure of an utterance: such as in the (difficulties of) selection of appropriate vocabulary while circumventing interruption, as part of the planning process (Maclay and Osgood, 1959), to maintain the speaker turn in dialogue and so on. Other perspectives, for e.g. a *computational* one (see Heeman and Allen (1999), Shriberg, Bates, and Stolcke (1997), and Bear, Dowding, and Shriberg (1992)) are primarily concerned with the recognition disfluent speech for other purposes, such as the improvement of automatic speech recognition (ASR) systems.

Clark (1996) and Clark and Fox Tree (2002) proposed that speakers are able to utilise disfluencies as *collateral signals* in communication, in addition to the *primary signal* of the message. Colloquially, the primary signal of the message can be thought of *what* was said (in essence) and the collateral signal as *how* it was said. Consider the following example, taken from Brennan and Williams (1995):

A: Can I borrow that book?
B: ... $\{F\ um\}$ ... all right. (1.1)

Here, speaker **B** used a filler $\{F...\}$ which causes **A** to note that **B** might have had a different intention compared to if **B** answered "all right" immediately. **B** in essence says "yes" to lending the book, but the way **B** said this indicates some uncertainty in lending the book, due to the collateral signal (a disfluency) used. So far, research in Spoken Dialogue Systems (SDS) has focused on the primary signal, discarding disfluencies as noise; for e.g. as seen in Tur and De Mori (2011) "Speech Summarization", where the goal of *intent classification* and *slot filling* is to reduce the input utterance into a *semantic* frame (Louvan and Magnini, 2020). However, these expressions do

not occur in a vacuum, they are part of a spoken communication amongst interlocutors, which contain informative collateral signals. If we hope to make advancements in the field, we need to focus on the information offered by disfluencies, as cleaning speech of disfluencies removes the naturalness of speech, as well as important information on the cognitive and communicative aspects of the speech. **Thus despite the rich literature available on the informativeness of disfluencies, their consideration in an SLU context has been so far rarely explored.**

1.2 Data Specificities and Constraints

In this section, we specify *what* kind of data is used in the thesis, *which* disfluencies we choose to study, and *why* we choose these disfluencies. These choices are motivated computationally; for e.g. taking into account the balance between the size of the dataset and the feasibility of the annotation of disfluencies. Additionally, in this thesis we focus on **SLU tasks concerned with social communication**.

Data. In this thesis, we choose to focus on *transcripts* of spontaneous speech. Verbatim transcripts of spontaneous speech include disfluencies. "I think i think i think the stand they took, i support that" is an example of a verbatim transcript, where the disfluencies are highlighted in teal. While there are works that study disfluencies using not naturally elicited speech (e.g. Bailey and Ferreira (2003) used sentences containing disfluencies read aloud to judge whether listeners were perceptive to them), our aim is to study disfluencies in completely spontaneous environments. This constrains the availability of datasets, as often large spontaneous speech datasets in SLU (such as the POM dataset (Park et al., 2014))), can consist of bad acoustic conditions, such as interlocutors using their own microphone, background noise and so on. Thus **we focus on the textual modality**, and do not consider the prosodic contexts in which disfluencies occur (acoustic modality), nor the visual contexts (i.e. the visual modality, such as the link between disfluencies and gestures).

The disfluencies. In this thesis, we focus on 2 specific kinds of disfluencies; *fillers* (specifically, "uh" and "um"), and the *discourse marker* "oh". While taxonomy and distinction of what constitutes a filler, discourse marker and so on greatly differs depending on field, **we focus on the *tokens* "uh", "um" and "oh'**. This is based on the present observations:

- **Frequency.** These tokens are very frequent in spontaneous speech corpora, occupying upto 4% of the tokens (Aijmer, 1987). Despite being so frequent, they are often considered noise in SLU.

- **Annotation constraints.** Furthermore these tokens require no complex annotation, and can easily be isolated from transcripts in simple string searches. Other kinds of disfluencies may not be single token disfluencies, and often require carefully curated annotations (following the general structure as given in Eg. 1.2).

- **Informative properties.** Psycholinguistic work indicates that these tokens may have *information sharing properties*, for e.g. they may precede unexpected words (i.e. low frequency or low context words) (Beattie and Butterworth, 1979). Following this, listeners may also expect to hear a new term after hearing a filler (Arnold et al., 2004). These properties could be useful in SLU, in terms of distinguishing important information.

- **Semantic meaning.** These tokens do not explicitly have semantic meaning (Meteer et al., 1995), and their meaning is highly dependent on context. However, there is currently a focus on the semantic meaning of an utterance in the field (Ettinger, 2020), neglecting pragmatic level of analysis.

$$\text{Archie} \quad \underbrace{[\,\text{likes}}_{\text{To be replaced}} \quad + \quad \underbrace{\{\text{F uh}\}}_{} \; \underbrace{\text{loves}\,]}_{\text{Replacement}} \quad \text{Veronica.} \tag{1.2}$$

1.3 Research Objectives

Since disfluencies are natural to spontaneous speech, the literature on them is vast. The treatment of disfluencies occupies two extremes; from being considered noise, to being an informative signal of communication. We narrow down the literature to identify two issues; **i) the challenges of annotation of disfluencies in SLU and ii) computational approaches to disfluencies distinguished from psycholinguistic approaches**. These two issues serve as the basis to motivate the research undertaken in the thesis. We briefly touched on the constraints of annotation in Sec. 1.2, which motivate which disfluencies are studied and what data is used in this thesis. In this section, we focus on the different computational and psycholinguistic perspectives that motivate the research objectives.

Computational and Psycholinguistic perspectives

In this thesis, we distinguish between **a psycholinguistic approach to disfluencies** (concerned with the role disfluencies in the *planning* and *comprehension* of speech), and **a computational approach to disfluencies** (where the intent is more from an automatic *recognition/processing* standpoint[3]). Note, that there are other linguistic perspectives that do consider disfluencies as noise (e.g. Maclay and Osgood (1959)), however, we choose to focus on works that treat disfluencies as informative signals in the flow of speech.

Computational perspectives on disfluencies From a computational perspective, the intent of automatic recognition/processing of disfluencies (such as in Heeman and Allen (1999), Shriberg, Bates, and Stolcke (1997), and Bear, Dowding, and Shriberg (1992)) in present-day is largely due to automatic Natural Language Understanding (NLU) systems not being robust to them (Ginzburg et al., 2014). In a standard SDS pipeline, the output transcripts of ASR systems are cleaned of disfluencies in post-processing, using disfluency detection systems (such as Hough, Schlangen, et al. (2015), Shalyminov, Eshghi, and Lemon (2018), and Rohanian and Hough (2021)). Then, further NLU/SLU can occur on the cleaned input, where often, (particularly for task oriented dialogue), the goal is in *intent classification* and *slot filling* (such as in Tur and De Mori (2011) "Speech Summarization"). As Ginzburg et al. (2014) state, this process occurs "prior to any semantic interpretation". Consider the disfluent utterance "A train ticket to Paris uh I mean to Berlin". If "to Paris uh I mean" is not removed (to become fluent; "A train ticket to Berlin"), it leads to confusion in the subsequent (semantic) processing of the utterance, with the need to identify the correct intent (i.e. book train tickets) and also slot (i.e "Berlin" not "Paris"). With the advancements of voice assistant technologies, there is an increasing awareness of the problems disfluent utterances may pose on other tasks, such as in Gupta et al. (2021), which show their role in the confusion of Question Answering (QA) systems.

With this is mind, there has been an interest in studying the characteristics of disfluencies, such as their distributional properties (Shriberg, 1994; Shriberg, 2001; Vasilescu, Rosset, and Adda-Decker, 2010b; Meteer et al., 1995) in corpora, and analysis at various linguistic levels; for e.g. phonetic

[3]In our distinction, we include certain linguistic studies in a computational approach, if the ultimate intent is on recognition/processing. For e.g., several phonetic works (such as Shriberg (1995) and Shriberg and Lickley (1993)) that characterise the acoustic environments that disfluencies occur in, in the ultimate goal of aiding ASR systems.

characteristics (Shriberg, 1999; Shriberg and Lickley, 1993) (including for e.g. work to make cross-lingual systems more robust to disfluencies (Candea, Vasilescu, and Adda-Decker, 2005)), morpho-syntactic characteristics (Goryainova et al., 2014), pragmatics (Shriberg et al., 1998), contextual occurrence (Vasilescu, Rosset, and Adda-Decker, 2010b) and so on. Some works also make a distinction between *informative* (e.g. as pragmatic markers of discourse structure) and *uninformative* disfluencies ("truly disfluent events") (Vasilescu, Rosset, and Adda-Decker, 2010a), thus distinguishing which disfluencies may be useful in subsequent NLU processing.

Considering another different dimension to the issue, from a perceptual standpoint, there is research to suggest that disfluencies are filtered out even by the human listener. For e.g. prosody can remain natural sounding even after the removal of certain disfluencies (Fox Tree, 1995), and listener's when asked are unable to pinpoint the exact locations of disfluencies (Lickley, Shillcock, and Bard, 1991; Lickley and Bard, 1998). However, Bailey and Ferreira (2003) point out that the idea of "filtering" out disfluencies is implausible given the *incremental* nature of speech processing; i.e. "the processing starts before the input is complete" (Schlangen and Skantze, 2011). The idea of filtering would indicate that processing has to wait, needs to occur *after* the removal of disfluencies.

Thus an overall problem in SLU (particularly SLU for SDS) is the narrow focus only on the semantic aspect of the utterance, which ignores the *implicit* meanings of disfluencies. The problem then, as stated in Clark and Fox Tree (2002), remains about how to merge the two signals – **and in the case of the thesis, how to *computationally* study the interaction of the two signals**. Thus there is a need to move towards an automatic but *holistic* analysis of the two; if we hope to move towards better models and understanding of spontaneous speech.

Furthermore, the meanings of disfluencies are *contextual*. In broader SLU tasks (where there is overlap with fields such as personality/affective computing – here the acoustic signal may be considered with other modalities; such as the textual/visual modality, for tasks such as emotion detection, detection of the engagement of the listener ...), disfluencies can be considered an informative *social signal* (see Schuller et al. (2019), Vinciarelli and Mohammadi (2014), Mairesse et al. (2007), and Ekman et al. (1980)). However, the drawback here is often the lack of *contextual* analysis in these works. One such example of this, is areas of research that may predetermine based on the context (such as in job interviews, where fluency is assumed to be desirable) whether disfluencies are positive or negative (see Rasipuram and Jayagopi (2016)). While indeed, the speaker's use of disfluencies may have an effect on hireablity; **it does not account for the nuances of the ways**

disfluencies may be used. In this aspect, the link between disfluencies and a wide variety of phenomena is to be expected, with Barr (2001) even describing fillers as *vocal gestures*. Thus, it may not always be a case of "paying attention to X is correlated with Y" (Boyd and Schwartz, 2021)[4]. It is also worth noting that neural architectures may not clean utterances of disfluencies per say, or indeed, do any preprocessing (end-to-end systems such as Serban et al. (2016), or even bidirectional models that simply utilise the entire acoustic/textual input), the same problem remains as to the *nuances* of the use of disfluencies, i.e, there is little interpretability or understanding of the disfluencies within the message, nor how disfluencies interact with the rest of the message. Thus it is not simply a question of ***whether* disfluencies are noise in SLU, but *what* are their properties that could be leveraged for SLU**.

Psycholinguistic perspectives on disfluencies A psycholinguistic perspective on disfluencies is concerned with the role of disfluencies in the planning/production (by the speaker) and comprehension (by the listener) of speech.

A common theme in the planning process is the idea of a *cognitive load*, i.e. the amount of cognitive effort required in the planning process. Disfluencies were found to occur at the start of an utterance, due to higher cognitive load in planning an utterance (Maclay and Osgood, 1959). Shriberg (2001) found that the longer the utterances, the more disfluencies they contain, also suggesting an increase in cognitive load of the speaker. Beattie and Butterworth (1979), suggested instead that there is an *element of choice* in the planning process based on their findings. They first establish that generally, speakers are disfluent both with low frequency words and improbable words in the context. However, even when frequency of a word was maintained, disfluency still indicated low contextual probability of a word.

Like this, there are two main positions behind the speaker's production of disfluencies. One is that disfluencies are accidentally caused in speech due to *cognitive burden* of the speaker (e.g. in Bard, Lickley, and Aylett (2001)). Other works study disfluencies as an important *communicative function* used in dialogue. Here, the speaker strategically updates the listener using disfluencies as cues, for e.g. as discussed in Beattie and Butterworth (1979). The distinction between the two views is that the former is an unconscious

[4]There are fields in SLU that do specifically study the informativeness of disfluencies, such as clinical SLU (Rohanian, Hough, and Purver, 2021). From a generation perspective (Text-to-Speech (TTS)), works do acknowledge that disfluencies are part and parcel of spontaneous speech (Leviathan and Matias, 2018; Skantze, Johansson, and Beskow, 2015). These are briefly discussed in Chapter 2.

by-product of an overburdened system, while the latter, an intentional and strategic production. The results for the production of disfluencies are often mixed, for e.g. in Nicholson (2007) and Yoshida and Lickley (2010a), with evidence suggesting that they occur in both cases.

Many works also focus on the *comprehension* of disfluent speech, i.e. taking into account the listener's understanding of the speaker's disfluencies (Corley and Stewart, 2008), and not on why the disfluency itself was produced (Nicholson, 2007). As Corley and Stewart (2008) state, "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech". Many works then study the effect of disfluencies on listener comprehension (e.g. (Swerts, 1998; Arnold et al., 2004; Corley, MacGregor, and Donaldson, 2007; Barr and Seyfeddinipur, 2010)). Bailey and Ferreira (2003) state that listeners *must* have developed a comprehension system to process disfluencies, given how frequent they are in spoken language. Thus, whether or not the speaker intentionally used disfluencies (Barr and Seyfeddinipur, 2010), the listener still *processes* them. The listener comprehension of disfluencies is a crucial step in understanding the incoming flow of speech, **our preliminary objective is to computationally study the role of disfluencies on the listener's perception**. Having established this, then we move onto further analysis on disfluencies in the listener-speaker dynamic.

This relates in part to a point made by Tran et al. (2017b); one of the main difficulties in modelling/integrating prosodic cues (in a conversational parsing task) is due to variation of speaker style. The same is the case for disfluencies (and indeed, Tran et al. (2017b) found the most improvements in parsing when adding prosodic cues to disfluent/longer sentences, compared to fluent ones). To leverage research from psycholinguistics for SLU on the informativeness of disfluencies, there is a **need to separate granularities on larger datasets**; i.e *generalisable patterns* of disfluencies may be overall true, and *nuances* of disfluencies applicable more contextually, such as in the extreme labels of a dataset. This is because **utterances build up as a *function of* discourse**, and thus there is a need to study disfluencies both taking into account fine-grained *utterance level* information and *global* discourse level information, which to our knowledge, has little been explored outside few works (e.g. see Mills and Healey (2006)).

In this thesis when we use the term **"computational study of disfluencies"**, we mean to develop methodologies that automatically process disfluencies to empirically observe i) their *impact on* the production and comprehension of speech, and ii) how they *interact with* the primary signal (the lexical, or what was said in essence). Since we use psycholinguistic perspectives as inspiration, here processing is not for the purposes of removal

(like the computational perspectives discussed), but rather, for the purpose of understanding their informativeness. Bearing these perspectives in mind, concretely, the research objectives are as follows:

$\mathcal{O}_1$: Monologues – What is the impact of disfluencies on the performance of SLU systems?

The production and perception of fillers in monologues. In this objective (see Part II), we would like to develop methodologies to study the different *production contexts* of fillers, and the effect they have on the listener's *perception*[5]. We specifically would like to study this using ML models, to see the impact of these automatically computed production contexts in the prediction of the listener's perception in SLU models. The objective is to first focus on a dataset of monologues, eliminating *dialogue-related disfluencies* such as backchannels, which allows us to focus uniquely on the information offered by fillers. Here, we study fillers at a *global* level.

$\mathcal{O}_2$: Monologues – What is the local level interaction of fillers with the primary signal?

Fillers in the process of information sharing in the flow of the discourse In this objective (see Part III), we would like to explore the *contextual* use of fillers, specifically in relation to the *primary signal*, or what was said in essence. We will propose computational methods based on statistical analysis in order to study the impact of fillers on the primary signal, from local to global. In this objective, we treat the message from the speaker as *an incoming source of information that builds up in the discourse*, and then we *observe the function of fillers in this process*. Thus, we computationally study at a micro-level, which fillers are informative in terms of the new information specific to the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception.

$\mathcal{O}_3$: Task-Oriented Dialogues – What are the Roles of Disfluencies in the Listener-Speaker Dynamic?

The function of disfluencies in the communication of the primary signal In this objective (see Part IV) we would like to computationally

[5]While psycholinguistic work uses the term "comprehension" of disfluencies, we focus on the how the listener's *perception* may change depending on the production context of the disfluency.

study the function of fillers and the discourse marker "oh" in terms of their interaction with the primary signal. Here, we utilise a dataset of children's dialogues, which was designed to encourage collaboration as they engaged in a collaborative learning activity. In this regard, while local utterance level information *also builds as a function* of global, discourse phenomena (here, task success), there is a much more *dynamic, interactive context to consider amongst interlocutors* in dialogues. Thus specifically when we speak of primary signals in this context, the we mean that the objective is to computationally study the function of disfluencies in the *alignment/ or development of shared representations* between interlocutors at different linguistic levels. We study the function of fillers in *verbal alignment* (or lexical alignment), and also the function of "oh" in behavioural alignment (we propose a new alignment context to mean when instructions provided by one interlocutor are *followed* with physical actions by the other interlocutor). We then, see how these local alignment contexts build into a function of dialogue level phenomena; i.e. success in the task.

Throughout the thesis, these objectives are subdivided into chapters based on methodological approaches. For e.g., in $\mathcal{O}_1$, the chapters are divided into i) a preliminary work consisting of statistical approaches and linear models, and ii) a chapter based on SOTA models and unsupervised methods. Please see Table 1.1 for a concrete breakdown of the chapters and from their methodological perspectives. Also note that throughout this thesis, we study disfluencies on the macro-level (i.e over the entire discourse of either monologues or dialogues) which we refer to as the *global/discourse* level, and the micro-level referred to as the *local/utterance* level.

1.4 Contributions

Table 1.1 shows the thesis structure in terms of each chapter's contributions to the main objectives[6].

The contributions of the thesis in regards to $\mathcal{O}_1$ point to the *importance of fillers in both the speaker's production of the message, and the listener's overall impression of the message* in SLU systems, studied using a dataset of monologues. We thus propose methodologies to study fillers in this production/comprehension context.

The contributions of the thesis in regards to $\mathcal{O}_2$ show that some *contextual use of fillers, specifically in relation to the* primary signal*, may not*

[6]See the Conclusions chapter for full account of the contributions.

	Chapters				
	C4	C5	C6	C7	C8
$\mathcal{O}_1$: *Monologues:*					
Impact of disfluencies on SLU performance?					
$\mathcal{O}_{1.1}$ *Global Level*					
$\mathcal{O}_{1.11}$ *Statistical and Linear Models*	X				
$\mathcal{O}_{1.12}$ *SOTA Models, Unsupervised Methods*		X			
$\mathcal{O}_2$: *Monologues:*					
Local Level of Fillers and the Primary Signal?					
$\mathcal{O}_{2.1}$ *Global, Local Level*					
$\mathcal{O}_{2.11}$ *Statistical Analysis*			X		
$\mathcal{O}_3$: *Task-Oriented Dialogues:*					
Disfluencies in the Listener-Speaker Dynamic?					
$\mathcal{O}_{3.1}$ *Verbal Alignment*					
$\mathcal{O}_{3.11}$ *Local Level: Rule-based Algorithms*				X	
$\mathcal{O}_{3.12}$ *Global Level: Statistical Analysis*				X	
$\mathcal{O}_{3.2}$ *Behavioural Alignment*					
$\mathcal{O}_{3.21}$ *Local Level: Incremental Algorithms*					X
$\mathcal{O}_{3.22}$ *Global Level: Statistical Analysis*					X

Table 1.1: Methodological contributions to the thesis objectives, where contributions are shown for each chapter.

be perceived by the listener. In this objective, we propose methodologies to treat the message from the speaker as an incoming source of information that builds up in the discourse, and then we empirically observe the role of fillers in this process. Specifically, we study at a micro-level, which fillers are informative in terms of the new information specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception.

Lastly contributions of the thesis in regards to $\mathcal{O}_3$ show *the communicative functions of disfluencies*, specifically their association with verbal and behavioural alignment. Overall, the study of disfluencies in the multi-level alignment theory of Pickering and Garrod (2004) is one rarely considered in existing literature (let alone, computational study). In dialogues while local utterance level information also builds as a function of global, discourse phenomena (here, task success), there is a much more *dynamic, interactive context to consider amongst interlocutors* in dialogues. We thus propose methodologies to see how these local alignment contexts (verbal and behavioural) build into a function of dialogue level phenomena; i.e. success in the task. Then, we empirically observe the function of disfluencies in this process.

Our contributions in terms of publicly available code and datasets are:

- **JUSThink Dialogue and Actions Corpus:** Dataset of transcribed children's dialogues based on the already collected JUSThink dataset (Nasir et al., 2020a; Nasir et al., 2020b), specifically to study alignment[7] (contains open source transcripts we provided, with aligned event logs and test responses, taken from the original dataset).
- **JUSThink Alignment Analysis:** Tools to study verbal and behavioural alignment in the JUSThink Dialogue and Actions Corpus; a children's dataset of situated dialogues[8].
- Tools to study the representation of fillers in BERT[9]
- Tools to do text processing on the widely used Persuasive Opinion Mining Dataset (POM)[10] (including notes on issues in the data).
- Tools to process the Sentiment annotated POM dataset (Garcia et al., 2019)[11] (including notes on issues in the data).

1.5 Challenges

In the duration of this thesis, we faced several challenges in studying disfluencies in such a context. The main challenges we encountered are:

- **What even are disfluencies? Can we agree on terminology?** Since disfluencies are so common to spontaneous speech, they are studied across corpora and domains; which leads to varied terminology. For e.g. some works consider paralinguistic sounds ("coughs", "laughter") as disfluencies; as they do in some way contribute to an *interruption* in the fluent flow of speech. This varied terminology in turn, leads to a lack of transparency on the findings of *the same phenomena under investigation, but under a different name.* In Chapter 2, we compare two predominantly used datasets in SLU, as a way to discuss terminological issues.

[7] https://zenodo.org/record/4675070
[8] https://github.com/chili-epfl/justhink-alignment-analysis
[9] https://github.com/eusip/SLU-Fillers/tree/v1.1
[10] https://github.com/tdinkar/fillers_in_POM
[11] https://github.com/eusip/POM

- **A hydra-headed research problem** There is a (over)saturation of works concerning disfluencies in both fields (SLU and linguistics), but not enough connectivity across studies. In this thesis, we try to be mindful of this issue by making a distinction between generalisable patterns of disfluencies, and also findings that may be more contextual and specific to the datasets investigated.

- **Which disfluencies and why?** In addition to the number of research questions that could be asked about disfluencies, a problem remains about *which specific disfluencies to study and why*. While we briefly outlined this in Sec. 1.2 based on the general annotation problems of disfluencies in SLU, we use this question as an opportunity to discuss the general complexities of transcribing disfluent speech in SLU in Chapter 2.

- **Issues with datasets** Related to the previous point, we faced several challenges related to datasets. A first challenge is that previous research has shown the difficulties of transcribing fillers (Le Grezause, 2017), and therefore, when selecting datasets, one needs to *carefully consider transcriber experience*. A second issue when selecting datasets, is that the more "natural" and "spontaneous" the dataset may be, the more it could have poor audio conditions. An example of this is the POM dataset (Park et al., 2014) that we predominantly study in this thesis. Here, speakers voluntarily recorded themselves giving a movie review (often, in the comfort of their own homes), which is desirable; as it allows for the study of disfluencies outside limited, laboratory contexts and on larger datasets. However, since personal equipment was used for the recordings, the forced alignment algorithms used on the dataset gave poor alignment results between transcripts and audio. Lastly, we are limited in terms of the *availability of annotator labels* (specific to psycholinguistic research) in the dataset. In part, our solution to this is to not be constrained only by labels from the dataset, for e.g., in Chapter 7 and Chapter 8 we propose rule-based algorithms to automatically annotate verbal and behavioural alignment. As mentioned previously, we also carefully considered these datasets, both for availability of good quality transcriptions but also relevant labels; before being used in the thesis.

1.6 Organisation of the Thesis

The outline of the thesis is as follows:

- In Chapter 2, we narrow down to the two main issues, expanding on our Research Objectives. We first describe the computational approaches to disfluencies distinguished from psycholinguistic approaches in Sec. 2.1. Then, we discuss the challenges of annotation of disfluencies in SLU in Sec. 2.2. Given these two issues, in Sec. 2.3 we lastly expand on why we focus on the tokens "uh", "um" and "oh" studied in this thesis.

- In Chapter 3, we describe the two datasets we choose to study in this thesis; a dataset of monologues and a dataset of task-oriented dialogues. Throughout the results chapters (Chapter 4 - Chapter 8), we briefly describe the dataset used, but refer readers back to this chapter. We also describe the annotation scheme of disfluencies used in these two datasets.

- In Chapter 4 we conducted a preliminary analysis from a *production and perception* perspective to study the relationship between fillers, and the listener's perception of the speaker.

- In Chapter 5 we do an analysis from a *production* perspective in an unsupervised task (spoken language modelling (SLM)), to study the representation of fillers in deep contextualised word embeddings. We then analyse from a *perception* standpoint the information offered from fillers without hand-crafted features on two supervised tasks; the prediction of a listener's estimation of i) a speaker's metacognitive state and ii) a speaker's stance.

- In Chapter 6 we focus more on the *contextual interaction of fillers with the primary signal*. We propose methodologies in order to study the interaction of disfluencies on the primary signal, from local to global. Specifically, we computationally study at a micro-level, which fillers are informative in terms of the information new and specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception.

- In Chapter 7 we move onto task-oriented dialogues. We propose rule-based algorithms to specifically extract verbal alignment contexts (or lexical alignment) in situated dialogues. Then, we do a i) statistical analysis to study local patterns of fillers in the context of verbal alignment between interlocutors; from the first time an interlocutor introduces an expression to the time an expression becomes part of a *shared vocabulary*, and ii) statistical analysis to show the role of fillers used in

the context of verbal alignment in relation to global variables of task success.

- In Chapter 8 we propose a novel, rule based algorithms to study behavioural alignment[12]. Then, we do a statistical analysis to study i) the local patterns of the discourse marker "oh" in the context of behavioural alignment between interlocutors and ii) the role of "oh" used in the context of behavioural alignment, and its relation to variables of task success.

- Lastly, in Chapter 9, we give a global conclusion about the work done, and give perspectives about the future work that could build upon this work.

1.7 Publications

The contents of this thesis are partially based on the following publications (listed according to appearance in thesis):

1. **Dinkar, T**., Vasilescu, I., Pelachaud, C., & Clavel, C. (2018, November). Disfluencies and teaching strategies in social interactions between a pedagogical agent and a student: Background and challenges. In *SEMDIAL 2018 (AixDial), The 22nd workshop on the Semantics and Pragmatics of Dialogue* (pp. 188-191). Laurent Prévot, Magalie Ochs and Benoît Favre. Paper.

2. **Dinkar, T**. Representations of fillers and discourse markers in SLU: a psycholinguistic approach. In *YRRSDS 2021, Young Researchers' Roundtable on Spoken Dialogue Systems*. Paper

3. **Dinkar, T.**, Vasilescu, I., Pelachaud, C., & Clavel, C. (2020, May). How confident are you? exploring the role of fillers in the automatic prediction of a speaker's confidence. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8104-8108). IEEE. Paper.

4. **Dinkar, T.***, Colombo, P.*, Labeau, M., & Clavel, C. (2020, November). The Importance of Fillers for Text Representations of Speech Transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7985-7993). Paper.

[12]Which we defined as when instructions given by one interlocutor are followed with an action by another interlocutor, in a timely manner.

5. **Dinkar, T.**, Biancardi, B., & Clavel, C. From local hesitations to global impressions of a speaker's feeling of knowing. In *SEMDIAL 2021, The 25th workshop on the Semantics and Pragmatics of Dialogue* Paper.

6. **Dinkar, T.**, Biancardi, B., & Clavel, C. From local hesitations to global impressions of a listener. In *ICNLSP 2021.* Paper.

7. *(Submitted: Dialogue and Discourse journal)* Norman, U.*, **Dinkar, T.***, Bruno, B., Clavel, C (March 2021). Studying Alignment in Spontaneous Speech via Automatic Methods: How Do Children Use Task-specific Referents to Succeed in a Collaborative Learning Activity? Preprint: Paper.

I | Background

2 | Perspectives and Challenges

Since disfluencies are ubiquitous to spontaneous speech, the treatment of disfluencies has many approaches. In this thesis, we narrow down the literature into two main issues; **i) computational approaches to disfluencies distinguished from psycholinguistic approaches and ii) the challenges of annotation of disfluencies in SLU**. Regarding the challenges of annotation, **though we focus on the tokens "uh", "um" and "oh" in this thesis**, we still discuss the broad complexities of annotation of disfluencies, and consequently, some difficulties of data collection for SLU. It is these complexities, that lead us to selecting the datasets and specific disfluencies studied in the thesis.

Thus in this chapter, we narrow down to the two main issues, expanding on our Research Objectives as given in Chapter 1[1]. We first describe the computational approaches to disfluencies distinguished from psycholinguistic approaches in Sec. 2.1. Then, we discuss the challenges of annotation in SLU in Sec. 2.2. Given these two issues, in Sec. 2.3 we lastly expand on why we choose these disfluencies studied in this thesis.

2.1 Computational and Psycholinguistic perspectives

We loosely adopt the distinction between psycholinguistic perspectives and computational perspectives as given in Lickley (1994).

Generally, a psycholinguistic perspective is concerned with the role disfluencies of in the *planning/production* of speech (by the speaker), and the *comprehension* of speech (by the listener). Broadly, we focus on psycholinguistic theories that deal with the communicative, and cognitive aspects of

[1]Throughout the thesis, we briefly describe the background relevant to that specific chapter, summarising from this chapter. Sometimes we include some additional *related works* that we have not discussed in this chapter, that are not specific to disfluencies, but may be relevant to the models/theoretical frameworks used in that chapter.

language. Specifically regarding speech, we focus on theories of speech production and comprehension, for e.g. in Levelt (1983), studying how speakers monitor and correct their speech, and in turn, how listener's are able to integrate new material correcting the previous material. Additionally, **our focus is more on fillers; highlighting works that study them as cues for integrating information**, but also discuss works on other disfluencies when relevant.

We then distinguish psycholinguistic perspectives on disfluencies with computational perspectives on disfluencies (for e.g. Heeman and Allen (1999), Shriberg, Bates, and Stolcke (1997), and Bear, Dowding, and Shriberg (1992)); i.e approaches more concerned with the *recognition/processing* of disfluencies. Within the computational perspectives, we distinguish between the treatment of disfluencies studied in broader SLU (i.e. SLU tasks concerned with social communication), and SLU as is traditionally considered (i.e. SLU for SDS). Note, in this thesis **we focus on SLU tasks concerned with social communication**.

2.1.1 General Psycholinguistic perspectives

We had discussed briefly the perspective of works that study the planning/production of speech by the speaker, and the comprehension of speech by the listener.

Note, **there are several other linguistic approaches not discussed in this thesis**, such as sociolinguistic approaches (e.g. studying the effect that gender, regional background etc. have on the the production of disfluencies in Shriberg (2001)). Some of these perspectives, including works from psycholinguistics, do consider disfluencies as noise (e.g. Maclay and Osgood (1959)). However, they are not relevant to this thesis, as **we are only concerned with works that treat disfluencies as signals of information in the flow of speech**.

We utilise both terms "planning" *and* "production", because it is not clear how intentional or unintentional disfluencies are as signals uttered by the speaker (e.g. in Nicholson (2007)). Additionally, we utilise the term "perception" in our results chapters, as we are concerned with the impact disfluencies may have on the perception of the flow of speech. However, psycholinguistic works often utilise the term "comprehension" to specifically study the *integration* of disfluencies by the listener, and thus term is utilised in this chapter when appropriate.

The Production of Disfluencies

Cognitive load A common theme in the planning process is the idea of a *cognitive load*, i.e. the amount of cognitive effort required in the planning process. The analysis of the production of disfluencies as a consequence in the planning process of speech has been researched at various linguistic levels.

For e.g., acoustic analysis of disfluencies identifies characteristics of the *planning process of the speaker*. O'Shaughnessy (1995) found that speakers tend to maintain a fixed speaking rate during most utterances, but often adopt a faster or slower rate, depending on the cognitive load. Acoustic analysis thus reveals that disfluent speech may be due to cognitive load; i.e speakers slow down when having to make unanticipated choices and accelerate when repeating some words. Plauché and Shriberg (1999) found that clustering prosodic features also revealed subsets of repetitions (when what was previously said is repeated exactly) that each reflect *different problems in planning*. Shriberg (2001) identifies two groups of speakers – *repeaters*, who produce more repetitions than deletions, and *deleters* (when previous material that was uttered is abandoned), who produce more deletions than repetitions. The *repeater-deleter* difference is not solely due to stylistic differences; acoustic analysis shows that deleters have a faster speaking rate than repeaters in terms of words per unit time. The interpretation suggested is that faster speakers (deleters) "get ahead of themselves", and recant what was said to begin again, while speakers with a slower speaking rate (repeaters) "take more time to plan", leading to an increase in repetitions.

Analysis of disfluencies at other levels also reveal characteristics of the planning process. In an acoustic-syntactic analysis, it was found that high-frequency monosyllabic function words (such as "the" or "I") are more likely to be longer or have a fuller form when there are neighbouring disfluencies (such as fillers), and indicate that the speaker was encountering problems in planning the utterance (Bell et al., 2003). Disfluencies were found to occur at the start of an utterance, due to higher cognitive load in planning an utterance (Maclay and Osgood, 1959). Shriberg (2001) found that the longer the utterances, the more disfluencies they contain, also suggesting an increase in cognitive load of the speaker.

Strategic modelling Based on their findings Beattie and Butterworth (1979), suggested that instead of cognitive load, there is an *element of choice* in the planning process. They first establish that generally, speakers are disfluent both with low frequency words and improbable words in the context. However, even when frequency of a word was maintained, disfluency still indicated low contextual probability of a word. Like this, there are two

main positions behind the speaker's production of disfluencies. One is that disfluencies are accidentally caused in speech due to *cognitive burden* of the speaker (e.g. in Bard, Lickley, and Aylett (2001)). Other works study disfluencies as an important *communicative function* used in dialogue. According to Nicholson (2007), this view is based on the *strategic modelling view*, where the speaker strategically updates the listener using disfluencies as cues, for e.g. as discussed in Beattie and Butterworth (1979). The distinction between the two views is that the former is an unconscious by-product of an overburdened system, while the latter, an intentional and strategic production. Often studies will look at both of these positions, by analysing the individual disfluencies of a speaker as well as the collective disfluencies produced by interlocutors. The results for the production of disfluencies are often mixed, for e.g. in Nicholson (2007) and Yoshida and Lickley (2010b), with evidence suggesting that they occur in both cases.

Throughout the thesis, we are inspired by the literature that suggests that disfluencies (focused on fillers) are linked in different ways to the planning process in speech, and specifically, the cognitive properties as discussed. We are not concerned with whether they were intentionally uttered or not (but for this reason, only use listener annotated labels when working with qualitative assessments). In Chapter 4, we design a set of filler features at a macro-level based on research to suggest the links between disfluencies and the planning process; for e.g. considering their position in the utterance, frequency in discourse and so on. In Chapter 5 we study the representations of fillers using deep contextualised embeddings, specifically focusing on the use of fillers to potentially inform about lexical and syntactic choices in production.

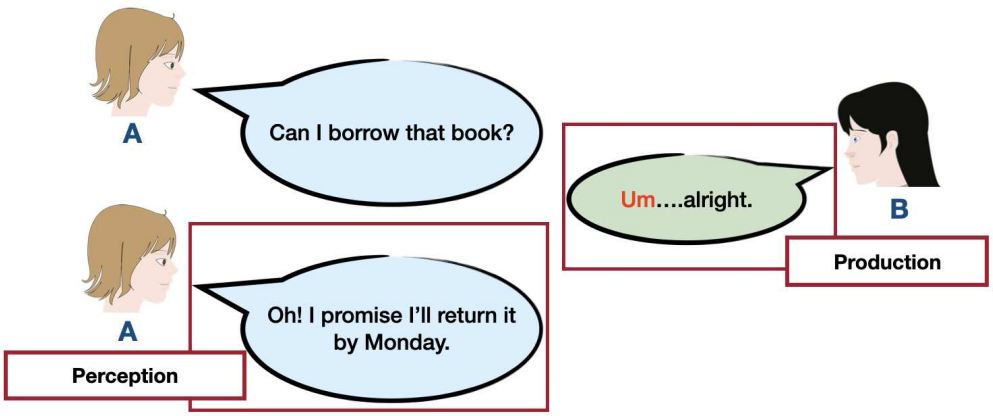


Figure 2.1: The production and perception of speech

The Comprehension of Disfluencies

Many works also focus on the *comprehension* of disfluent speech, i.e. taking into account the listener's understanding of the speaker's disfluencies (Corley and Stewart, 2008), and not on why the disfluency itself was produced (Nicholson, 2007). As Corley and Stewart (2008) state, "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech". Many works then study the effect of disfluencies on listener comprehension (e.g. (Swerts, 1998; Arnold et al., 2004; Corley, MacGregor, and Donaldson, 2007; Barr and Seyfeddinipur, 2010)). Bailey and Ferreira (2003) state that listeners *must* have developed a comprehension system to process disfluencies, given how frequent they are in spoken language.

As cues for integrating information What is fascinating to us, is the works that reveal that listeners use disfluencies as signals to *understand and resolve incoming information* in the flow of speech, regardless of whether the speaker intentionally used disfluencies in that way. Indeed, this is a different attitude towards disfluencies in SLU, for e.g. their role has been studied in the *confusion* of SLU systems (Gupta et al., 2021). Research suggests that fillers helped in the faster recognition of a target word for listeners, indicating that they cause listeners to pay more attention to the upcoming flow of speech (Fox Tree, 1995). The use of fillers also *biases* listeners towards new referents rather than ones already introduced into the discourse (Arnold et al., 2004). Arnold, Kam, and Tanenhaus (2007) additionally showed that listeners have expectations on the upcoming material to contain *difficult to describe*/unconventional referents when preceded by disfluencies. Listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle) (Barr and Seyfeddinipur, 2010). This this result was found to be *speaker specific*, it depended on what was old and new for the current speaker; not just what was old or new for the listener.

Barr and Seyfeddinipur (2010) suggests that this is evidence for the *perspective taking account of language comprehension*; it is clear that listeners interpret fillers as delay signals, and infer plausible reasons for the delay by taking the speaker's perspective. These results are exciting, as they suggest that fillers can have *metacognitive (i.e. assessment of knowledge state) effects*, with the listener using fillers as cues to interpret the speaker's *metacognitve state*. Fillers and prosodic cues were also found to impact listener's attributions of a speaker's metacognitive state, specifically the estimation of a speaker's level of certainty on a topic (Brennan and Williams, 1995).

These processes of comprehension are continuous and *incremental*. Research has shown that the comprehension of disfluencies are part of the incremental processing of the flow of speech; e.g. Bailey and Ferreira (2003) shows disfluencies can affect the internal syntactic parser of the listener, and Brennan and Schober (2001) shows that listener's may use disfluencies as cues to avoid integrating what they deem to be incorrect material in online processing.

Disfluencies as cues to understand new information has even been shown *neurologically*. Corley, MacGregor, and Donaldson (2007) studied the effect of hesitation ("um") on the listener's comprehension using the N400 function of an Event-related potential (ERP). The $N400$ effect can be observed during language comprehension, typically occurring 400ms after the word onset; and exhibits a negative charge recorded at the scalp consequent to hearing an unpredictable word. They first established the $N400$ effect in unpredictable words compared to predictable words. Then, when using hesitations preceding the unpredictable word, *the* $N400$ *effect in listeners was visibly reduced*. In a subsequent memory test on the listener, words preceded by hesitation were more likely to be remembered. Thus, whether or not the speaker intentionally used disfluencies (Barr and Seyfeddinipur, 2010), the listener still *processes* them.

Since the listener comprehension of disfluencies (particularly fillers) is a crucial step in understanding the incoming flow of speech, a primary objective is to **computationally study the role of fillers on the listener's perception in SLU systems**. Furthermore, effects of comprehension range from positive (e.g. may help in the faster recognition of target words (Fox Tree, 1995)) to negative (e.g. may lead to estimations of a speaker's level of knowledgeability in a topic (Brennan and Williams, 1995)). There might be, as stated in Vasilescu, Rosset, and Adda-Decker (2010a), *informative* versus *uninformative* disfluencies, and thus *context* is very important to distinguish these speech events. A second objective, given conflicting literature on listener comprehension is to **study whether different contexts of fillers may have different effects on the listener's perception** (see Part II and Part III).

Processing time hypothesis One might question, if all these effects are simply due to the *processing time hypothesis*, i.e. is disfluent speech more memorable/noticeable simply because disfluencies add more time to the speech utterance (causing a listener to invariably give more attention to the utterance)? Fraundorf and Watson (2011) examined this in a study on how fillers affect the memory of the listeners; by comparing fillers versus

coughs of equal duration spliced into fluent speech. They found that fillers facilitated recall, and coughs negatively hampered recall accuracy. Disfluent speech is hence more likely to be remembered by the listener, and this is not solely based on the additional time added to the utterance.

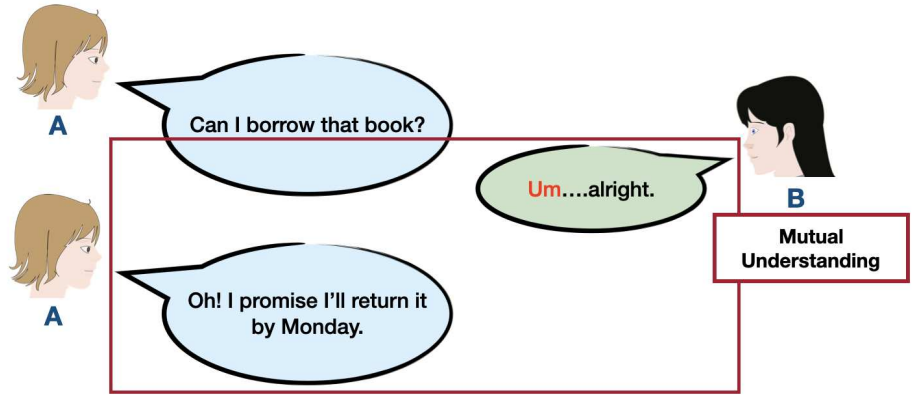
Fraundorf and Watson (2011) also manipulated the location of the fillers in speech, to study the effect of *position* of fillers on comprehension. This was based on the findings of Swerts (1998), who found that following fillers, listeners may expect a speaker to shift topics, as they carry information about larger topical units – therefore, acting as cues for discourse structure. However, Fraundorf and Watson (2011) found that fillers benefit listener's recall accuracy regardless of it's typical or atypical location. Tottie (2014) found that fillers are noticed more when overused or used in the wrong context, so while they may facilitate recall regardless of location, they still may be more noticeable in atypical locations. Indeed, we do find in the thesis that the positions of fillers do influence other perceptions the listener has of the speaker; such as their overall estimation of the speaker's metacognitive state.

The Listener-Speaker Dynamic

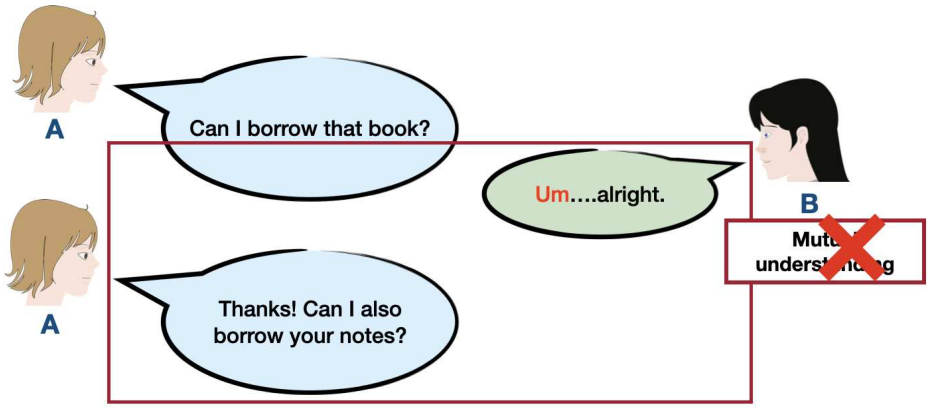
So far, we have spoken about the process of planning and comprehension, not speaking about how it is often (taking the context of dialogues) an *iterative and dynamic* process. There are theoretical psycholinguistic works on this listener-speaker context which are concerned with *mutual understanding*; meaning that any contribution to a conversation should be mutually understood by interlocutors (Clark and Wilkes-Gibbs, 1986). This involves not only a continual *assessment of one's own knowledge state* (i.e *metacognition*) (Brennan and Williams, 1995), but an estimation of the knowledge state of the other interlocutor. Speakers ideally design their utterances taking into account this *joint responsibility* of interlocutors, and monitor listeners for evidence of their understanding (Brennan and Williams, 1995).

In Figure 2.2, if **A** had replied to **B**'s "...{*Fum*} ...all right" with "Thanks! Can I also borrow your notes?", we could say that **A** did not pick up on **B**'s reluctance to lend the book, and mutual understanding did not occur. **B** might in turn reply trying to clarify her reluctance to lend the book, in an effort to be mutually understood. It is important that the listener correctly interprets the utterance (Clark and Schaefer, 1989; Clark and Brennan, 1991). The mutual belief that what the speaker meant has been understood by listener is what Clark and Schaefer (1989) and Clark and Brennan (1991) refer to as *grounding*.

We have currently defined this notion imagining that interlocutors contribute to their common ground, i.e. build up a shared mutual understanding.



A picks up on **B**'s reluctance



A does not pick up on **B**'s reluctance

Figure 2.2: According to Clark and Wilkes-Gibbs (1986), contributions to conversations should be mutually understood by both interlocutors.

However, Pickering and Garrod (2004) stress on the importance of an automatic *priming* mechanism, i.e. when a listener encounters an utterance from the speaker, it is more likely that subsequently the listener will produce an utterance by using this representation. Here, the listener need not explicitly assent-to/accept the speaker's contribution; instead, there can simply be an implicit common ground unless a misunderstanding requires changing the representation.

In this thesis, we **computationally study the role of disfluencies in alignment contexts; both at a verbal level, and at a behavioural level** (see Part IV). By behavioural alignment, we mean when instructions given by one interlocutor are followed or not followed by the other interlocutor with concrete actions. While in the previous section, we discussed disfluencies

and their impact on the listener's processing of the incoming flow of speech (including findings that could lead to mutual understanding), we propose automatic rule-based algorithms to understand these alignment processes at a micro level.

2.1.2 General Computational Perspectives

Previously we had briefly discussed a computational perspective; i.e. works that computationally deal with the processing/recognition of disfluencies with the intent of removing them so that the utterance is more "fluent" and text like. While it is a general trend to consider disfluencies as noise in this perspective, in this section we also describe certain works that have found disfluencies informative in SLU (e.g. to predict turn-taking (Saini, 2017), in clinical SLU (Rohanian and Hough, 2021), ...). Thus we distinguish between what is traditionally considered SLU (SLU for SDS), and broader SLU (SLU tasks concerned with social communication). We make this distinction, because the treatment of disfluencies varies greatly – from *removal to integration* – depending on the SLU task. As stated previously, in this thesis we focus on SLU tasks concerned with social communication.

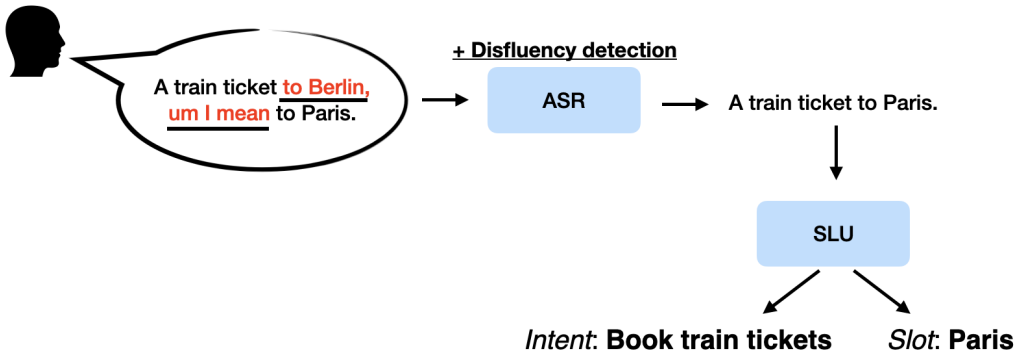


Figure 2.3: Example of SLU for SDS, where the disfluent part of the utterance is highlighted in red and underlined. As shown, the disfluencies are removed in post-processing using disfluency detection systems, after ASR systems transcribe the input speech. Then, the input is further collapsed into a semantic frame, consisting of intent and slot.

In SLU for SDS the objective is to collapse the input utterance into a semantic frame, consisting of an *intent* and a *slot* (Louvan and Magnini,

2020). Disfluencies are in this regard, discarded as noise. In a standard SDS pipeline, the output transcripts of ASR systems are cleaned of disfluencies in post-processing as shown in Figure 2.3, using disfluency detection systems. Thus from a computational perspective, the intent of automatic recognition or processing of disfluencies in present-day is largely due to automatic Natural Language Understanding (NLU) systems not being robust to them (Ginzburg et al., 2014). However, **removing disfluencies from the utterance, could remove important information about the social aspects of communication**.

As discussed previously, there has been an interest in studying the characteristics of disfluencies for recognition purposes, such as their distributional properties in corpora (Shriberg, 1994; Shriberg, 2001; Vasilescu, Rosset, and Adda-Decker, 2010b; Meteer et al., 1995), and analysis at various linguistic levels. Identifying these properties of disfluent speech *enables a comparison with its fluent counterparts*. For e.g., at a phonetic level, research shows that for repetitions (when what was said is exactly repeated), the repeated part of the utterance is found to have similar pitch contours to its original, but just stretched out over time (Shriberg, 1999). Shriberg (1999) found that the vowels in fillers gave much longer durations than the same vowels in fluent contexts, and Shriberg and Lickley (1993) found that the intonation of fillers is not independent of prior prosodic context. These studies were done with the aim of eventually *incorporating linguistic features for ASR to be more robust to disfluencies* (such as in Heeman and Allen (1999), Shriberg, Bates, and Stolcke (1997), and Bear, Dowding, and Shriberg (1992)) with mixed results. Other linguistic levels have also been studied, for e.g. morpho-syntactic characteristics (Goryainova et al., 2014), pragmatics (Shriberg et al., 1998), contextual occurrence (Vasilescu, Rosset, and Adda-Decker, 2010b) and so on.

The issue is important given that voice assistant technology is becoming more sophisticated, and the kinds of input utterances more complex. However, disfluency detection as a task today is predominantly based on perfect text transcriptions, with criticisms that they could not work online nor with transcripts from ASR. This is because ASR systems are not robust to transcribing disfluencies correctly, and even if such is the case, many (text based) disfluency detection systems are still conditioned on appropriate segmentation. From Rohanian and Hough (2021), works "are almost *exclusively* conducted on pre-segmented utterances of the Switchboard (SWBD) corpus of telephone conversations"; contradicting the feasibility of such systems. Now works are exploring the feasibility of detection with ASR transcripts, such as in Rohanian and Hough (2021), or the incorporation of prosodic cues (Zayats et al., 2019a; Dutrey et al., 2014). Thus, systems are still not robust

enough to perform gold-standard disfluency detection on completely spontaneous speech. This may not be the case for all voice assistant technology, for e.g. concerned with task oriented dialogue. The the user may be aware of the limitations of the agent, and may restrict the domain and complexity of the utterance to have one intent per utterance/turn (an assumption that is made by most SOTA dialogue systems (Zhao and Kawahara, 2019a)). The issue of automatic processing of disfluencies also leads to a lack of availability of corpora to study disfluent phenomena. Note, in this thesis we do not further discuss the literature on disfluency detection, as it is not relevant to the thesis. However, **we do discuss the feasibility of our work tested using ASR transcripts where possible** (see Part IV).

There is also research to indicate that disfluencies are filtered out by the human listener. In experiments, Fox Tree (1995) noted that removing repetitions from the utterance digitally did not affect the naturalness of speech. However, this could be explained from a later work (as mentioned previously) from Shriberg (1999), who found that repetitions have similar pitch contours, but just stretched out over time. Lickley, Shillcock, and Bard (1991) found that listener's could not pinpoint the exact interruption point of a disfluency, and tended to point it out later (up to one word) in the flow of speech. Lickley and Bard (1998) studied a listener's ability to identify disfluencies to find that they are not reliably predictable unless a noticeable pause or abandoned word is apparent. However, Bailey and Ferreira (2003) point out that the idea of "filtering" out disfluencies (despite their prosody remaining intact, location not easily remembered by listeners) is implausible given the *incremental* nature of speech processing; i.e. "the processing starts before the input is complete" (Schlangen and Skantze, 2011). Filtering would indicate that processing needs to occur *after* the removal of disfluencies. Furthermore, we subsequently discuss disfluencies from a TTS perspective, where humans evaluate the same synthesised speech *with or without disfluencies*, to find that they do affect the listener's perception in different ways.

From a *broader SLU perspective*, the effect of disfluencies has also been studied in higher dimensions. Often, these tasks are concerned with social aspects of communication. An example of such a task, is emotion detection. Moore, Tian, and Lai (2014) and Tian, Moore, and Lai (2015), found that disfluency features achieve higher accuracy for emotion detection than lexical or acoustic features alone. They conclude that disfluencies could possibly capture high level features in emotion detection that lexical/ acoustic features might omit.

The drawback of works can be their lack of *contextual* analysis. The link

between disfluencies and a wide variety of phenomena is to be expected, with Barr (2001) even describing fillers as *vocal gestures*. It may not always be a case of "paying attention to X is correlated with Y" (Boyd and Schwartz, 2021). In this thesis, we would like to study the *nuances* disfluencies, i.e, with a focus on understanding how disfluencies interact with the rest of the message. Thus, there *are* already existing works in SLU that specifically do study the informativeness of disfluencies in SLU, including advancements in clinical SLU (Rohanian, Hough, and Purver, 2021). We distinguish the work in this thesis from previous works as **we specifically focus on how disfluencies interact holistically with the lexical contents of the message** (not just, for e.g. a cue in turn taking (Saini, 2017), which does not fully explore the relationship between disfluencies and the *contents of* what was said in the turn). For e.g., in Part IV, we study the interaction between the lexical message and fillers; between *what was said* and *how it was said*.

Generation and Text-to-Speech (TTS) Similar to psycholinguistic approaches, it is to be noted that there are works that study the *perception* of disfluencies from a generation standpoint, i.e. using artificially synthesised disfluencies in speech. While this may not be considered part of SLU as such, we briefly describe some works since it is related to computationally studying disfluencies, and from a perspective that disfluencies may affect perception in some way.

In a work directly motivated by psycholinguistic perspectives of comprehension (see Figure 2.1.1), Wollermann et al. (2013) explore the listener's perception of disfluencies using TTS. This was based on theories by Brennan and Williams (1995), which discuss the listener's evaluation of how uncertain they think the speaker is regarding a topic. They had the system exhibit "uncertain" behaviour through disfluent TTS responses in a question-answering context. They found that disfluencies in combination with prosodic cues (e.g. delays + fillers) increased a listener's perception of uncertainty towards the system's answers.

In general, studies may not be so specific to psycholinguistic perspectives. For e.g. works that utilise disfluencies in TTS may be to enhance the *naturalness* of the synthesised speech. Pfeifer and Bickmore (2009) for e.g. evaluate an agent that uses fillers "uh" and "um" in speech. The motivation behind this was to improve the naturalness of speech in an Embodied Conversational Agent (ECA), as ECAs often try to emulate humans in gestures and facial expressions, yet speak in fluent sentences. Results are mixed, with some participants saying that fillers enhanced the naturalness of the conver-

sation, while others expected that an agent should speak fluently, and fillers were deemed inappropriate. However, recently Leviathan and Matias (2018) introduced Google Duplex: an AI system for accomplishing real world tasks over the phone. A key component to the naturalness of the system was *in the incorporation of disfluencies* (such as fillers and repairs) in the TTS responses during human-agent interaction. Székely et al. (2019) discuss approaches for treating fillers in TTS tasks, for e.g. suggesting methods that will result in them being synthesised naturally (both distributionally and perceptually) in the generated output. Disfluencies may also be used as a *communicative strategy*. For e.g., Skantze, Johansson, and Beskow (2015) studied how a system can use multi-modal turn-taking signals (including fillers) as a *time-buying measure*, i.e. to buy time for generating a response as the next move of the robot is decided.

These studies ground the generation of disfluencies in an artificial agent, to find that humans are perceptive to disfluencies; and indeed perceive them as signals even when they are artificially synthesised. While we do not work on the generation of disfluencies in this thesis to observe the effect on perception, we do study whether disfluencies can be a discriminative feature in the *prediction* of a listener's perception (see Part II).

2.2 The Challenges of Annotation of Disfluencies in SLU

In this section, we discuss the challenges of annotation of disfluencies in SLU – considering mainly terminological and transcription issues surrounding disfluencies. In the previous section, we highlighted that within the computational perspective, there are two branches of SLU (SLU for SDS and broader SLU tasks); and that the treatment of disfluencies depends on the task. In this section, we explore a commonly used annotation scheme for each, to highlight the *confusing terminology surrounding disfluencies* in SLU. We then discuss the known complications of transcribing disfluent speech, showing that *obtaining annotations of disfluencies is not a straightforward task*.

Furthermore, it is important to highlight these issues, as **the quality of annotations – and consequently, the validity of the findings – depends on their consideration**. This section aims to shed light on these challenges. The criteria to select a corpus containing disfluencies for SLU is very different from other (non-computational) domains. For example, in SLU, SOTA methodologies are not always readily available to automatically process smaller datasets (so larger datasets are desirable), and it can be dif-

ficult to study phenomena that occur very infrequently, as they do not have generalisable properties. Thus, these challenges make it non-trivial to automatically process disfluencies, and **affect the availability of disfluency annotations in corpora**.

2.2.1 Terminological issues

In the subsequent paragraphs, we describe two annotation schemes that are commonly used in SLU research. It is important to note that these are just two schemes out of many in SLU (and that others exist, e.g. Purver, Hough, and Howes (2018), but are not as common). One scheme is used more for the purposes of detection/recognition of disfluencies, and the other for an easy way to integrate disfluencies for subsequent automatic processing. End-to-end systems such as Serban et al. (2016), may skip the issue of transcription/annotation all-together. The issue of annotation has also been extensively studied from a non-computational, linguistic perspective (Grosman, 2018), with extensive summaries of terminology from different linguistic works (Lickley, Shillcock, and Bard, 1991; Nicholson, 2007), and with new and emergent annotation schemes proposed (Crible et al., 2015).

In general, the task of transcribing and annotating spontaneous speech is not straightforward considering that spontaneous speech is filled with disfluencies, complex sentence structures (e.g. from turn-taking) and non-speech sounds. Non-speech sounds could include paralinguistic sounds[2] the interlocutor makes, such as *laughter* or *coughing*, and other general sounds such as *children playing outside*. Shriberg (1994) additionally shows that even the kind of corpus has an effect on the rate of disfluencies within the corpus. Other studies have found this to be the case, e.g. speakers are more disfluent in dialogues compared to monologues (Oviatt, 1995), in human-human conversations than human-machine conversations (Oviatt, 1995), and disfluencies are affected by dialogue role and domain (Colman and Healey, 2011). Transcribers require guidelines that take into account the characteristics of the data and the purpose of the dataset itself. This in turn may affect the kind of disfluency annotation scheme used.

The disfluency annotated Switchboard (SWBD) Corpus also known as the Penn Treebank3 release of SWBD, takes the original Switchboard-1 Telephone Speech Corpus of American English conversations (Godfrey, Hol-

[2]Paralinguistic sounds may still broadly be considered disfluencies in some works; as they contribute to an *interruption* in the fluent flow of speech produced by the speaker, rather than being produced externally.

liman, and McDaniel, 1992; Godfrey and Holliman, 1993), and adds annotations of part-of-speech (POS) tags, utterance segmentation within turns, and importantly, disfluency annotations (Calhoun et al., 2010). The disfluency annotations were done following guidelines proposed by Meteer et al. (1995).

We focus on this annotation guideline rather than other annotation schemes of disfluencies (e.g. Levelt (1983)) because currently, it is one of the **largest open conversation corpora with disfluency annotations**. It remains **a widely used benchmark for SLU tasks**; disfluency detection (e.g. (Hough, Schlangen, et al., 2015; Shalyminov, Eshghi, and Lemon, 2018; Zayats et al., 2019a)), dialogue act (DA) classification and segmentation (Tran, 2020; Zhao and Kawahara, 2019a), parsing (Tran et al., 2017b), ASR[3] .... There are also other non-standard tasks in SLU that utilise this dataset; for e.g. the specific role of disfluencies to predict turn-taking (Saini, 2017) and coherence modelling for dialogue (Cervone, 2020).

Thus the SWBD corpus contains quality annotations at different linguistic levels, but is large enough that it is often studied computationally. Similar versions of these annotations can be seen in disfluency detection papers (e.g. (Hough, Schlangen, et al., 2015; Dutrey et al., 2014)), or other disfluency focused datasets such as "DUEL: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter" (Hough et al., 2016), or "DISFL-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering" (Gupta et al., 2021).

$$\text{Archie } \underbrace{[\text{ likes}}_{\text{RP}} + \underbrace{\{\text{F uh }\}}_{\text{IM}} \underbrace{\text{loves }]}_{\text{RP}} \text{ Veronica.} \tag{2.1}$$

Levelt (1983) originally proposed the notion of *some kind of erroneous speech that is to be replaced by corrected speech*, which is used as a starting point for annotation (and especially, when considering the task of disfluency detection). This is because, the idea is to only keep the corrected speech in the utterance for further SLU processing (see Figure 2.3), and discard the rest as noise.

[3]Here we are referring to the original SWBD corpus (Godfrey, Holliman, and McDaniel, 1992) and the various layers of annotations that have been added to it (see Calhoun et al. (2010) for a detailed description). For example, the Switchboard Dialogue Act (SWDA) corpus extends the SWBD corpus with annotations of communicative functions, is subsequently used in DA classification.

The HCRC Map task corpus (Anderson et al., 1991; Thompson et al., 1993) of task-oriented dialogues (where speakers must collaborate verbally to reproduce on one participant's map a route printed on the other's) is similarly annotated for a wide variety of linguistic behaviours, and then consequently was used by, "computational linguists for training machine classifiers". However, is not as large as the SWBD corpus (18 hours compared to 260 hours of speech).

Disfluency	Type	Schema	Example
Restarts	Restarts+repair	$[RM + RR]$	
	Restarts+repair+NSE	$[RM + \{\}RR]$	
	Restarts w/o repair	$[RM +]$	
	Complex restarts	*usually go from left to right*	
NSE	Filler	$\{F...\}$	"uh", "um"
	Explicit editing term	$\{E...\}$	"I mean", "sorry"
	Discourse marker	$\{D...\}$	"you know", "well"
	Coordinating conjunction	$\{C...\}$	"and", "and then"
	Aside	$\{A...\}$	*aside in sentence/ restart*

Table 2.1: The annotation schema for disfluencies as given in Meteer et al. (1995) for the Penn Treebank3 release of SWBD. NSE stands for Non-Sentence Elements. Note, other annotations such as *laughter*, *coughing*, and so on have not been included in this table.

To briefly describe the SWBD disfluency annotations, we refer to Eg. 2.1 and Table 2.1. Formally, there is: i) the *reparandum phase* (RM), i.e or the entire region to be deleted, and ii) the *repair phase* (RP), i.e. what replaces the *RM*. This was adopted by Shriberg (1994), who also proposes the term iii) *interregnum phase* (IM) (denoted in the table as "Restarts+repair+NSE"), which is an optional interruption point. Non-sentence elements (such as fillers) can occur within this structure or outside of this structure, sometimes called *isolated edit terms/ single edit tokens*.

Shriberg (1994)'s annotation scheme only spans one speaker turn and can only be initiated by the speaker making the correction. In other Conversation Analysis (CA) works, repair could span different speaker turns and different interlocutors (see a detailed discussion in Purver, Hough, and Howes (2018)), such as clarification requests – "I'm sorry could you repeat that?").

The Linguistic Inquiry and Word Count (LIWC) by Pennebaker, Francis, and Booth (2001) is a text analysis software, where transcriptions can be formatted using the guidelines given in the LIWC documentation. Then, a psychological text analysis of the tokens would be mapped and returned using the software (e.g. tokens "think" and "know" in the transcript are mapped to a category "Insight"). The annotation guidelines proposed

by the LIWC **quickly deal with and integrate disfluencies**, while not spending time on curated transcriptions.

The general idea behind the LIWC, as stated in Boyd and Schwartz (2021), is that subtle, imperceivable differences in a speaker's language use could be analysed using statistical software predominantly based on word frequencies to compute larger characteristics of the speaker, such as personality traits. To quote Pennebaker, Francis, and Booth (2001), though the LIWC (2001) was not originally intended for spoken language, they "found it to be useful in analysing conversations and interviews" and thus outlined some conventions of transcription. It is **often used in guidelines for research that would fall under *broader SLU tasks* or "personality/affective computing"** (e.g. as shown in the work 'Recent trends in deep learning based personality detection" by Mehta et al. (2020)).

Comparing the two guidelines The SWBD guidelines take into consideration linguistic conventions, and were intended for the specific purpose of studying speech at different linguistic levels (from lower to higher): syntax to discourse and for computational purposes (e.g. ASR). Additionally, the disfluency annotated SWBD guidelines outline in detail how to annotate disfluencies, for the purposes of recognition. On the other hand, the LIWC's main purpose is to do psychological text analysis for higher dimensions of personality/affective computing. The broader guidelines in the LIWC simply define methods to transcribe spontaneous speech, which includes conventions to annotate disfluencies; which may be inaccurate, but require relatively less resources (e.g. transcriber experience).

While disfluencies are not the *central focus* of the LIWC, we believe it is important to highlight some characteristics of such guidelines. Other works concerning disfluencies may highlight the *depth* of work on the topic disfluencies and and various annotation schemes. However, **so far there has been little consideration for the *breadth*; there are *general purpose guidelines* used to transcribe/annotate spontaneous speech**; where the data by nature, *will* include disfluencies. There is not necessarily the awareness of every nuance of disfluencies in every SLU task, but there are invariably disfluencies that will occur in the spontaneous speech dataset utilised. The LIWC is just one example (of many,) where the disfluency guidelines are not detailed, but is still is a very much used resource to automatically parse spontaneous speech datasets. Thus discussing the LIWC gives the opportunity to point out **the frequent terminological issues in the field arising from general (and not very detailed) disfluency guidelines**.

One example of this, is where the LIWC recommends the transcriber to utilise a simple, but not as accurate transcription of disfluencies. They recommend to the transcriber to insert as many fillers as deemed appropriate in place of a more complicated disfluency structure (roughly, the $RM+\{IM\}$ phase), and then to keep the RP. Thus, one may think that the dataset selected for an SLU task has annotations of some disfluencies. But in fact, **these annotations could be inaccurate given the guidelines followed**, and is important to keep in mind if the objective of the work is to specifically study disfluencies.

> *"Hm", "hmm", "uh", "uhh", "uhm", "um", "umm", and "er" are part of the nonfluency dictionary. . . . Stuttering can be accommodated by altering the stuttering part of a phrase to a nonfluency marker. For example, "`The, the bo-`, the boat went into the water" could be changed to "`Uh`, the boat went into the water." The transcriber will have to decide how many uh's would be appropriate.*
>
> *Pennebaker, Francis, and Booth (2001)* (1)

This guideline is shown in quote 1 (for "nonfluencies" – "uh", "um", "er" and so on), with instructions for what they call "stuttering". Here, "stuttering" means some form of restarts, or to broadly denote the general phenomena of being disfluent. Though the LIWC is a psychological text analysis software, the term is not to be confused with the clinical sense of "stuttering" (in clinical literature, the term commonly used is "stutter-like disfluencies" (SLDs)). Rather, the term seems borrowed from Mahl (1956), whose interest was in speech of schizophrenic patients; and used the term "speech disturbances" to describe different kinds of disfluencies, with "stuttering"[4] in particular to refer to repetition of partial words (Lickley, 2015).

Thus the general purpose guidelines for spontaneous speech proposed by the LIWC compared to the specific disfluency guidelines proposed in SWBD leads to disfluency annotations that are i) faster to transcribe/annotate ii) sacrifices on accuracy given that the transcriber is free to substitute a complex disfluency for a single edit token iii) can be utilised by less experienced transcribers, and iv) thus can be utilised in the case of subpar acoustic data (in case of noisy environments, when it is very unclear what was said by the speaker).

[4]This term is also observed in personality computing, including in the Persuasive Opinion Mining (POM) (Park et al., 2014) dataset studied in this thesis, that annotates "stutters" in the dataset (please refer to Chapter 3 for a detailed description of the datasets), and indeed cites Mahl (1956) and overall, uses the broad term "speech disturbances".

Same subject, many lenses: confusing terminology around disfluencies. In this thesis, we study the the tokens "uh", "um" and "oh", which are conventionally called fillers (Clark and Fox Tree, 2002) and discourse markers (Schiffrin, 1987) respectively. Narrowing the two annotation schemes to compare just fillers and discourse markers, the problems of terminological issues and tokens in each category become apparent (please refer to Table 2.2 for a summary of these differences).

As shown in the table, works on fillers (such as Clark and Fox Tree (2002)) contain tokens ("uh", "um", "er", ...) that would be categorised fillers in the SWBD Corpus, but would be categorised as "nonfluencies" by the LIWC. Additionally, the disfluency annotated SWBD guideline proposes a category of "discourse markers" which contains tokens such as "you know", "anyway" and "like". This category confusingly has overlapping tokens with "Filler Words" in the LIWC. While we study the token "oh" more inspired by work in Schiffrin (1987) who considers it a discourse marker, it is not considered a discourse marker in the SWBD guidelines. Furthermore, discourse markers are not always considered disfluencies in all works, and the relationship between discourse markers and disfluencies is complex (Crible, 2018). This is also why, we have frequently used the term of "tokens" to discuss the disfluencies studied this thesis, rather than the conventional categorisation of these disfluencies (fillers, discourse markers, ...). Further details about their selection can be found in Sec. 2.3.

Thus, in this section, we introduced the general idea of disfluency structure for disfluency detection that was originally proposed by Levelt (1983); i.e. some kind of erroneous speech (RM) that is replaced by corrected speech (RP). Then we highlighted the confusions in terminology, by discussing two commonly used guidelines (for very different sets of tasks) in SLU that greatly differ in their consideration of disfluencies.

2.2.2 Transcription issues

General errors in transcription Despite having detailed annotation structures for disfluencies, problems may arise when transcribing the (disfluent) audio itself. Zayats et al. (2019b) highlighted the problems that arise when transcribing spontaneous speech automatically. They did a comparison of the words that appeared in the original Switchboard-1 Telephone Speech Corpus (SWBD) (Godfrey, Holliman, and McDaniel, 1992; Godfrey and Holliman, 1993), but not the cleaned up SWBD Mississippi (MS)-State transcripts

Token(s)	Scheme	Category
"uh", "um",	LIWC	Nonfluencies
	SWBD	Fillers
"uhh", "umm"	LIWC	Nonfluencies
	SWBD	*No distinction between "uh" and "uhh"*
"you know"	LIWC	Filler Word (*youknow*)
	SWBD	Discourse Marker
"oh yeah"	LIWC	Nonfluencies(?) – *depends on the transcriber*
	SWBD	Fillers – *only non-isolated "oh"*
The, the bo-, the boat	LIWC	Stuttering (*Uh, the boat*)
	SWBD	Restart (*The[, the bo-, the + boat]*)

Table 2.2: Examples of some differences in terminology between the LIWC and the disfluency annotated SWBD dataset. The round brackets in the category column depict the way the tokens would be annotated given the respective guidelines. As shown, the LIWC recommends that "stutters" (see Table 2.2.1) events be converted to a token in the nonfluency dictionary so that the text analysis software can process them as such, like "um".

(Deshmukh et al., 1998)[5]. The manually corrected MS-State transcripts contained considerations such as:

1. Corrections of word alignment timings (Zayats et al., 2019b).
2. An indication of the transcription error and type of error (an inserted, deleted and substituted word) in the original transcript.
3. Asking the annotators to remain more faithful to the variations in pronunciations of words compared to the previously standardised version (e.g. "gonna" instead of "going to").
4. Asking the annotators to transcribe word fragments more completely even if they were unsure of what the completed word was (e.g. "w[ent]-" instead of "w-").

Additionally there were corrections in segmentation with the aim to preserve prosodic continuity and linguistic context; as according to Deshmukh

[5]In 1997, the The Institute for Signal and Information Processing conducted a cleanup campaign of the SWBD transcripts (Deshmukh et al., 1998) (referred to as the MS-State transcript), due to the original transcripts containing errors; which is expected from human transcribers (Zayats et al., 2019b).

et al. (1998), the original transcripts were split at "counter-intuitive" points that did not follow the lengthy pauses or natural boundaries present in the conversations. Segmentation was thus done at locations where there were *clear* silences separating the speech, i.e. remaining faithful to the speaker's "train-of-thought".

In Zayats et al. (2019b), they refer to these words that appear in the original SWBD transcripts but not the MS-State transcripts as "hallucinations" (or inserted words), and words that appeared in the MS-State transcripts but not the original SWBD are referred to as "missed" (deleted) words. They divided words into categories such as "lexical", "functional" and so on. Words in the "other" category are words characteristic of spontaneous speech, such as fillers, backchannels and interjections. They found that words that fall into the "other" category have a high frequency of transcription errors (both hallucinated and missed), which they **discuss could be due to many words in these categories not being standard**.

Transcription errors of fillers Le Grezause (2017) further investigates the effect that these transcription errors have on the two fillers "uh" and "um" to find that the types of transcription errors of the two fillers are not proportionate. They found that "uh" is the most likely token to be both deleted and inserted, whereas deleted and inserted "um"s are less frequent. Substitutions are often from monosyllabic words in the "functional" ("and", "a") or the "other" ("yeah", "huh") category. The filler "um" is substituted the most frequently with "uh" compared to another word, and transcribers tend to misperceive "um" as "uh" with a much higher frequency than the other way around. Notably, the filler "um" was found to be substituted more in conversations that were rated difficult to transcribe. Given that "um" is inserted/deleted less frequently than "uh", and "uh" substituted for "um" more than the other way around, the authors conclude that "um" carries a higher informational load, and thus a more important role in the discourse.

Le Grezause (2017) hypothesised that the more natural sounding a conversation is, the less transcriber's would pay attention to disfluencies. An interesting finding is that in conversations that were rated less natural sounding, the transcribers missed "uh" less frequently, and in conversations that were rated more natural sounding, the transcribers substituted "um" more frequently. It is also plausible when considering fillers and the listener's perception (here, transcribers) that the speaker's may have used fillers in atypical locations (i.e. locations not expected by the transcriber), causing the transcriber to notice their use more closely (taking from the findings of Tottie (2014)). It was ultimately found that transcribers who had transcribed

a few conversations, tended to have substantially higher transcription errors uniformly for all categories, compared to transcribers that had transcribed a large number of conversations.

2.3 Disfluencies studied in the thesis

In this thesis, we predominantly focus on fillers "uh" and "um", based on psycholinguistic work. Firstly, while many psycholinguistic works use the generic term "disfluencies", it is notable that most of the studies we discussed, particularly related to listener comprehension, were in fact based on fillers (such as Fraundorf and Watson (2011), Swerts (1998), Vasilescu, Rosset, and Adda-Decker (2010a), Brennan and Williams (1995), Barr and Seyfeddinipur (2010), Bailey and Ferreira (2003), Corley, MacGregor, and Donaldson (2007), Arnold et al. (2004), Arnold, Kam, and Tanenhaus (2007), Corley and Stewart (2008), and Bell et al. (2003)). What interests us in particular, are that these works specifically study how listeners use **fillers as cues to understand and resolve incoming information in the speech stream**. There are also computational considerations in choosing these disfluencies, which are subsequently discussed.

Note, that fillers may be also called "filled pauses' '(Maclay and Osgood, 1959) and "hesitations". Clark and Fox Tree (2002) introduced the term "fillers", to distinguish them from "filled pauses"; as term "filled pauses" was previously used to contrast with the term "silent pauses". "Filled pauses" thus seem indicate that there is a pause in speech *filled* by some sort of (meaningless) sound. Throughout this thesis, we utilise the term "fillers" for consistency, including when citing previous research that may use the term "filled pauses".

Disfluencies with no semantic content Vasilescu, Rosset, and Adda-Decker (2010b) give a generalised definition for distinguishing these disfluencies (particularly applied to fillers) which we adopt in this thesis; i.e. there are a category of disfluencies can be broadly classified into common conversational/speech events that do not contribute directly to the final message when considering purely the lexical level. For our purposes, there are disfluencies which contribute to the *semantic* sense of the message, and disfluencies which do not change the semantic sense of the message, but certainly could affect other levels of language, such as the *pragmatic* sense. In Figure 2.2 **B** in essence says "yes" to lending the book, keeping the same semantic sense regardless of presence or absence of filler. However, the way **B** said this ("... {*Fum*} ... all right") *implicitly* indicates some uncertainty or hesitation,

changing the pragmatic sense of the utterance. In SLU for SDS, preserving disfluencies that add to the *semantic* sense of the message are of interest, but not non-semantic disfluencies.

> *Filled pauses have unrestricted distribution and* ***no semantic content****. A few common examples are "uh", "um" and "huh". There are also rarer filled pauses, such as "eh", "oops" etc . . .*
>
> *Meteer et al. (1995)* (2)

The SWBD disfluency annotation guidelines for fillers are in agreement with this definition (see quote 2), i.e. that they contain little to no semantic content and thus do not have *explicit* meaning. These disfluencies, can **occur with high frequency in spontaneous speech datasets** compared to more complex disfluencies that could contain lexical information (such as "Restarts" in Table 2.1).

Overlapping functions, different tokens In Vasilescu, Rosset, and Adda-Decker (2010a), a further distinction is made when considering these non-lexical disfluencies; that is i) truly "disfluent events", which can include fillers, and ii) discourse markers, which act as as a set of linguistic expressions that bracket units of talk (Schiffrin, 1987). As stated in Vasilescu, Rosset, and Adda-Decker (2010a), "classifying such elements is not straightforward as inter-class boundaries are permissive and taxonomy is context and/or corpus dependent". Fox Tree and Schrock (1999) discuss that spontaneous speech differs from scripted speech *because* of discourse markers such as "well", "like", and like fillers, have *several* functions, such as "helping listeners recover from repair, follow a speaker's train of thought . . .". Schiffrin (1987) also shows that discourse markers serve to mark discourse structure of a speaker's speech stream, thus separating one utterance from the previous, and aiding the listener in comprehension. It is also hard to distinguish fillers here, as they also *function as discourse markers* (Swerts, 1998), such as when the filler is used sentence-initially (filling some criteria in Schiffrin (1987)). There are many works that try to formalise discourse markers and the confusing terminology around them (for e.g. Crible (2018), Degand and Simon (2009), and Degand, Cornillie, and Pietrandrea (2013)). In this thesis, we described confusing taxonomies to emphasise that we focus on the non-lexical *tokens* themselves; specifically "uh", "um" and "oh". We colloquially refer to "uh" and "um" as *fillers*, as we base our work on literature that describes these markers as fillers. We refer to "oh", an *information management marker*, based on work that studies "oh" as such. We focus on fillers and the informa-

tion management "oh", because of their role (as described) in helping resolve incoming information in the speech stream[6].

Positional roles Segmentation of utterances can have a big impact on the performance of SLU systems. From quote 2, fillers can have unrestricted distribution, and thus could be placed at any point in the sentence. However, when a filler ends a sentence the reader gets the impression of an incomplete utterance ("And wh-who am I *{F uh}* ... (*incomplete utterance?*)"). This then traverses to annotation guidelines, for e.g. in Meteer et al. (1995), i.e. "... when a filler is separated from the end of a unit by a comma, consider the filler part of the previous unit. If it is separated by a period however, consider it part of the following unit ...", and in the POM dataset (Park et al., 2014), there is the practice of annotating a filler as sentence-initially, if it occurs in between sentences. This annotation guideline may be applicable to some tokens in discourse markers ("well", "like") and not applicable to some ("you know", depending on prosody – "you know?") – but still, all bracketing units of talk and aiding in segmentation of speech. It is to be noted that for "oh", Fox Tree and Schrock (1999) give examples of how the token *may not* be used in some environments (e.g. in an idiom; "John kicked ... oh ... the bucket"). Thus overall, fillers and discourse markers can aid in the segmentation of utterances based on the annotation guidelines – i.e organising the units of speech. In this thesis we study the different functions of fillers, including when serve to act as discourse markers (by broad definitions); studying their positional functions.

Frequency While fillers contain no semantic content, they occur with high frequency in speech datasets, compared to often sparse and specific structures of disfluencies. Aijmer (1987) in a corpus study found "oh" to be one of the most frequent tokens in the corpus. Shriberg (2001) shows that the number of fillers per word across corpora exceed other kinds of disfluencies. For example, $\approx 2\%$ of the Switchboard (SWBD) (Godfrey, Holliman, and McDaniel, 1992) dataset consists of fillers. In the POM dataset of monologues, $\approx 4\%$ of the tokens consist of fillers. Evidence for the sparsity of other disfluencies can be seen in Shalyminov, Eshghi, and Lemon (2018) and Moore, Tian, and Lai (2014), which makes them harder to model. Tottie (2014) also finds that "uh" and "um" are some of the most frequently occurring tokens in both British and American English, which is relevant to the datasets studied in this work. They are so ubiquitous that open source ASR systems now have functions to transcribe them. For e.g. CMU Sphinx speech recogniser includes fillers

[6]We mainly focus on the *tokens* "uh" and "um", based on psycholinguistic perspectives.

in their lexicon, and also has a boolean method `isFiller()` to determine is the unit of speech is a filler, and Google APIs such as YouTube transcription services now include fillers as part of the automatic "closed captioning" (transcription) system.

They occur in an intersection between speech and text Fillers are a common property of spontaneous speech and are shown to have *distinct acoustic characteristics/ paralinguistic properties* (Shriberg, 1999; Vasilescu, Rosset, and Adda-Decker, 2010b), yet they can also be transcribed in text, without more complex disfluency annotations. They can also occur in scripted speech, such as movie dialogues or books, for example to show a particular thought processes of a character (typically hesitation e.g. "uh"), illustrated by the following example:

> **Gilderoy Lockhart:** Hello. Who are you?
> **Ron:** [Taken aback] *{F Um}*... Ron Weasley.
> **Gilderoy Lockhart:** Really! And, *{F uh}*, wh-who am I?
> **Ron:** Lockhart's memory charm backfired! He hasn't got a clue who he is!
> — *from Harry Potter and the Chamber of Secrets (movie)*

This is meaningful for the present work, as it was observed (both by work done in this thesis and other researchers) that deep contextualised text representations of fillers do exist (Tran et al., 2019a), despite being trained on written text (Wikipedia, BooksCorpus (Zhu et al., 2015), Word Benchmark (Chelba et al., 2014)).

Conclusion

Thus in this chapter, we narrow down the literature to identify two relevant issues for disfluency research in SLU, expanding on our Research Objectives as given in Chapter 1. We described the relevant computational approaches to disfluencies distinguished from psycholinguistic approaches in Sec. 2.1, and then described the challenges of annotation of disfluencies in SLU in Sec. 2.2. Given these two issues, in Sec. 2.3 we expanded our reasons for choosing to **predominantly focus on "uh" and "um" and "oh" in this thesis**. In the next chapter, we describe in detail the main datasets that are used (see Chapter 3). Note, throughout the thesis, we briefly describe the background relevant to that specific chapter, summarising from this chapter. Sometimes we include some additional *related works* that we have not discussed in this

chapter, that are not specific to disfluencies, but may be relevant to the models/theoretical frameworks specific to that chapter.

3 | Datasets

In this thesis, we use two main datasets; a dataset of *monologues* called the Persuasive Opinion Mining dataset (POM) (Park et al., 2014), and a dataset of task-oriented *dialogues* called the JUSThink dataset (Nasir et al., 2020a; Nasir et al., 2021). The POM dataset loosely follows the annotation guidelines from the LIWC (an adaptation of Mahl (1956)'s terminology), though it is not explicitly referred to in the original work. Please refer to Sec. 3.1 for further details about this dataset. For the dataset of dialogues, we transcribe a subset of the JUSThink dataset (which we title the "JUSThink Dialogue and Actions Corpus"), keeping in mind the issues with the transcriptions of disfluencies. Please refer to Sec. 3.2 for further details about the JUSThink dataset and the transcription campaign.

3.1 The Persuasive Opinion Mining Dataset

In this thesis, we study the role of fillers in a dataset of monologues by studying:

1. **At an utterance level:** the *interaction between fillers and the primary signal*; i.e. the lexical, or what was said in essence.

2. **At a discourse level:** How the *speaker's production of disfluencies could affect to the listener's perception* in monologues. Specifically, we use macro-level disfluency features (production contexts) and listener annotated labels of the perception of the speaker's i) certainty or commitment to their utterance/review (using the label of confidence) and ii) stance (using the label of sentiment).

For this, we choose the POM dataset (Park et al., 2014), a dataset of 1000 (American) English monologue movie review videos[1]. Please refer to Chap-

[1]This dataset is an open-source dataset, freely available at `https://github.com/A2Zadeh/CMU-MultimodalSDK`.

ter 4, Chapter 5 and Chapter 6 for our experimental findings using this dataset.

Scenario Speakers recorded themselves (video and audio) giving a movie review, which they rated from 1 star (most negative) to 5 stars (most positive). The movie review videos are freely available on ExpoTV.com, and are completely in the wild; speakers were simply reviewing a movie without the knowledge that their review would eventually be annotated for such a context. For the data collection, only movies that were rated 1-2 stars, or 5 stars were selected for annotation. The reasoning behind this was to capture more persuasive reviews, as well as study the effect of strong sentiment on persuasion; i.e. speaker's that liked a movie enough to rate the movie 5 stars would likely persuade the (unseen) listener to watch the movie. Speaker's recorded themselves in their own location of choosing using their own recording equipment, leading to natural and voluntary reviews, but compromising on the audio/video quality. Park et al. (2014) utilise Amazon Mechanical Turk (AMT) (Mason and Suri, 2012), a popular online crowd-sourcing platform for the transcription of the videos, and annotation of high level attributes at the dialogue level.

Discourse level: Measuring high level attributes of the speaker

50 English speaking workers from AMT based in the United States were selected to annotate the reviews for high level attributes, such as "persuasiveness", "confidence" and so on. There were three annotators per video. According to the original paper (Park et al., 2014), "To minimise gender influence, the task was distributed such that the workers only evaluated the speakers of the same gender". For confidence, annotators were asked after watching the entire review "How confident was the reviewer", and had to rate the speaker on a Likert scale of 1-7 with given labels: 1 (not confident), 3 (a little confident), 5 (confident) and 7 (very confident). For stance, the annotators were asked "How would you rate the sentiment expressed by the reviewer towards this movie?", and were asked to give a label from 1 (strongly negative) to 7 (strongly positive).

Transcription of audio and annotation of disfluencies

Dataset Description Park et al. (2014), thus collect 500 positive movie reviews (5 star rating) consisting of 315 males and 185 females, and 500

negative movie reviews; (1-2 star rating, due to a lack of 1 star rated reviews), of which 216 are 1 star reviews consisting of 151 males and 65 females and 284 are 2 star reviews consisting of 212 males and 72 females. We utilise from this dataset:

1. **The transcript files**: For the transcription of the dataset, 18 English speaking workers from AMT based in the United States transcribed the videos. The transcripts were then reviewed and edited by "in-house" experienced transcribers for accuracy. This was taken into consideration when choosing this dataset, as we discuss in Chapter 2, Zayats et al. (2019b) and Le Grezause (2017) that found that transcriber experience matters in the annotation disfluencies. From Park et al. (2014)", they state that they "obtained verbatim transcriptions, including pause-fillers and stutters". These specific disfluencies are discussed in item 3.1.

2. **The metadata files**: The meta-data files are *.csv* files consisting of each individual score the annotator gave for each attribute the review, and meta-data about the movie, such as title, actors, directors and so on.

Annotation of disfluencies The descriptive paper about the the POM dataset refers to Mahl (1956), and uses the same term "speech disturbances" to refer to both fillers (which they call "pause-fillers") and restarts (annotated as "stutters").

While Park et al. (2014) do not explicitly cite the LIWC, it appears that some of those guidelines are adapted in the transcription of the dataset and annotation of disfluencies. In this section, we illustrate which disfluencies are annotated, and where they deviate from the guidelines given in the LIWC.

Firstly, fillers are accurately transcribed and are not used in place of the $RM + \{IM\}$ phase, which is suggested by the LIWC (see quote 1 in Chapter 2). For this phase, instead, they mark "stutters" in the POM dataset; with a similar *sense* of stuttering described in quote 1, i.e. where the transcriber instead of using a complex repair structure, instead uses a simpler one. To reiterate, "stutter" markings do not denote the clinical sense of stuttering, merely some form of repair (see Table 2.1, Chapter 2), and is taken from Mahl (1956)'s terminology. The same sentence "The, the bo-, the boat" from the quote may be transcribed as "The, the (stutter), the boat" without extra fillers inserted as suggested by the LIWC ("Uh, the boat"). Fillers are annotated as shown in Eg. 3.1 (with spelling "umm" and "uhh"), along with other disfluencies i.e. "stutters", (though each stutter can mean a different

Hi there, today we're going to be reviewing the dvd of gladiator which is a uh FILLER big Russell Crowe film from uh FILLER late nineteen-nineties . um FILLER it won uh FILLER academy awards and it was quite a popular movie. um FILLER it tells the story of the gladiator who is played by Russell Crowe and his attempts sort of to gain freedom for himself and resist um FILLER the emperor at the time. um FILLER it's a really good movie. It's long uh FILLER that's a primary complaint against it it's over two two and a half hours so you need to have some time to sit down and watch gladiator but russell crowe does a really good job and he's really believable in the role. um FILLER and it's really thoroughly entertaining from start to finish. um FILLER it's shot in a very cinematic style I guess. I guess all movies can be shot in cinematic style but the photography in this one seemed in particular excellent to me um FILLER and it really was just i thought a fantastic movie. uh FILLER you can the dvd pretty cheap now it shouldn't be too expensive it's been out uh FILLER for several years. But if you haven't seen it and you somehow missed it the first time around it really is a great story um FILLER to check out. Definitely not for the kids it's an adult movie but um FILLER yeah. If you fit in the age group and haven't seen gladiator go get it. Five stars out of five. Thanks.

Figure 3.1: Example transcript from the POM dataset

kind of restart). Stutters can also indicate tokens that the annotators were unable to catch. In Figure 3.1, we give an example of a transcript from the dataset (with the fillers spelt as "uh" and "um", consistent in this thesis).

The (uhh) the movie (umm) did not live up to the book.
– POM dataset
[The {F uh } + the] movie {F um } did not live up to the book.
– Disf. SWBD
(3.1)

Additionally, the POM dataset contains "xxxx" entries that are described in quote 3 from the LIWC. These "xxxx" entries likely exist due to poor quality acoustic data, as the speakers used their own (non-standardised) recording equipment to record their movie reviews. Therefore, other disfluencies likely exist in the (acoustic) dataset, but they are either not not completely transcribed ("xxxx" in place) due to poor quality audio, or annotated fully (with "stutters" used in place). Thus it is not feasible to study other specific types disfluencies in the POM dataset.

> *LIWC (2001) is designed only for spoken language. Transcribers often insert remarks, such as `[subject laughs]`, `[shaky voice]`, `[whispers]`. We recommend removing these. Occasionally, the transcriber cannot understand a word or passage. Rather than writing `[can't understand word]` or `[?]`, the transcriber should put a nonsense word, such as "xxxx" in its place. LIWC (2001) will count the xxxx as a spoken word but not assign it to a dictionary.* (3)
>
> *Pennebaker, Francis, and Booth (2001)*

Selecting the POM dataset

We think this dataset is particularly relevant for the following reasons:

- **Since this is a dataset of monologues, it allows us to focus uniquely on the role of fillers (Swerts, 1998)**. This is because the speaker is conscious of an *unseen* listener, but is not interrupted by the listener with other dialogue related disfluencies, such as backchannels ("Uh-huh"). This also minimises some turn-taking properties of fillers, such as when they are used by the speaker to hold the speaker turn. Additionally, the annotators were never asked specifically to pay attention to the speaker's use of fillers.

- **Filler annotations of "uh" and "um" have been manually and accurately transcribed**. Each transcription of a movie review video was reviewed by experienced transcribers for accuracy after being transcribed via Amazon Mechanical Turk (AMT) (Park et al., 2014). The experience of the transcriber is important, as Zayats et al. (2019b) shows that transcribers tend to misperceive disfluencies and indeed, this can affect the transcription of fillers (Le Grezause, 2017). The filler count of this dataset is high (roughly 4% of the transcriptions, for comparison, the Switchboard (Godfrey, Holliman, and McDaniel, 1992) dataset of human-human dialogues, consists of $\approx$ 1.6% of fillers (Shriberg, 2001)). Sentence markers have been manually transcribed, with the practice of the filler being annotated sentence-initially, if the filler occurs between sentences (in this dataset, utterance segmentation is not available, and is interchangeable with sentence).

- **The inter-annotator agreement for high level attributes is high**; with confidence (label we use to denote the listener's perception of the speaker's metacognitive state) of Krippendorff's $\alpha = 0.73$ (Park et al., 2014). Additional details can be found in the original (Park et al., 2014). We do an extensive analysis of fillers and the dataset in Chapter 4, so we do not give additional details in this chapter.

3.2 The JUSThink Dialogue and Actions Corpus

In this thesis, we study the role of fillers and "oh" in a task oriented dialogue by studying:

1. **At an utterance level:** The relationship between disfluencies and local alignment contexts[2] specifically (a) *verbal* and (b) *behavioural* alignment of the interlocutors (children) taking part in the task and

2. **At a dialogue level:** The relationship between disfluencies and the (a) *performance* and (b) *learning outcomes* of the interlocutors as a dialogue level measure of task success.

For this, we utilise the JUSThink dataset (Nasir et al., 2020a; Nasir et al., 2020b), where we selected 10 representative dialogues to transcribe and utilise for specific purpose of studying alignment in spontaneous speech (called the

[2]or the "development of similar representations" (Pickering and Garrod, 2006)

"JUSThink Dialogue and Actions Corpus"). Please refer to Chapter 7 and Chapter 8 for our experimental findings using this dataset. We contribute this dataset of anonymised, transcribed children dialogues, event logs of their task progression, and code – which are all made publicly available to either reproduce our results or to use for further research.

The JUSThink dataset

JUSThink is a collaborative problem solving activity for school children, collected by researchers at the École Polytechnique Fédérale de Lausanne (Nasir et al., 2020a; Nasir et al., 2020b). It aimed to improve children's Computational Thinking (CT) skills by exercising their abstract reasoning on graphs. Recent research on educational curricula stresses the need for learning CT skills in schools, as going beyond simple digital literacy to developing these CT skills becomes crucial (Menon et al., 2019). With this in mind, the objective of the activity was to expose school children to minimum-spanning-tree problems.

Scenario A humanoid robot, acting as the CEO of a gold mining company, presents the activity to the children as a game, asking them to help it collect gold, by connecting gold mines one another with railway tracks. They are told to spend as little money as possible to build the tracks, which change in cost according to how they connect the gold mines. The goal of the activity is to find a solution that minimises the overall cost, i.e. an optimal solution for the given network[3].

Children participate in teams of two to collaboratively construct a solution, by drawing and erasing tracks. Once all gold mines are reachable, i.e. in some way connected to each other, they can submit their solution to the robot for evaluation. They must submit their solution together, and can submit as many times as they want. The robot then reveals whether their solution is an optimal solution or, if not, how far it is from an optimal solution (in terms of its cost). In the latter case, children are also encouraged by the robot to try again. They can submit a solution as many times as they want until the allotted time for the activity is over. In this thesis, we treat this triadic activity as a dyadic dialogue. The children are initially prompted

[3]In the network used in the JUSThink activity, there exist 10 nodes: { 'Luzern', 'Interlaken', 'Montreux', 'Davos', 'Zermatt', 'Neuchatel', 'Gallen', 'Bern', 'Zurich', 'Basel' } and 20 edges. A description of the network, with the node labels (e.g. "Mount Luzern"), x, y position of a node, possible edges between the nodes, and edge costs, is publicly available online with the dataset (JUSThink Dialogue and Actions Corpus), from the Zenodo Repository DOI: `http://doi.org/10.5281/zenodo.462710410.5281/zenodo.4627104`.

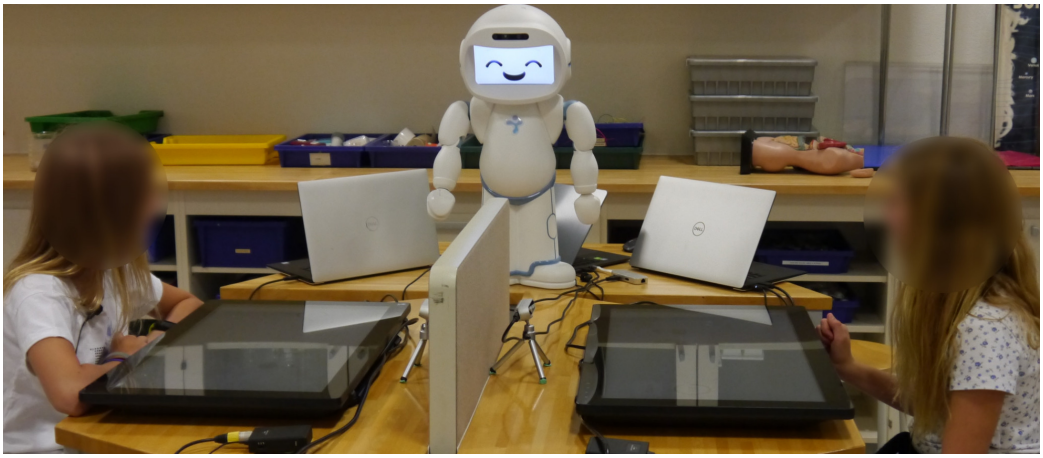
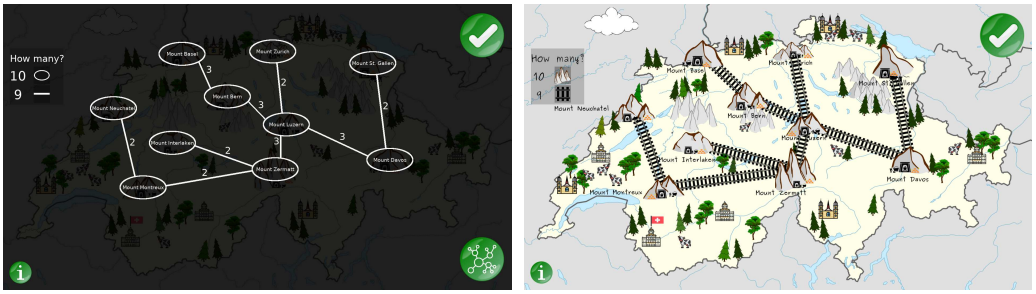


Figure 3.2: The JUSThink activity setup

by the robot to work with each other, and later simply given the cost of their sub-optimal solution. However, almost all of the exchanges are between the two interlocutors. After careful observation of the dialogues in the dataset, we observe the tendency to ignore the robot unless submitting a solution.

Setup Two children sit across each other, separated by a barrier. A touch screen is placed horizontally in front of each child. Children can see each other, but cannot see the other's screen, as shown in Figure 3.2. They are encouraged by the robot to verbally interact with each other, and work together to construct a solution to the activity.

The screens display two different views of the current solution to the children. One view is an *abstract view*, where the gold mines are represented as nodes, and the railway tracks that connect them as edges (see 3.3a). The other view, or the *visual view*, represents the gold mines and railway tracks with images (see 3.3b). A child in the abstract view can see the cost of built edges, but cannot act upon the network. Conversely, in the visual view, a child can add or delete an edge, which is a railway track, but cannot see its cost. The views of the children are swapped every two *edit actions*, which is any addition or deletion of an edge. Hence, after every two edit actions, the child that was in the abstract view moves to the visual view and vice versa. A turn is thus the time interval between two view swaps, i.e. in which one child is in the visual view and the other child is in the abstract view. A turn lasts for two edit actions. This design aims at encouraging children (interlocutors) to collaborate.



(a) An example abstract view. (b) Corresponding visual view.

Figure 3.3: interlocutors' views during the JUSThink activity

Dialogue level: Measuring Success in the Task

In this Ssec., we briefly describe the scores of task success, i.e performance and learning, taken from the original dataset (Nasir et al., 2020a; Nasir et al., 2020b).

Measuring task performance. We adopt a measure that is based on the costs of the submitted solutions compared to the cost of an optimal solution (which is always the same). For each proposed solution, we calculate the normalised cost as the difference between the cost of the solution and the cost of an optimal solution, normalised by the cost of the optimal solution. We define *error* as the smallest normalised cost value among all submitted solutions, which represents the team's closest solution to an optimal solution. We use *error* to measure task performance[4].

Measuring learning outcomes. Learning measures commonly build upon the difference between the post-test and pre-test results, e.g. in (Sangin et al., 2011); which indicates how much an interlocutor's knowledge on the subject has changed due to the activity. We measure the learning outcomes on the basis of the relative learning gain ($learn_P$) of an interlocutor P, which essentially is the difference between pre-test and post-test, normalised by the margin of improvement or decline (Sangin et al., 2011). We use *learn*, the average relative learning gain of both interlocutors, to measure a team's

[4]We process logs (folder *logs/* in the dataset at DOI: 10.5281/zenodo.4627104) with a script (*tools/1_extract_performance_and_other_features_from_logs.ipynb* in the tools at DOI: 10.5281/zenodo.4675070) to compute the *error* for each team (in table *processed_data/log_features/justhink19_log_features_task_level.csv*, available with the tools).

learning outcomes[5].

Utterance level: Alignment context for creating the JUSThink Dialogue and Actions Corpus

In this thesis, we select this dataset in order to study at an utterance level the relationship between disfluencies and alignment. Firstly, the activity is particularly suited to study alignment, since it is designed in such a way to create interdependence, i.e. a mutual reliance to further the task, between the interlocutors. This interdependence requires interlocutors to align with each other on multiple levels, e.g. how to refer to the environment and how to represent the activity, in order to succeed. Concretely, the activity enables:

- **Swapping and visual view control.** Since a turn changes every two edit actions, if an interlocutor has a particular action they want to take, they have to either wait for their turn in the visual view to implement the desired change, or instruct the other interlocutor. Here, we utilise an idealised perspective on the activity: at a given time, the interlocutor in the abstract view is an Instruction Giver (IG) who describes their instructions for the task by using specific expressions to refer to the task (referring expressions), and the other (in visual view) is the Instruction Follower (IF) who executes the action (this is akin to the Map Task (Anderson et al., 1991))[6]. The activity design creates a frequent swapping of views. This aims to discourage interlocutors from working in isolation or in fixed roles of IG and IF, which could potentially happen in other collaborative tasks.

- **Routine expressions and alignment in the task.** Since the robot uses brief and general instructions to present the activity and its goal, the interlocutors must figure out for themselves the way to approach the activity. The expressions specific to the task, which are the names of the gold mines (cities in Switzerland, a multi-lingual country), are potentially unfamiliar to interlocutors. Interlocutors must refer to the task, and then align with the other to form expressions as part of the

[5]We process test responses (folder *test_responses/* in the dataset) with a script (*tools/2_extract_learning_gain_from_test_responses.ipynb* in the tools) to compute the *learn* for each team (in table *processed_data/learning_features/justhink19_learning_features.csv*, available with the tools).

[6]In fact, the IF could also be following their own intuitions and ignoring the IG. Yet, ultimately we think this is a justified assumption, as only the IF is in control of the actions.

shared lexicon. By aligning, they establish a shared lexicon, and thus align their representations of the activity with each other.

- **Submission of solutions.** Since the interlocutors have to submit their solution together by pressing the "submit" button that is present in both views, they have to, at least, align in terms of their intent to submit. Alternatively, one interlocutor must convince the other to reach a common intent.

The JUSThink Dialogue and Actions Corpus: Transcribing a subset of dialogues

Dataset description The JUSThink Dataset consists of 76 children in teams of two (41 females: $M = 10.3$, $SD = 0.75$ years old; and 35 males: $M = 10.4$, $SD = 0.61$). There is one problem solving session, i.e. task, per team. 8 out of the 38 teams ($\approx$ 21%) found an optimal solution to the activity. The teams were formed randomly, without considering the gender, nationality, or the mother tongue. They include mixed and same gender pairs, and this information is available but not used in this thesis. The study was conducted in multiple international schools in Switzerland, where the medium of education is in English, and hence students are proficient in English. The dataset contains:

1. **The recorded audio files**: Audio was recorded as two mono audio channels synchronised to each other, with one lavalier microphone per channel. The interlocutors were asked to speak in English. The microphones were clipped onto the interlocutors' shirts.

2. **Event log files**: Event log entries consist of timestamped touch and button press events, application status changes, and the interlocutor's edits to the tracks.

3. **Pre-test and post-test**: Interlocutors' responses to the items in the pre-test and post-test.

At an utterance level, to study the relationship between disfluencies and *verbal alignment, or alignment of expressions* (Chapter 7), we focus on the recorded audio files and transcribe a representative subset of the dataset. To study the relationship between disfluencies and *behavioural alignment, or alignment of actions* (Chapter 8), we combine this with the edit actions from the log files the with the transcripts.

At a dialogue level, we utilise the two measures of task success collected for this dataset; i.e. performance and learning outcomes. For *performance* in

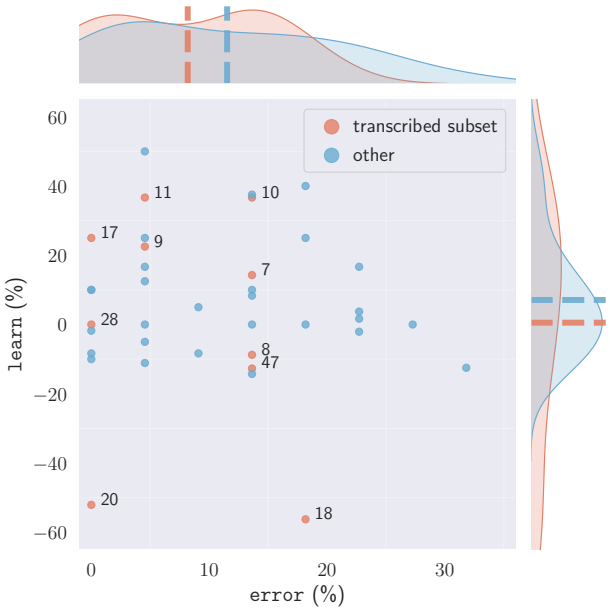


Figure 3.4: Scatter plot of the transcribed teams (red dots) and non-transcribed teams (blue dots) in the learning outcome (measure used is *learn*) vs. task performance space (measure used is *error*). The mean of a set (transcribed or other) is shown as a dashed line. Lines indicate the fit of a univariate kernel density estimate for the corresponding set. Numbers denote the ID of teams.

the task, we use the cost of the solution submitted at each attempt, extracted from the event log files. To compute a measure of *learning outcomes*, we use interlocutors' scores in the pre-test and the post-test (see Figure 3.2 for the way this is calculated).

Transcribing a representative subset. In order to study referring and alignment of the interlocutors, we selected a *subset* of 10 teams of the dataset. We relied on manual transcription, due to the poor performance of state-of-the-art automatic speech recognition systems on this dataset – which consists of children's speech with music playing in the background. The transcripts are publicly available online[7]. The teams were chosen to be a representative sample of the overall dataset (see Figure 3.4), keeping in mind the task success distribution; which we measure through performance in the task and learning outcomes observed from the pre-test and the post-test. We chose

[7]In the `transcripts/` folder, available from the Zenodo Repository, DOI: `http://doi.org/10.5281/zenodo.462710410.5281/zenodo.4627104`

the subset to transcribe based on the following considerations:

- The percentage of successful teams (30% compared to 21% of the whole dataset).
- The distribution of teams in performance and learning outcomes. Figure 3.4 shows how the teams are distributed in terms of performance (*error*) and learning (*learn*). As the figure shows, the means of performance and learning of the transcribed subset are similar to those of the whole dataset.
- The distribution of the number of attempts (i.e. submissions) and turns. Figure 3.5 shows how the teams are distributed in terms of the number of attempts and turns. The mean number of attempts and turns of the transcribed subset is similar to that of the whole dataset.

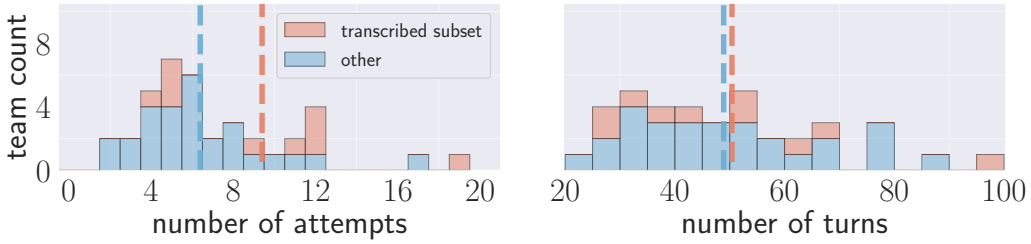


Figure 3.5: The distribution of transcribed tasks (red bars) and non-transcribed tasks (blue bars) in terms of the number of attempts (i.e. submissions) and number of turns. The mean of a set is shown as a vertical dashed line.

Table Table 3.1 provides further details about the transcribed subset. As shown, the mean duration of the task is ≈ 23 minutes, and the transcriptions account for ≈ 4 hours of data. The transcripts report which interlocutor is speaking (either *A* or *B*) and the start and end timestamps for each utterance, beside the utterance content. Utterance segmentation is based on Koiso et al. (1998)'s definition of an *Inter Pausal Unit (IPU)*, defined as "a stretch of a single interlocutor's speech bounded by pauses longer than 100 ms". We also annotated punctuation markers, such as commas, full stops, exclamation points and question marks.

The annotation of disfluencies Fillers, such as "uh" and "um", were transcribed, as well as 'oh'[8]. The above 3 spontaneous speech phenomena occur

[8]aligned with (Schiffrin, 1987) as an information management marker.

Table 3.1: Descriptive statistics for the transcribed teams ($N = 10$). SD stands for standard deviation.

	mean	SD	min	max
number of submitted solutions	9.4	4.7	4	19
number of turns in task	50.4	21.4	28	98
total duration (mins)	23.7	7.1	11.2	36.0
time per submission (mins)	2.6	2.3	0.7	12.8
duration of a turn (secs)	25.0	30.3	1.1	240.2
length of utterance (in tokens)	6.6	5.4	1	32

frequently in the dataset {“um”: 236, “uh”: 173, “oh”: 333}. Other phenomena, such as ‘ew’ or ‘oops!’ were also transcribed, however, their frequency is too low for analysis. Transcription included incomplete elements, such as “Mount Neuchat-” in “Mount Neuchat- um Mount Interlaken". When we say that we annotated the dataset following mixed guidelines from the LIWC and the SWBD disfluency annotation guidelines, we mean that we specifically annotate fillers (and in total, completed totally 4 checks of the fillers (and other non-semantic disfluencies) using PRAAT). This ensured that the disfluencies under consideration were accurately transcribed. However, we did not annotate other complex restart disfluencies. We also standardised pronunciation variants (subsequently discussed).

Pronunciation differed among and within interlocutors (for example, for the word ‘Montreux’, pronouncing the ending as /ks/ or /ø/), due to the unfamiliarity of the interlocutors with the referents, and individual accents. As our methodology is dependent on matching surface forms (refer to Chapter 7), we did not transcribe pronunciation variants of a word. We standardise variations of pronunciation in the transcriptions, and we do not account for e.g. variations in accent. A graduate student completed two passes on each transcript, which were then checked by another native English speaker graduate student for accuracy (particularly with the annotation of disfluencies) with experience in transcription/annotation tasks. Eg. 3.2 gives an example dialogue taken from the transcriptions.

***A*:** What about Mount Davos to Mount, Saint Gallen?
***B*:** Because what if, you did if we could do it?
***A*:** What about Mount um Davos to Mount Gallen?
***B*:** Mount ...
***B*:** oh Mount Davos
***A*:** yeah to Mount Gallen.
***B*:** to Mount Gallen yeah do that.
***B*:** Okay my turn.

(3.2)

Thus throughout the thesis we utilise these two datasets. We briefly describe the datasets in the results chapter, but frequently point to this chapter for additional clarity.

II | The Impact of Disfluencies on Performance in SLU Systems

Production of Fillers to Perception In Monologues using Discourse Level Features

4 | Computational Study on the Link Between the Production of Fillers and the Listener's Perception

4.1 Introduction and Background

From Part I, we see that a ubiquitous part of spoken language are disfluencies. It is well known that spontaneous speech is rarely fluent (Shriberg, 2005), and we observed this in Chapter 3, by surveying the primary datasets used in the thesis. In the past few years, there has been a widespread interest in SLU. Yet, methodologies in SLU to specifically study the disfluencies of the speakers and the information they can provide often remain overlooked. In this chapter, we conduct a preliminary analysis on the link the *production contexts of fillers*, and their consequent link with the listener's *perception*; specifically, using listener's assessments of how knowledgeable they perceive the speaker to be. We do so to assess the impact of fillers on performance in SLU models.

In communication, speaker's assess their own certainty about their knowledge, and communicate this knowledge to the listener. The listener in turn tries to assess the speaker's estimation about this information, i.e their *metacognitive state*. This entire process contributes to the status quo of *mutual understanding*[1]. In this part (Part II), we focus on a dataset of monologues, where annotations of the listener's the impression of the speaker's metacognitive state is available. In this preliminary analysis, we eliminate the component of immediate listener feedback that would be present in dialogues; i.e. the speakers are conscious of an *unseen* listener, but dialogue

[1] As we use Pickering and Garrod (2006)'s alignment theory of a common ground being implicit in a discourse, unless a misunderstanding requires a change of representation.

related disfluencies are not present. Thus we can focus uniquely on fillers, and the linguistic literature that shows the links between fillers and metacognitive states, as discussed in the Chapter 2. Metacognition in general may have wide applicability to for cross-lingual analyses. For example, previous research by Le Grezause (2017) shows that the presence and position of two fillers "uh" and "um" are associated with stance strength. Stance strength may be inter-linked with the phenomena of metacognition; i.e. the speaker's commitment to their stance; perhaps resulting in a use fillers to tone down the strength of the assertion.

We find in our analysis, that the integration of filler features (that account for different production contexts) allows for an improvement in the prediction of the listener's perception, and that preliminary results suggest that different functions of fillers (that we design based on psycholinguistic literature) correlate differently with the listener's perception of the speakers knowledgeability. While this task would be considered a broad SLU understanding task, we believe it has widespread applicability given the discursive properties of metacognition itself, and the increasing processing of speeches and reviews. In this broad SLU context, these findings give a first impression that fillers need not be removed as noise. Please see Chapter 5 for work on studying from a psycholinguistic perspective the role of fillers in production and perception using deep contextualised word embeddings.

In the following few paragraphs, we outline the relevant research for our preliminary analysis to study whether disfluencies are noise in SLU. For a more detailed discussion (such as the vast research on fillers, particularly from a psycholinguistic perspective, and the links between disfluencies and metacognition/ mutual understanding) please refer to Part I.

The speaker's use of fillers Fillers, as stated previously, are a type of disfluency that can be a sound ("um" or "uh" in English) filling a pause in an utterance or conversation. As detailed in Chapter 2, we observed that there is a vast amount of research on fillers that conclude that they are informative in understanding spoken language, for e.g. in Clark and Fox Tree (2002), Yoshida and Lickley (2010a), Brennan and Williams (1995), and Corley, MacGregor, and Donaldson (2007). There are several roles that fillers play within the verbal message. The speaker can use a filler to indicate a pause in speech (Clark and Fox Tree, 2002) or hesitation (Pickett, 2018). A speaker can use fillers to inform about the linguistic structure of their utterance, such as in their (difficulties of) selection of appropriate vocabulary while maintaining their turn (in dialogue). Importantly, fillers are linked to the metacognitive state of the speaker. It was observed that fillers and prosodic

cues are linked to a speaker's *Feeling of Knowing* or "expressed confidence", that is, a speaker's certainty or commitment to a statement (Smith and Clark, 1993).

The listener's perception of the speaker However, the meanings of fillers are contextual, and dependent on the perception of the listener (Grice, Cole, Morgan, et al., 1975; Clark and Fox Tree, 2002). The speaker *encodes* the meaning into their speech, while the listener *decodes* (or perceives) this meaning depending on the context. Hence, studies have also looked at the comprehension of fillers, that is, by taking into account the listener's understanding of speech uttered by the speaker (Corley and Stewart, 2008). Brennan and Williams (1995) found that listeners can perceive a speaker's metacognitive state, by using fillers and prosody as cues to study what they refer to as the *Feeling of Another's Knowing*; or the listener's perception of a speaker's expressed confidence/ certainty in their speech.

Research on metacognitive states The idea of metacognitive states is applicable a wide variety of communicative contexts. In this context (outside of the link between fillers and metacognition (Smith and Clark, 1993; Brennan and Williams, 1995)), research on the speaker's metacognitive state has focused on its' link to prosody (Pon-Barry, 2008; Smith and Clark, 1993; Brennan and Williams, 1995), facial expressions and gestures (Swerts and Krahmer, 2005), and overt lexical cues (Jiang and Pell, 2017) – words that explicitly mark uncertainty/certainty, such as ("I'm unsure", "definitely"...). In a related field, there are studies that automatically predict points of uncertainty in speech (Schrank and Schuppler, 2015; Dral, Heylen, and Akker, 2011).

Fillers used as features in SLU In broad SLU (predominantly studies that would overlap with affective computing), fillers (such as filler count) are commonly used as an attribute to study persuasiveness (Park et al., 2014) and big 5 personality traits (Mairesse et al., 2007). The Computational Paralinguistics Challenge 2013, focused on detecting fillers, which they considered to be a "social signal" (Schuller et al., 2019), acknowledging the importance of fillers in the listener-speaker dynamic. Along these lines, research in personality computing has the most consistent correlation, with observations made from speech; including paralanguage, such as fillers (Ekman et al., 1980; Vinciarelli and Mohammadi, 2014). However, there is a missing link between the computing of linguistic level properties of fillers that may ultimately contribute to these areas in personality computing. One

such example of this, is areas of research that may predetermine based on the context (such as in job interviews, where fluency is assumed to be desirable) whether the impact of fillers is positive or negative (such as in Rasipuram and Jayagopi (2016)). While indeed, the speaker's use of fillers may have an effect on hireablity given the linguistic research that has consistently shown that listener's comprehension is affected by fillers; it does not account for the nuances of the ways fillers may be used, and the dimensions of the message; from formulation and articulation by the speaker to then comprehension by the listener (and in the case of dialogues; feedback).

Thus fillers are commonly used as a feature in broad SLU, but research is lacking in a focused analysis of them. To our knowledge, only a few studies use specifically focus on the informativeness fillers as the *focal* feature in other computing tasks (unless in disfluency detection); for e.g. are found to be successful in stance prediction (stance referring to the subjective spoken attitudes towards something (Haddington, 2004)) (Le Grezause, 2017) and turn-taking prediction (Saini, 2017).

The rest of the chapter is organised as follows: In Sec. 4.2, we outline the drawbacks of current research and our research questions. In Sec. 4.3, we describe our filler features based on linguistic literature, the salient points of the dataset used, and the models used. Sec. 4.4 gives the results and discussion, while Sec. 4.5 discusses the conclusion of the experiments.

4.2 Research Questions

Drawbacks of existing works So far, there has not been a mainstream interest in the prediction of the perception of the speaker's metacognitive state from an SLU perspective (though smaller tracks in conferences may contain such works, for e.g. "Psycholinguistic influences on dialogue system design"), nor studies that holistically focus on the informativeness of disfluencies.

Related studies on uncertainty detection are limited to a narrow range of question-answering (QA) tasks (Schrank and Schuppler, 2015) and do not account for a discourse-level analysis, as uncertainty detection typically detects utterance level points of uncertainty based on overt lexical cues, and not the listener's overall impression. Points of uncertainty in speech, may very well lead to a general impression of confidence of the speaker. Linguistic work that shows the links to fillers and metacognition are similarly limited to QA datasets (Brennan and Williams, 1995) or single utterance evaluations (Fraundorf and Watson, 2011), but not overall discourse levels. The speaker may be typically asked to give an answer (usually the length of a sentence)

to a question with/without a filler inserted, based on a predetermined script. The listener is then asked to form a perception based on this answer, and may explicitly be asked to rate how certain the speaker seemed in their utterance.

This raises the question whether the link between fillers and metacognition exists in (more broader forms of) spontaneous speech itself. In natural conversation, listener's may not be aware of the use of fillers, unless overused or used in the wrong context (Tottie, 2014). We study spontaneous speech in a monologue where the speakers voluntarily and naturally recorded themselves (and thus has naturally occurring fillers). The annotations that the listener gives are impressions created after hearing the entire monologue (relevant details highlighted in Sec. 4.3 in this chapter, but for a detailed analysis of the POM dataset please refer to Chapter 3), without explicitly being asked to pay attention to the speaker's use of fillers.

Thus despite the rich literature regarding fillers, from an SLU perspective, fillers remain mostly unexplored, or overlooked as noise. The aim of this chapter is to do a preliminary feature study of fillers, to see whether the fillers in different production contexts, can contribute to the prediction of perception of the speaker's expressed confidence. We use the annotator's (listener's) labels for this task. The research questions are as follows:

RQ1: Are the different production contexts of fillers correlated differently to the listener's perception of the speaker's metacognitive state? Based on the previous research, we would like to automatically compute a set of filler features based on different production contexts, and observe whether they correlate differently with the perception of the speaker's metacognitive state in a preliminary statistical analysis.

RQ2: Can fillers be informative in the prediction of the listener's perception of the speaker's metacognitive state? Using the filler features we designed in Sec. 4.2, we would like to see whether these features impact the prediction of the listener's perception of a speaker's metacognitive state using linear models.

4.3 Methodology

Designing Filler Features

Production context to model filler features We design a set of psycholinguistic inspired features based on the *filler-as-word*[2] hypothesis proposed by Clark and Fox Tree (2002). In this hypothesis, a filler functions as an interjection, therefore its meaning is highly dependent on the context. Fillers are hence distinguished by their basic meaning and their implied meaning, or implicature (Clark and Fox Tree, 2002). We use this hypothesis of basic and implicature interjections as a basis for our feature representation. This is done to account for some of the vast literature available on the production of fillers. The basic meaning of a filler, is to announce the initiation, at the time of the filler $t(filler)$ by the speaker, of what is expected to be a delay in speech (Clark and Fox Tree, 2002). Although fillers have one basic meaning; they have several implicatures. A filler can have an implicature that would broadly fall under a speaker being nervous, hesitating or uncertain (Pickett, 2018). A filler can be used by the speaker to indicate pausing to (re)formulate thoughts at discourse boundaries (Swerts, 1998) (whether this is consciously done by the speaker or not remains an open question). Fillers can have an implicature linked to a speaker's stance, both stance polarity and stance strength (Le Grezause, 2017). Plauché and Shriberg (1999) found that subsets of repetitions (type of repair) reflect *different problems in planning*. Fillers are associated with the cognitive load of the speaker (Shriberg, 2001); disfluency rates increase for longer sentences, suggesting an increase in cognitive load of the speaker.

We thus design two sets of filler features; a basic set and an implicature set. The basic set includes three features: *num*, *uh*, *um*, corresponding to the total number of fillers, the total number of filler "uh" and the total number of filler "um", respectively. The implicature set includes the following features (trying to account for various psycholinguistic findings of fillers in production):

- *init*, *med*: the number of fillers that occur sentence-initially (start of the sentence) and sentence-medially respectively (within the sentence) as fillers to denote syntactic/positional marking.

- *s_stance*, *w_stance*: the number of fillers occurring in sentences that contain strong stance tokens (i.e. very positive or very negative) (where

[2]Although this view has been contested, we utilise the idea to mean there are different production contexts of fillers that may affect the listener's impression differently, while a basic meaning of a filler used to initiate a pause in speech remains constant.

token refers to both fillers and words), or weak stance tokens (i.e. positive or negative) respectively. This allows us to distinguish between fillers that mat be used to tone down the force of assertion. Token-level stance and polarity labels for the dataset are taken from Garcia et al. (2019).

- *stutter*: the number of fillers present in sentences that contain "stutter" markings (please note as described in the Chapter 3, that "stutter" markings do not denote the clinical sense of stuttering, merely some form of repair (see Table 2.1)). Park et al. (2014) describe these fillers as "speech disturbances" (following Mahl (1956)'s terminology).

- *s_len*, *r_len*: the length of the sentence in tokens, and the length of the review in tokens respectively.

We compute textual features as a representation of the whole video to be used in statistical analysis and as input to our linear models. We compute *num*, *um*, *uh*, *init*, *med*, *stutter*, *s_stance* and *w_stance* by counting all instances respectively, and normalising by the total number of words/tokens spoken in each video, as was originally done in Park et al. (2014). Both *s_len* and *r_len* are normalised by the average sentence length and the average review length of the videos.

Dataset

Persuasive Opinion mining (POM) dataset For this work, we choose the POM dataset (Park et al., 2014)[3], a dataset of 1000 (American) English monologue movie review videos. Speakers recorded themselves (video and audio) giving a movie review, which they rated from 1 star (most negative) to 5 stars (most positive). The movie review videos are freely available on ExpoTV.com, and are completely in the wild; speakers were simply reviewing a movie without the knowledge that their review would eventually be annotated for such a context. 3 annotators (or listeners) per video were then asked to label the movie reviews for high level attributes, such as confidence. We think this dataset is particularly relevant for the following reasons:

- Since this is a dataset of monologues, it allows us to focus uniquely on the functions of fillers (Swerts, 1998). This is because the speaker is conscious of an *unseen* listener, but is not interrupted by the listener

[3]Note that relevant details of the dataset are highlighted in this section for a quick reference, but for further details of the POM dataset please refer to Chapter 3.

with other dialogue related disfluencies, such as backchannels ("Uh-huh"). This also minimises some turn-taking properties of fillers, such as when they are used by the speaker to hold the speaker turn. Additionally, the annotators were never asked specifically to pay attention to the speaker's use of fillers.

- Filler annotations of "uh" and "um" have been manually transcribed. Each transcription of a movie review video was reviewed by experienced transcribers for accuracy after being transcribed via Amazon Mechanical Turk (AMT) (Park et al., 2014). The experience of the transcriber is important, as Zayats et al. (2019b) shows that transcribers tend to misperceive disfluencies and indeed, this can affect the transcription of fillers (Le Grezause, 2017). The filler count of this dataset is high (roughly 4% of the transcriptions, for comparison, the Switchboard (Godfrey, Holliman, and McDaniel, 1992) dataset of human-human dialogues, consists of $\approx 1.6\%$ of fillers (Shriberg, 2001)). Sentence markers have been manually transcribed, with the practice of the filler being annotated sentence-initially, if the filler occurs between sentences (in this dataset, utterance segmentation is not available, and is interchangeable with sentence).

- The inter-annotator agreement for several attributes is high; with confidence (label we use to denote the listener's perception of the speaker's metacognitive state) of Krippendorff's $\alpha = 0.73$ (Park et al., 2014). For confidence, annotators were asked "How confident was the reviewer", and had to rate the speaker on a Likert scale of 1-7 with given labels: 1 (not confident), 3 (a little confident), 5 (confident) and 7 (very confident). Additional details can be found in the original (Park et al., 2014). Summary statistics, are given in Table 4.1.

We take the Root Mean Squared (RMS) value for the 3 annotations per video as the final label, to reflect higher annotation scores[4]. We remove the labels in the 1-2 range, due to sparsity of these labels in this preliminary analysis.

[4]Though the inter-annotator agreement for confidence is high, we choose RMS as a way to handle disagreement between annotators. For example, annotation labels $\{3, 5, 7\}$ would result in mean value of 5, not highlighting that one annotator found the reviewer particularly confident. The RMS value however (≈ 5.3), slightly enhances the high confidence label.

Description	Value
Reviews that contain fillers	**792**
Total number of review used	892
Total "um" fillers in the corpus	4969
Total "uh" fillers in the corpus	4967
Total fillers in the corpus	**9936**
Number of tokens in the corpus	230462
% of tokens that are fillers	**4.31**
Average length (in tokens) of a review	255.9

Table 4.1: Brief summary statistics of the POM dataset.

Models

The contextual information provided by fillers, may contain information that is essential for predicting the listener's perception of the speaker's expressed confidence. We incorporate such information into our experiments through feature representation that is based on state of the art linguistic literature (Clark and Fox Tree, 2002; Le Grezause, 2017; Pickett, 2018; Shriberg, 2001; Smith and Clark, 1993; Brennan and Williams, 1995), statistical analysis and linear machine learning models. We use the transcripts of the dataset as our input. We do not to utilise audio features already provided by the CMU-Multimodal SDK, due to the poor results of the forced alignment (Yuan and Liberman, 2008; "Gentle forced aligner [computer program]") algorithms. We are thus not able to pinpoint specific audio regions for fillers.

RQ1: Are the different production contexts of fillers correlated differently to the listener's perception of the speaker's metacognitive state? In a preliminary analysis, we take the RMS value of the confidence labels provided by the three annotators as the final rating of the speaker giving the review. We then consider reviews that are categorised as high-confidence (HC) and low-confidence (LC). Since confidence ratings are positively skewed[5], we take ratings of 3.5 (a little confident) and below to denote LC speakers, and 6 and above to denote HC speakers. The resulting size of the categories are 171 HC and 133 LC speakers. We then calculate a Pearson's correlation coefficient r between all the filler features for HC and LC reviews and the final rating of confidence using the Benjamini-Hochberg procedure for multiple testing correction. We repeat this procedure for all

[5]this is shown both in the annotation guidelines as discussed in Sec. 4.3, and the ratings itself, as annotator's may have hesitated to rate the speaker 1 (not confident), and preferred instead to use the label 3 (a little confident)

reviews in the dataset (not only HC/LC reviews) using the same RMS value as our final rating of the three annotators.

RQ2: Can fillers be informative in the prediction of the listener's perception of the speaker's metacognitive state? We use the original standard training, testing and validation folds provided in the CMU-Multimodal SDK (*CMU-Multimodal SDK*). We use the filler features as defined in Sec. 4.3, to predict the final rating of expressed confidence. Our baseline is a Random Vote (RV); where 100 random draws respecting the train dataset balance were made. We use a mean squared error (MSE) to evaluate our models. For RV, the MSE is averaged over these 100 samples. We take 2 classic Machine learning algorithms, that is Random Forest (RF) and Ridge regression (RR), with respective hyper-parameter searches on the validation set. We choose RR as we have multi-collinear features, and both RF and RR have easy interpretability for feature importance.

4.4 Results and Discussion

4.4.1 RQ1 Are the different production contexts of fillers correlated differently to the listener's perception of the speaker's metacognitive state?

The average use of fillers in HC/LC reviews The box plots comparing a speaker's average use of fillers (*num*) compared to the HC/LC categories is given in Figure 4.1. Inspecting the box plot, it is likely that speaker's in the HC category use less fillers overall than speaker's in the LC category (median filler rate of 3.049 and 4.82 respectively, with $U = 7329.5$ and $p < .0001$ by two sided Mann-Whitney U test with Bonferroni correction).

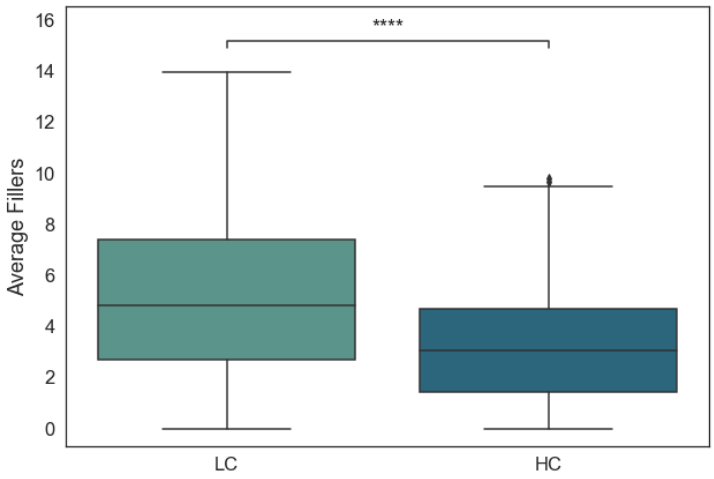


Figure 4.1: Box plots showing the speaker's average use of fillers *num* for HC/LC reviews, where $**$ denotes $p < .0001$. The median *num* is 4.82 and 3.049 for LC and HC respectively, with $U = 7329.5$ by two-sided Mann-Whitney U test with Bonferroni correction.

When we look at the speaker's use of each individual filler *uh* (Figure 4.2) and *um* (Figure 4.3) respectively, an initial impression is that the two fillers are are not used differently (at least by comparing rate of each filler with HC/LC) by the speakers. Firstly, the speakers (American English speakers) on average marginally use more *um* than *uh*, in both HC and LC categories (with median *um* filler rate of 1.39 and 1.97 respectively, compared to median filler rate of *uh*; 1.22 and 1.83 respectively). This difference is small also considering that the overall number of *uh* and *um* fillers in the dataset are $\approx$ the same, as shown in Table 4.1. However, looking at the box plots, we see that there is greater evidence to support the relationship between the filler *um* ($U = 9143.5$ and $p < .01$) and HC/LC categories compared to the filler *uh* ($U = 9687.0$ and $p < .05$) by two sided Mann-Whitney U test with Bonferroni correction, with HC speakers on average tend to use less *um* than LC speakers. Following Figure 4.1, the average rate for each individual *um* and *uh* filler is correspondingly higher for the LC category, and lower for the HC category.

The placement of fillers in HC/LC reviews The box plots for fillers that occur sentence initially (*init*) and sentence medially (*med*) are given in Figure 4.4 and Figure 4.5 respectively. There is evidence to suggest a rela-

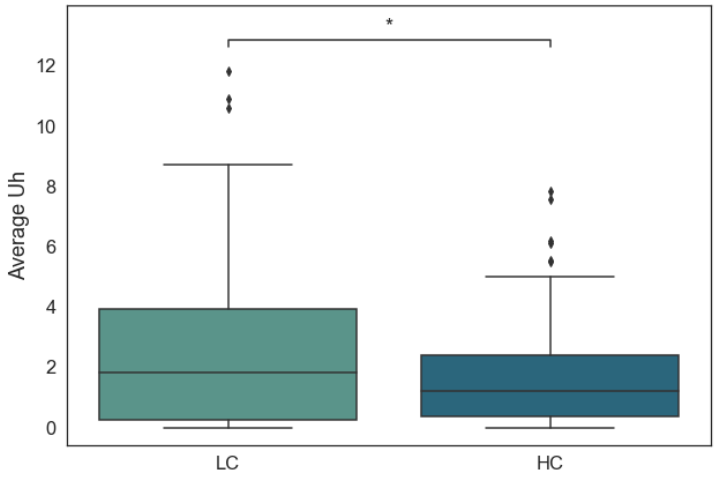


Figure 4.2: Box plots showing the speaker's average use of fillers *uh* for HC/LC reviews, where $*$ denotes $p < .05$. The median *uh* is 1.83 and 1.22 for LC and HC respectively, with $U = 9687.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

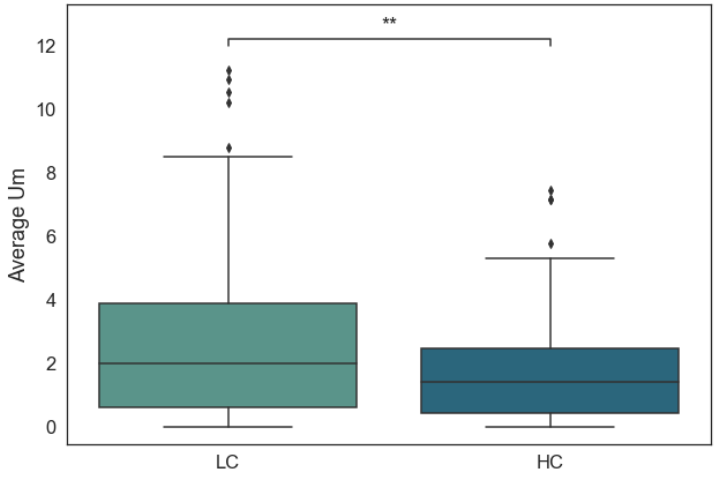


Figure 4.3: Box plots showing the speaker's average use of fillers *um* for HC/LC reviews, where $**$ denotes $p < .01$. The median *um* is 1.97 and 1.39 for LC and HC respectively, with $U = 9143.5$ by two-sided Mann-Whitney U test with Bonferroni correction.

tionship between both *med* fillers ($U = 7443.0$ and $p < .0001$) and HC/LC categories and *init* fillers ($U = 9505.0$ and $p < .05$) with HC/LC categories, but stronger evidence for *med* fillers and confidence. Speakers use fillers more in the *med* position than *init* position, in both LC and HC categories (with median *med* filler rate of 3.31 and 1.86 respectively, compared to median filler rate of *init*; 1.34 and 0.90 respectively). This is expected considering that *med* fillers may still occur at natural prosodic boundaries, such as after after punctuation and silent pauses. The results show that a bulk of fillers occur sentence initially (with 40% of fillers used by LC speakers occurring sentence initially, and 48% for HC speakers) compared to other locations. This is consistent with many other works that find the distribution of fillers commonly at discourse boundaries (Shriberg, Bates, and Stolcke, 1997; Swerts, 1998; Shriberg, 2001; Swerts and Geluykens, 1994). Given Figure 4.1, we expect the rate of fillers in the LC category to be higher than the HC category. However, LC speakers on average use ≈ 2.5 times more *med* fillers compared to *init* fillers (median of $med = 3.31$ and $init = 1.34$), while HC speakers use ≈ 2 times more *med* fillers compared to *init* fillers (median of $med = 1.86$ and $init = 0.90$ respectively). Thus, HC speakers use fillers more fillers sentence initially than LC speakers.

Fillers compared to the length of HC/LC reviews The box plots for the average sentence length (in tokens) of the review (s_len), and the length of the review (in tokens) (r_len) are given in Figure 4.6 and Figure 4.7 respectively. Inspecting the box plots, we see that HC speakers on average give longer reviews (in tokens) compared to LC speakers (median 207 and 266 tokens respectively). It is likely that there is a relationship between HC/LC and the r_len ($U = 6148.5$ and $p < 0.0001$ by two-sided Mann-Whitney U test with Bonferroni correction). Interestingly, the median for s_len between HC and LC speakers is $\approx$ the same, with 15.84 and 15.86 respectively. We cannot conclude whether there is a relationship between HC/LC and s_len, with $U = 11346.5$ and $p = 0.49$. This gives further evidence that despite HC speakers on average giving longer reviews; they use less fillers per sentence compared to LC speakers, given that the median s_len for both HC/LC are the same.

Fillers around other disfluencies in HC/LC reviews The box plot for the average use of fillers that are part of other disfluent structures (*stutter*) (called "stutters" in the POM dataset (Park et al., 2014), following terminology from Mahl (1956)) is given in Figure 4.8. We see that HC speakers use *stutter* fillers on average less than LC speakers (with median of 0.26

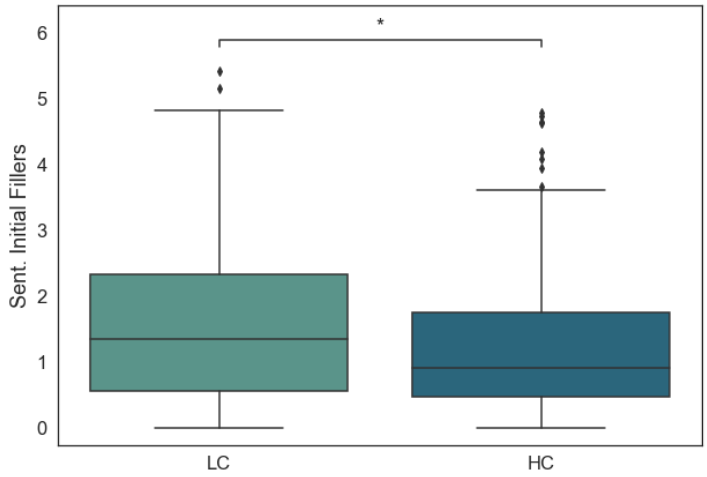


Figure 4.4: Box plots showing the speaker's average use of sentence initial fillers (*init*) for HC/LC reviews, where $*$ denotes $p < .05$. The median *init* is 1.34 and 0.90 for LC and HC respectively, with $U = 9505.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

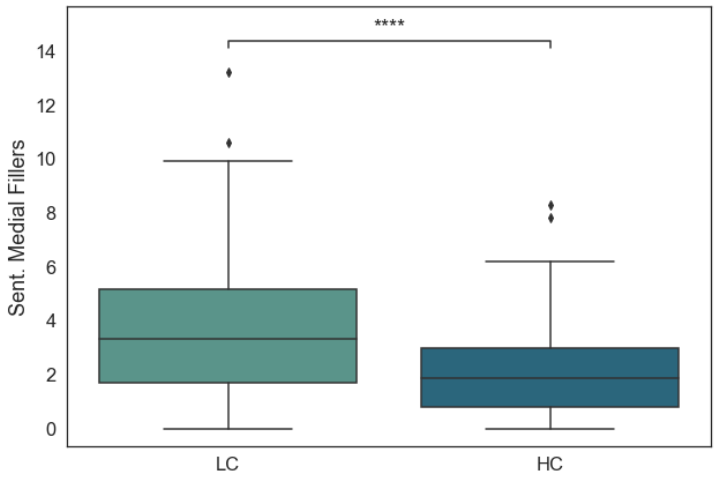


Figure 4.5: Box plots showing the speaker's average use of sentence medial fillers (*med*) for HC/LC reviews, where $****$ denotes $p < .0001$. The median *med* is 3.31 and 1.86 for LC and HC respectively, with $U = 7443.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

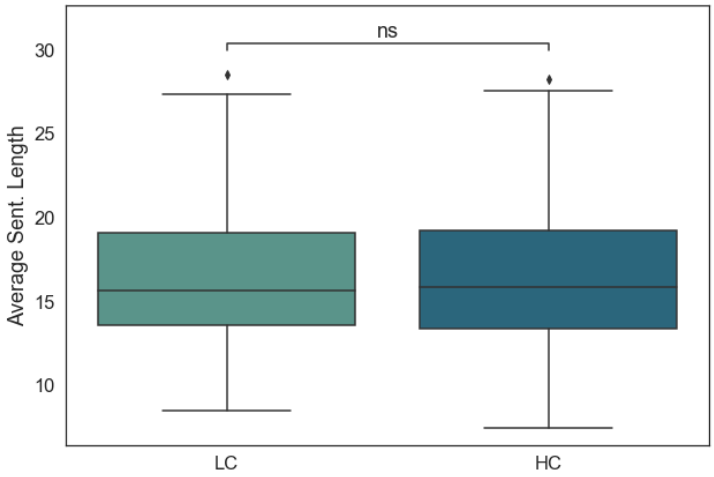


Figure 4.6: Box plots showing the speaker's average length of the sentence (s_len) for HC/LC reviews, where ns denotes $0.05 < p$. The median s_len is 15.86 and 15.84 for LC and HC respectively, with $U = 11346.5$ by two-sided Mann-Whitney U test with Bonferroni correction.

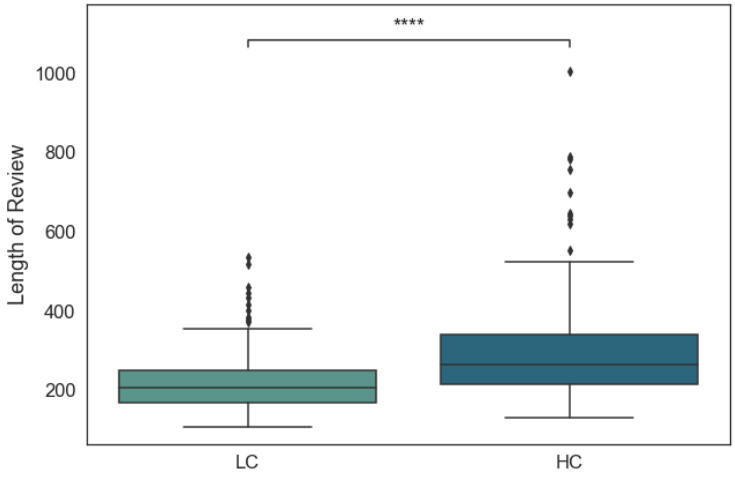


Figure 4.7: Box plots showing the speaker's review length (r_len) in tokens for HC/LC reviews, where $****$ denotes $p < .0001$. The median med is 266 and 207 tokens for LC and HC respectively, with $U = 6148.5$ by two-sided Mann-Whitney U test with Bonferroni correction.

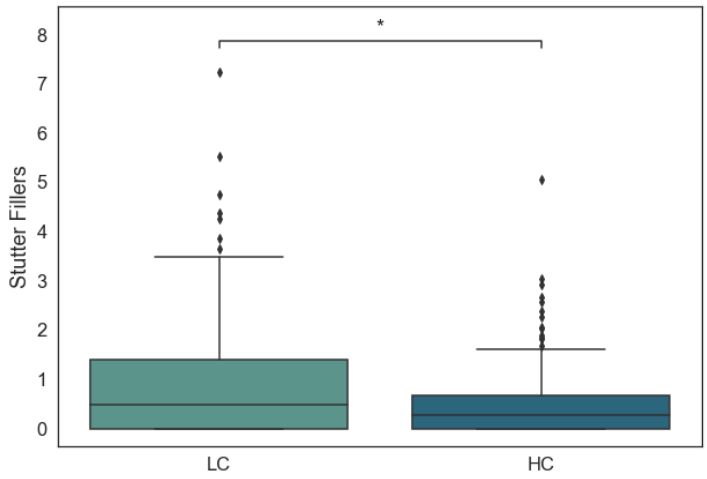


Figure 4.8: Box plots showing the speaker's use of fillers around other disfluencies (*stutter*) for HC/LC reviews, where $*$ denotes $p < .05$. The median *med* is 0.47 and 0.26 for LC and HC respectively, with $U = 9864.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

and 0.47 respectively, and $U = 9864.0$, $p < .05$ by two-sided Mann-Whitney U test with Bonferroni correction.). LC speakers are also more varied with their use of *stutter* fillers, with standard deviation $SD = 1.27$ compared to HC speakers, who vary considerably less, with $SD = 0.72$. The box plot also is reflective of the average use of fillers as shown in Figure 4.1; with the likeliness that LC speakers use more fillers (and consequently more *stutter* fillers) than HC speakers.

Stance fillers in HC/LC reviews The box plots for the speaker's average use of fillers in an utterance containing a weak stance (*w_stance*) and the speaker's average use of fillers in an utterance containing a strong stance (*s_stance*) for HC/LC reviews are given in Figure 4.10 and Figure 4.9 respectively. The use of *w_stance* fillers follows the same trend of the speaker's average use of fillers for HC/LC reviews; with HC speakers on average using less *w_stance* fillers compared to LC speakers (median 0.40 and 0.74 respectively), suggesting a possible link between confidence and *w_stance* fillers (with $U = 8965.0$ and $p < 0.01$ by two-sided Mann-Whitney U test with Bonferroni correction.). We cannot conclude a link between *s_stance* fillers and confidence (with $U = 10226.0$ and $p = 0.05$ by two-sided Mann-Whitney

Table 4.2: Re-calculated Pearson's r correlation coefficient for the filler features with confidence for only HC and LC reviews, using the Benjamini-Hochberg procedure for multiple testing correction. $*$ means $p \leq 0.05$, $**$ means $p \leq 0.01$, $***$ means $p \leq 0.001$ and $****$ means $p \leq 0.0001$.

	r	pValue	pValue Corr.	Reject
med	-0.34	0.0***	0.0***	True
num	-0.33	0.0***	0.0***	True
um	-0.24	0.0***	0.0***	True
uh	-0.21	0.0***	0.0***	True
w_stance	-0.21	0.0***	0.0***	True
stutter	-0.20	0.001***	0.001***	True
init	-0.16	0.006***	0.007***	True
s_len	0.01	0.896	0.896	False
s_stance	0.05	0.393	0.432	False
r_len	0.35	0.0***	0.0***	True

U test with Bonferroni correction). Notably, speakers regardless of HC/LC categories use fillers less when uttering a strong stance assertion (*s_stance*) in this dataset (median 0.16 and 0 for HC and LC speakers respectively), compared to *w_stance* fillers. An initial impression is that fillers may be used to tone-down the force of an assertion, with more fillers occurring in weak stance assertions compared to strong stance assertions.

Roles of fillers in HC/LC reviews Table 4.2 gives the Pearson's correlation coefficient r with the filler features (described in Sec. 4.3) and confidence for HC/LC reviews, using the Benjamini-Hochberg procedure for multiple testing correction. For visualisation, the corresponding heatmap is provided in Figure 4.11. We choose Pearson's correlation coefficient r, to measure the association between two continuous variables (in our case, each filler feature as outlined in Sec. 4.3 and the final rating of confidence calculated). To interpret the results, the coefficient r ranges from -1 to 1, where 0 would mean no correlation found between the two variables using the test; whereas values of -1 and 1 would indicate a perfect negative or positive correlation between the two variables. The magnitude of Pearson's correlation coefficient r can be interpreted by using the guidelines as given in Hemphill (2003), i.e. $|r| < 0.20$ as "lower third" or "small" correlation, $20 < |r| < 0.30$ as "middle third or "medium" correlation, and $30 > |r|$ as "upper third" or "strong" correlation (keeping in mind sample size).

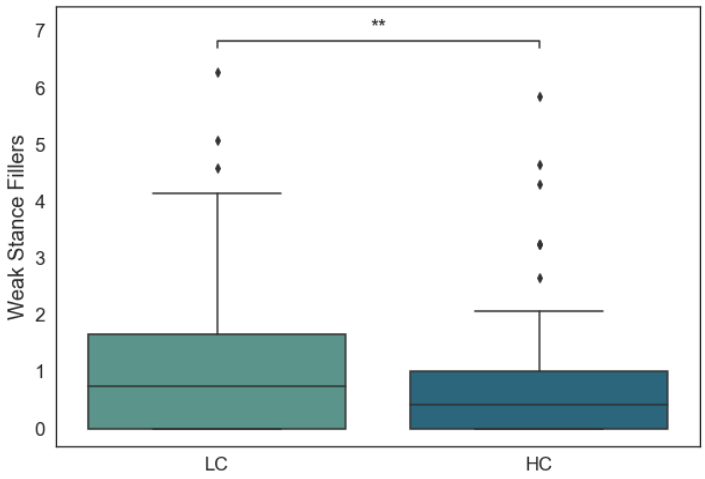


Figure 4.9: Box plots showing the speaker's average use of fillers in an utterance containing a weak stance (w_stance) for HC/LC reviews, where $**$ denotes $p < 0.01$. The median w_stance fillers is 0.74 and 0.40 for LC and HC respectively, with $U = 8965.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

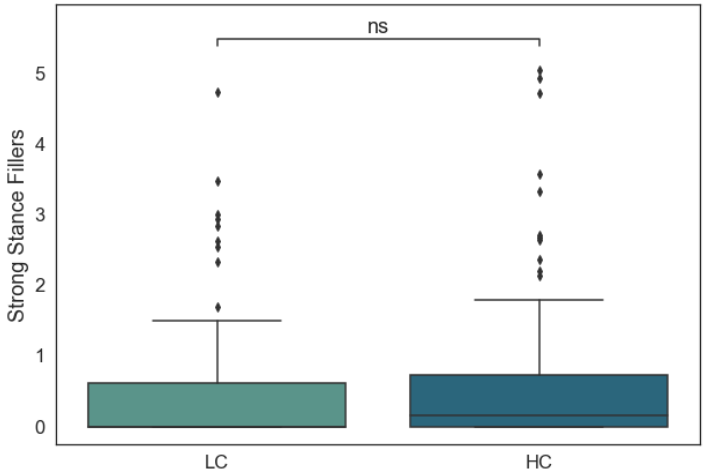


Figure 4.10: Box plots showing the speaker's average use of fillers in an utterance containing a strong stance (s_stance) for HC/LC reviews, where ns denotes $0.05 < p$. The median s_stance fillers is 0 and 0.16 for LC and HC respectively, with $U = 10226.0$ by two-sided Mann-Whitney U test with Bonferroni correction.

As shown in Table 4.2, we see that there exists a large negative correlation between the basic set of features; i.e. the total number of fillers and the label of confidence (*num*, where $r = -0.33$ and corrected $p \leq 0.001$), as well as a medium negative correlation with the number of individual fillers *uh* (where $r = -0.21$ and corrected $p \leq 0.001$) and *um* (where $r = -0.24$ and corrected $p \leq 0.001$) with confidence[6]. By Pearson's correlation coefficient, as the number of fillers increase, the level of confidence decreases.

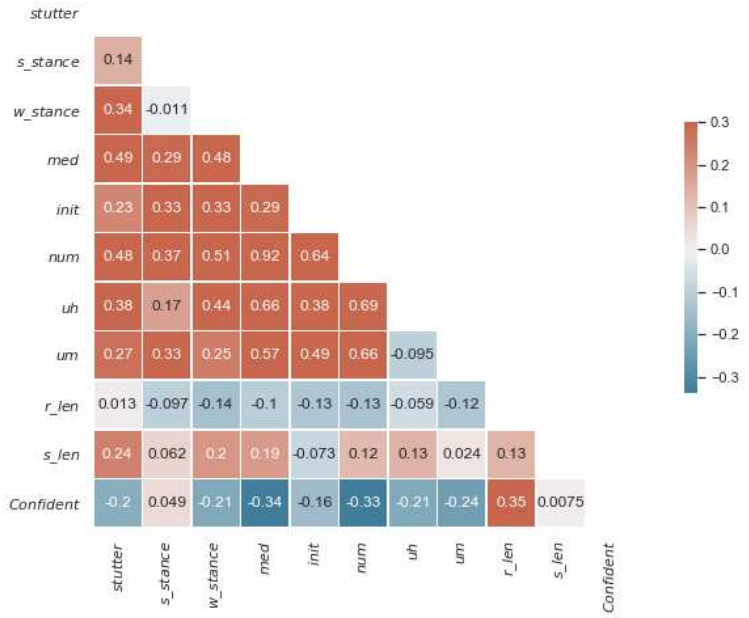


Figure 4.11: Heatmap giving the Pearson's r correlation with fillers features as described in Sec. 4.3 and the final rating of confidence for HC/LC reviews. The corresponding table with the pValues and recalculated pValues using the Benjamini-Hochberg (BH) procedure for multiple testing correction is Table 4.2.

Looking at the implicature set of filler features, we observe the same effect for fillers that occur sentence medially (*med*, where $r = -.34$ and corrected $p \leq 0.001$), indicating that as the number of *med* fillers increase, the level of confidence decreases. Given the results of the test, there is evidence to support that the placement of fillers could be important when

[6]This is to be expected given the multicollinearity of the features. However, given that RQ1 is an exploratory analysis, we follow the guidelines given in `https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/`. We utilise Variance Inflation Factors (VIF) for the regression analysis RQ2

considering the variable of confidence; with a small negative correlation with fillers that occur sentence initially (*init*) and confidence (with $r = -0.16$ and corrected $p \leq 0.001$), compared to the medium negative correlation between *med* and the label of confidence. Thus there may be a distinction between the placement of fillers and the relationship with confidence.

Interestingly, we see a small negative correlation with fillers that occur in utterances that contain weak stance tokens (*w_stance*, where $r = -0.21$ and $p \leq 0.001$) and confidence, but cannot conclude on the relationship between fillers that occur in utterances that contain strong stance tokens (*s_stance*, where $r = 0.05$ and $p = 0.4$) and confidence. As shown in Figure 4.10, HC speakers on average tend to use less *w_stance* fillers compared to LC speakers. Lastly, following the box plots shown in Figure 4.6 and Figure 4.7, there is greater evidence to support the relationship between the length of the review in tokens (*r_len*) and confidence; with a medium positive correlation where $r = 0.35$ and $p \leq 0.001$. However, we cannot conclude on the relationship between the average length of the sentence in tokens (*s_len*) with confidence, with $r = 0.01$ and $p = 0.39$. Thus the longer the reviews (i.e. an increase in *r_len*), the higher the rating of confidence. *stutter* fillers also have a small negative correlation with confidence, with $r = -0.20$ and corrected $p \leq 0.001$, indicating that listeners do perhaps perceive these filler-stutter occurrences as speech disturbances as hypothesised in Park et al. (2014).

Functions of fillers in all reviews Table 4.3 gives the Pearson's correlation coefficient r with the filler features (described in Sec. 4.3) and confidence for all reviews, using the Benjamini-Hochberg procedure for multiple testing correction. For visualisation, the corresponding heatmap is provided in Figure 4.12. Please refer to the previous paragraph for the guidelines to interpret the results.

Inspecting Table 4.3, we firstly see that even using all the reviews, there is evidence to support a negative correlation between the basic filler feature variables (*num*, *um*, *uh*) and all implicature variables except for *s_stance* and *s_len*. The shift from "medium" to "small" correlation is to be expected given the larger sample size used for this test (i.e. all reviews, not simply HC/LC categories). Given the results of this test and the results as given in Table 4.2, we can infer that there may be a relationship between the average number of fillers in a review and the final rating of confidence that the listener gives the speaker. It is unclear whether the two fillers are used differently when considering the rating of confidence; despite the findings of Le Grezause (2017). We can also infer that the placement of the filler in

Table 4.3: Re-calculated Pearsons r correlation coefficient for the filler features with confidence for all reviews using the Benjamini-Hochberg procedure for multiple testing correction. Included for each feature is the Variance Inflation Factor (VIF). The guidelines for significance ($*$) are the same as in Table 4.2.

	r	pValue	pValue Corr.	Reject	VIF
num	-0.23	0.0***	0.0***	True	> 10
med	-0.22	0.0***	0.0***	True	> 10
um	-0.16	0.0***	0.0***	True	> 10
w_stance	-0.14	0.0***	0.0***	True	2
uh	-0.14	0.0***	0.0***	True	> 10
stutter	-0.12	0.0***	0.001***	True	2
init	-0.11	0.001***	0.001***	True	> 10
s_len	0.00	0.893	0.893	False	9
s_stance	0.04	0.249	0.274	False	2
r_len	0.27	0.0***	0.0***	True	6

the sentence/utterance could be correlated with the rating of confidence, with greater evidence to support the negative correlation between sentence medial fillers and the rating of confidence. According to Tottie (2014), listener's are typically not aware that fillers have been used, unless overused, or used in the wrong context. It is plausible that the higher the filler count is in "atypical" locations, i.e. not at discourse boundaries, the more conscious the listener may be of them, and thus impacting the rating of confidence. Lastly, it would appear that there is a distinction between the way fillers are used in utterances that contain tokens with weaker stance, compared to utterances that contain tokens with stronger stance. Fillers are used more frequently in utterances that contain weaker stance, and there may be a relationship with fillers used in this way with the variable of confidence. Furthermore, these preliminary results are encouraging in terms of an overall goal of this thesis; to study disfluencies in a holistic way – i.e. considering several of their cross-linguistic/ discursive roles such as in metacognition (Brennan and Williams, 1995; Swerts, 1998), stance (Le Grezause, 2017; Levow et al., 2014) and so on.

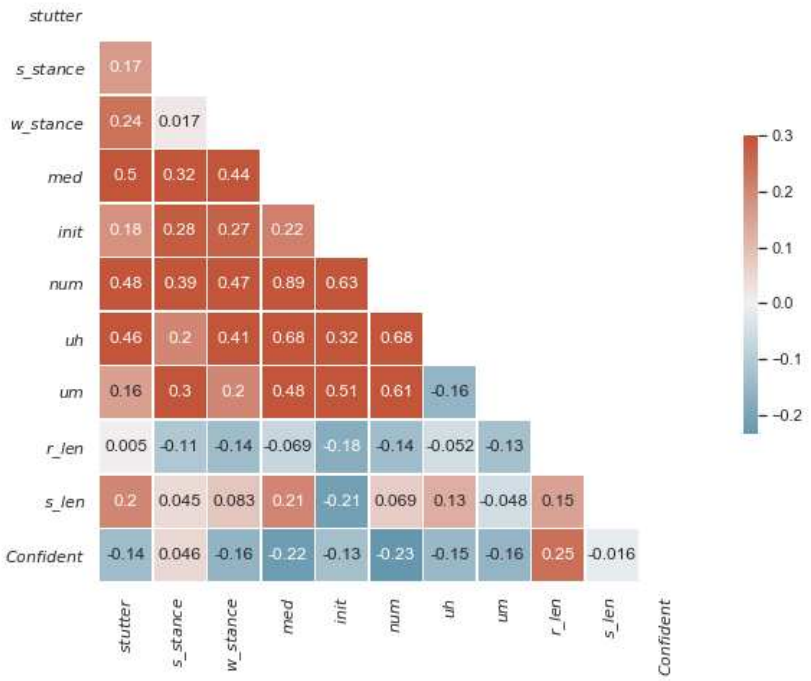


Figure 4.12: Heatmap giving the Pearson's r correlation with fillers features as described in Sec. 4.3 and the final rating of confidence for all reviews. The corresponding table with the pValues and recalculated pValues using the Benjamini-Hochberg (BH) procedure for multiple testing correction is Table 4.3

4.4.2 RQ2: Can fillers be informative in the prediction of the listener's perception of the speaker's metacognitive state?

As stated in Table 4.3, we use the original standard training, testing and validation folds provided in the CMU-Multimodal SDK (*CMU-Multimodal SDK*). We use the filler features as defined in Sec. 4.3, to predict the final rating of expressed confidence. Our baseline is a Random Vote (RV); where 100 random draws respecting the train dataset balance were made. We use a mean squared error (MSE) to evaluate our models. For RV, the MSE is averaged over these 100 samples. We take 2 classic Machine learning algorithms, that is Random Forest (RF) and Ridge regression (RR), with respective hyper-parameter searches on the validation set. We choose RR as we have multi-collinear features (i.e. which independent variables/ filler

Model	Features	MSE
RV baseline		2.90
RF	Basic (num)	1.04
	Basic ($uh + um$)	1.02
	Imp. (all)	**0.95**
	Basic + Imp. ($uh + um$, removing $med + init$)	0.98
RR	Basic (all)	1.15
	Imp (all)	1.14
	Basic+Imp.	1.13

Table 4.4: Results of the models described in Table 4.3. Imp. stands for implicature features.

features are correlated with the other)[7], and both RF and RR have easy interpretability for feature importance.

Firstly, for the RF model, we identify the multicollinear features. From RQ1, Table 4.3 gives the Pearson's correlation coefficient r with the filler features (described in Sec. 4.3) and confidence for all reviews. In the column "VIF", we list the Variation Inflation Factors for the features, which identifies the filler features that are affected by multicollinearity and the strength of the correlation. We utilise the guideline that features with VIF > 10 cannot be used together in the RF model. Unsurprisingly, the pair of features $uh + um$, and $med + init$ have a $VIF > 10$, as each pair combined give us the total fillers num used in the dataset. We take this into account in the RF model, for example either utilising num or $uh + um$ for the basic features but not both.

The main experiments and their results are listed in Table 4.4. Please refer to Sec. 4.3 for the description of the filler features. In order to get further insight into which features seem relevant to our task, we utilise a RF feature importance, as shown in Figure 4.13.

[7]Following the guidelines as given in `https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/`, "Multicollinearity makes it hard to interpret your coefficients, and it reduces the power of your model to identify independent variables that are statistically significant".

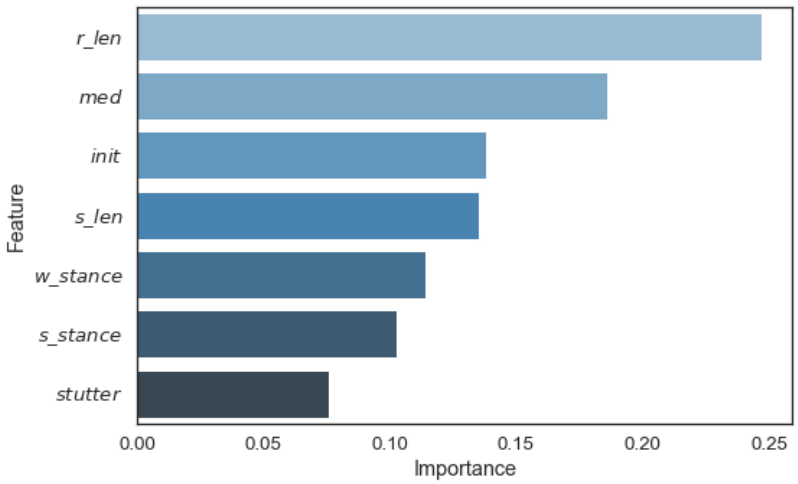


Figure 4.13: Top features calculated for the best RF model in Table 4.4, i.e. using only the implicature features.

In Table 4.4, the MSE for the models that use basic filler features, is lower than the RV baseline in predicting the the listener's impression of the speaker's expressed confidence. As shown, utilising basic filler features alone (i.e, the number of fillers in the review *num*, or the individual fillers $uh + um$) is sufficient to reasonably predict the final label of confidence ($MSE = 1.02 - 1.04$). We can see that the impact of adding the implicature features for both models, decreases the MSE. Looking at Figure 4.13, we see that the length of the review (r_len) is the most important feature for the RF model, followed by the position filler features, i.e. *med* and *init*. Both of these sets of features are features that relate to the planning and structure of the speech. What is interesting is that the features related to stance were ranked lower in importance for the RF model, despite the preliminary results observed in Figure 4.10. However, considering that both HC and LC speakers alike seem to use fillers more when asserting a weak stance/opinion than a strong one, it is plausible that it is not as strong a predictor of the rating of confidence but merely a characteristic of fillers itself. From the results given in Table 4.4, we can summarise that the length of the review is a good predictor of confidence, as well as the placement of the filler in the utterance, and the number of fillers in the review itself.

4.5 Conclusion

In this chapter, we presented a preliminary feature study on the role of fillers in the prediction of the listener's impression of the speaker's metacognitive state. The broad objectives in this chapter were i) to understand whether there is a link between the speaker's use of fillers (production context) and the listener's impression of the speaker's metacognitive state in more naturalistic forms of spontaneous speech, ii) to study whether there there are different functions of fillers that may influence the listener's impression differently and iii) to establish in an ML task the impact of fillers on the performance of SLU systems.

From RQ1, we can conclude that it is likely that there is a relationship between the number of fillers that were uttered by the speaker, and the listener's impression of the speaker's expressed confidence; i.e. the higher the fillers the lower the confidence. Here, we establish this in a large dataset that is not limited to QA contexts. Indeed the number of fillers used can contribute to the *overall* (global) impression the listener has of the speaker (and not only the utterance level – as was studied in previous research).

In similar broad SLU (/affective computing) work that utilises fillers as features, a common assumption may be that fillers contribute to disfluency so therefore they are not desirable in speech. An important finding of RQ1 is that not all fillers may contribute to the listener's impression of the speaker's confidence; i.e. while overall the higher the number of fillers the lower the confidence, it may not be the case for fillers if we tease them apart into different features. For example, the positional aspects of fillers could be more important in the final rating of confidence; particularly fillers that occur sentence medially compared to fillers that occur sentence initially. Listeners thus may find it useful when fillers are used sentence initially to give prosodic structure to the review. Generally, a bulk of fillers ($\approx 40-48\%$) occur sentence initially, regardless of HC/LC ratings (and indeed, this is consistent with many other works that find the distribution of fillers commonly at discourse boundaries (Shriberg, Bates, and Stolcke, 1997; Swerts, 1998; Shriberg, 2001; Swerts and Geluykens, 1994)). It also seems that stylistically, speaker's use more fillers when uttering opinions with a weaker stance compared to opinions with a stronger stance. This finding does not contradict findings on fillers and metacognition, indeed speakers may use fillers to tone down the force of their assertion/ opinion. However, this may occur independently of the global confidence rating the listener gives the speaker. Additionally, we found that the length of the review itself is a strong indicator of the final rating of confidence, with HC speakers in general being more "informative" in their

review. From our methodology, we cannot conclude whether the individual fillers "uh" and "um" affect the rating of confidence differently, despite recent research (Le Grezause, 2017) that shows they are not interchangeably used. In this dataset itself, the number of "uh" and "um" fillers are approximately the same.

Lastly, from RQ2 we show that fillers can be used as a feature in the prediction of the final rating of confidence; and are informative in such a context. This suggests that fillers are informative for SLU tasks that are to do with social communication, and further methodologies need to be developed in order to integrate them.

5 | Representation of Fillers in SOTA Language Models: Psycholinguistic Perspectives

5.1 Introduction and Related Work

The **aim of this chapter** (following the previous chapter), is to study the informativeness of fillers in the same dataset of spontaneous speech monologues, and assess their impact on SOTA SLU models – but here, instead of using hand-crafted filler features, learning contextualised representations of fillers using unsupervised methods. In this regard, we address the matter of representing fillers with deep contextualised word representations (Devlin et al., 2019). We also investigate *without hand-crafted features*, the information carried by fillers in targeted tasks based on the findings of the previous chapter.

For *mutual understanding*, fillers may be used *in production* by the speaker to inform the listener about their difficulties in selecting the appropriate vocabulary, to signal a less predictable word given the context (Shriberg et al., 1998), to mark discourse boundaries (such as (Swerts, 1998; Maclay and Osgood, 1959; Shriberg, 1994)), and thus, inform about overall lexical and syntactic choices that they make, A question remains as to whether this information could be leveraged in a Spoken Language Modelling (SLM) task using SOTA LMs. Thus in this chapter, we show that **fillers contain useful information that can be leveraged *specifically* by deep contextualised embeddings to better model spoken language**. We also study which filler representation strategies are best suited to our task of SLM and investigate the learnt positional distribution of fillers. Additionally, we show *without handcrafting features* on the same spontaneous speech corpus of monologues, that fillers are a discriminative feature in predicting the *perception* of expressed confidence of the speaker, and the *perception* of

a speaker's stance. In the following section, we briefly summarise the salient points from our previous results that we consider in the present work, and a brief overview of disfluencies and spontaneous speech in NLP systems. For a more detailed discussion (such as the vast research on fillers, particularly from a psycholinguistic perspective, and the links between disfluencies and metacognition/ mutual understanding) please refer to Part I.

Production and Perception From the previous chapter (see Chapter 4), we observed that there is a potential relationship between the average number of fillers that are uttered by the speaker in a monologue, and the listener's overall impression of the speaker's metacognitive state; i.e. the impression of the speaker's expressed confidence. We also could not conclude on whether the individual fillers "uh" and "um" are differentiated in use (despite recent work to suggest that this is the case (Le Grezause, 2017) and that they are not interchangeably used), and overall, the count of each filler in the dataset is roughly equivalent. Results also suggest that stylistically, i) regardless of the listener's impression of confidence, speakers produce a bulk of fillers sentence initially ($\approx 40-48\%$) compared to other locations in the review, (consistent with many works that study the distribution of fillers to find that they occur at discourse boundaries (Shriberg, Bates, and Stolcke, 1997; Swerts, 1998; Shriberg, 2001; Swerts and Geluykens, 1994)) and ii) speakers tend to use more fillers when asserting a weaker stance, compared to stronger stance opinions. The latter point confirms their potential in stance prediction, as was already shown by other researchers (Le Grezause, 2017; Levow et al., 2014), but on a different dataset and at an utterance level. A question remains as to whether fillers can be predictive of stance on a global level, given that many negative opinions could still lead to a positive review.

Spontaneous speech and disfluencies in NLP With the increasing popularity of voice assistant technologies and dialogue systems, NLU research often overlaps with SLU to consider the *textual processing of speech transcripts* (as discussed in Ruder (2020) as "Speech first-NLP"). However, when considering this automatic processing of text, **discrepancies may arise in the processing of speech transcripts compared to processing grammatically written text**, if utilising the same (NLP) systems. For e.g. Barriere, Clavel, and Essid (2017) showed that pre-trained word embeddings such as Word2vec (Mikolov et al., 2013), have poor representation of spontaneous speech phenomena such as "uh", as they are trained on written text and do not carry the same meaning as when used in speech. Specifically in an opinion mining task, they found that the representation of

"uh" might have a negative connotation arising from written text (where a filler may deliberately be used by a writer for example, to show hesitation of a character), compared to it being casually (and more neutrally) used in spontaneous speech. Tran et al. (2017a) present an attention-based encoder-decoder model for parsing conversational sentences. They obtain SOTA performance on parsing speech transcripts, which are further improved when including word level *acoustic-prosodic* features. An important finding from this work was that the integration of these acoustic-prosodic features showed the most gains over disfluent and longer sentences compared to fluent ones (though overall adding these acoustic-prosodic cues only lead to marginal improvements in the $F1$ score). This empirically shows a discrepancy between transcripts that are more "speech-like" (i.e. disfluent) compared to transcripts that are more grammatical and like written text (i.e. fluent). There is a need to further test SOTA NLP systems to determine performance on speech transcripts.

Previously, fillers and all **disfluencies were removed in pre-processing as noise**, as NLP models achieved highest accuracy on "fluent"/grammatical utterances (though with pre-trained word embeddings and depending on the task, this pre-processing strategy is not commonly used anymore). Recently Gupta et al. (2021), showed the role of disfluencies in the confusion of QA systems. Although they do acknowledge that disfluencies are indeed, ubiquitous to spontaneous speech and that NLP models are not robust to them, the intent of the work is on treating disfluencies as noise to the task. This contradicts psycholinguistic studies, which show that fillers play an informative role of planning and comprehension in spoken language (such as Clark and Fox Tree (2002) and Corley, MacGregor, and Donaldson (2007)). Indeed, our previous findings were based on the work of Brennan and Williams (1995) and Smith and Clark (1993), which showed that fillers can play a role in a speaker's estimation of their certainty, and the listener's estimation of the speaker's certainty; *specific to a QA context.*

So far, the specific information carried by fillers has only been studied using hand crafted features, for example in Le Grezause (2017) and Saini (2017). We utilised supervised methods and hand-crafted filler features in the previous chapter in our preliminary analysis. In terms of unsupervised methods for learning general characteristics of spontaneous speech, in recent work Tran et al. (2019b) showed that **using contextualised embeddings pre-trained on large written corpora, can be fine-tuned on smaller spontaneous speech datasets** to improve constituency parsing on conversational speech transcripts. It is worth noting that constituency parsing attributes *structure* to spontaneous speech transcripts (i.e. annotated punctuation, input being pre-segmented sentences ...). Indeed, in other

spontaneous speech tasks such as joint dialogue act segmentation and classification[1] in Zhao and Kawahara (2019b), they choose not to initialise word embeddings with pretrained GloVe vectors (Pennington, Socher, and Manning, 2014) due to degraded performance when doing so. Tran et al. (2019b) also found that fine-tuning the GloVe embeddings on a spontaneous speech corpus, before being used in the conversational parsing task, results in a negligible difference when compared to using original GloVe embeddings. The type of task/embedding plays a role in the representation of spontaneous speech when using pretrained word embeddings, with results being mixed. To the best of our knowledge, **we are the first to specifically investigate the learned representations of fillers using deep contextualised word representations**.

In order to understand existing representations of spontaneous speech phenomena in SOTA NLP systems, **tasks designed to test *specific* characteristics of spontaneous speech phenomena (such as fillers) are required** – rather than broader and more generalised observations about "learning representations of spontaneous speech". The intent here differs, as it allows for a targeted comparison between the linguistic capabilities of deep contextualised word representations and our knowledge of human behaviour, rather than a focus on beating SOTA NLP systems. Thus we focus specifically on the informativeness of fillers, and observe the impact based on psycholinguistic theories of disfluencies (please see the research questions in Sec. 5.2). The rest of the chapter is organised as follows: In Sec. 5.2, we describe our research questions, dataset and methodology. Sec. 5.3 gives the results and discussion, while Sec. 5.4 discusses the conclusion of the experiments, and the global conclusions of Part II.

5.2 Research Questions and Methodology

Research Questions Thus the present chapter is **motivated by the following observations:** i) Fillers play an important role in spoken language. From a psycholinguistic perspective, speakers can use fillers to inform the listener about the linguistic structure of their utterance, such as in their *(difficulties of) selection of appropriate vocabulary* while informing the listener about a pause in their upcoming speech stream. ii) Fillers and prosodic cues have also been linked to a speaker's *Feeling of Knowing* or *expressed confidence*, that is, a speaker's certainty or commitment to a statement (Smith and

[1] Where the input is the verbatim transcripts by speaker turn without any punctuation, and the output is the speaker turn segmented according to dialogue acts within the turn, and also classification of the dialogue act.

Clark, 1993). Brennan and Williams (1995) observed that fillers and prosodic cues contribute to the listener's perception of the speaker's expressed confidence in their utterance, which they refer to as the *Feeling of Another's Knowing*. In the previous chapter, we investigated this in an ML context, and showed using linear models and our hand-crafted textual filler features, that fillers can be a discriminative feature in the prediction of the listener's overall impression. iii) Recent work has shown that fillers have been successful in *stance* prediction (stance referring to the subjective spoken attitude towards something (Haddington, 2004)) (Le Grezause, 2017). Our previous results suggest link between fillers and stance, with a tendency for speakers to use more fillers when asserting a weaker stance, compared to a stronger one. We want to verify that our observations are still valid when we represent fillers in an automatic and efficient way. Hence, our research questions are as follows:

- **RQ1: (Production) Can the information contained by fillers be leveraged to model spoken language?**
- **RQ2: (Perception) Can fillers be a discriminate feature specifically in the prediction of confidence and in stance prediction?**

Dataset Please refer to Chapter 4 for the relevant points pertaining to the dataset used. We use the transcripts from the POM dataset as our textual input to the models, and compute the same RMS score as the final label of confidence as described. In addition to this, we use labels of stance. For stance, the annotators were asked "How would you rate the sentiment expressed by the reviewer towards this movie?", and were asked to give a label from 1 (strongly negative) to 7 (strongly positive). For stance labels, we simply take the average of the three annotator scores as our final label of stance (as unlike the confidence labels, the stance labels are not skewed). For a full description of the dataset, please refer to Chapter 3.

Models

For our work, we consider the two fillers "uh" and "um". To obtain contextualised word embeddings for fillers, we use bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019), as it has achieved SOTA performance on several NLP benchmarks and (as discussed previously), are better than Word2Vec for word sense disambiguation by integrating context (Bartunov et al., 2015).

Token.	Output Tokenizer
Raw	(um) Things that (uh) you usually wouldn't find funny were in this movie.
$\mathcal{T}_1$	['um', 'things', 'that', 'uh', 'you', 'usually', 'wouldn', "", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']
$\mathcal{T}_2$	['$[FILLER_{UM}]$', 'things', 'that', '$[FILLER_{UH}]$', 'you', 'usually', 'wouldn', "", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']
$\mathcal{T}_3$	['[FILLER]', 'things', 'that', '[FILLER]', 'you', 'usually', 'wouldn', "", 't', 'find', 'funny', 'were', 'in', 'this', 'movie', '.']

Table 5.1: Filler representation using different token representation strategies.

RQ1: (Production) Spoken Language Modelling For SLM, we use the masked language modelling objective (MLM). It consists of masking some words of the input tokens at random, and then predicting these masked tokens. The MLM objective is classically used to pre-train and then fine-tune BERT. Here, we use this MLM objective to fine-tune a pretrained BERT on our spoken language corpus of monologues. Each experiment requires a token representation strategy $\mathcal{T}_i$ and a pre-processing strategy $\mathcal{P}_{Si}$ (please refer to algorithm 1 for the procedure used for the language modelling task).

The **token representation strategies** are particularly important for our task, for BERT to learn the distribution of fillers. The three token representation strategies ($\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$), are described as follows: In $\mathcal{T}_1$, no special treatment is done to the fillers, i.e BERT will use its a priori knowledge of the fillers "uh" or "um" to model the language. In $\mathcal{T}_2$, "uh" and "um" are distinguished from other tokens by a special filler tag, and are represented as two different tokens respectively; this strategy aims at forcing BERT to learn a new embedding that focuses both on the position and the context of the fillers. Setting a special filler tag in BERT is motivated by trying to learn in an unsupervised manner the hand-crafted features we proposed in the previous chapter (Chapter 4). For e.g., we specifically defined *init* and *med* to focus on sentence-initial and sentence medial fillers (position attributes), or *uh* and *um* to distinguish between the two fillers. In $\mathcal{T}_3$, both fillers are represented as the same token, suggesting that they have the same implicit meaning(s) and are interchangeable. A concrete example is given in Table 5.1.

Pre-processing strategies, ($\mathcal{P}_{S1}, \mathcal{P}_{S2}, \mathcal{P}_{S3}$), are as follows: In $\mathcal{P}_{S1}$, the sentences have all fillers removed, both during training and inference. In $\mathcal{P}_{S2}$, the sentences have the fillers kept during training, but are removed at

Algorithm 1: Spoken Language Modelling

Input : $\mathcal{P}_{Si}$, $\mathcal{T}_i$, Pret. BERT $\mathcal{LM}$
Output: $(\mathcal{LM}, Perplexity)$
$(\mathcal{D}_{train}, \mathcal{D}_{dev}, \mathcal{D}_{test}) \leftarrow$ (train, dev, test set) according to $(\mathcal{P}_{Si}, \mathcal{T}_i)$
if *Do Finetuning* **then**
  $\mathcal{LM} \leftarrow \mathcal{LM}(\mathcal{D}_{train})$ using (MLM).
Evaluate: $Perplexity \leftarrow \mathcal{LM}$ on $\mathcal{D}_{test}$

inference. In $\mathcal{P}_{S3}$, the fillers are kept both during training and inference. For each pre-processing and token representation strategy, we optionally fine-tune BERT using the same Masked Language Model (MLM) objective as in the original paper Devlin et al., 2019. Note, if we do not fine-tune, the training dataset ($\mathcal{D}_{train}$) is not used and therefore $\mathcal{P}_{S1}$ and $\mathcal{P}_{S2}$ are equivalent. For language modelling we report the perplexity (*ppl*) measure to evaluate the quality of the model.

Algorithm 2: Confidence prediction

Input : $\mathcal{P}_{Si}$, $\mathcal{T}_i$, $\mathcal{LM}$ from algorithm 1
Output: $(CONF_p, MSE)$
$(\mathcal{D}^{labelled}_{train}, \mathcal{D}^{labelled}_{dev}, \mathcal{D}^{labelled}_{test}) \leftarrow$ (train, dev, test set) according to $(\mathcal{P}_{Si}, \mathcal{T}_i)$
$CONF_p \leftarrow \mathcal{LM} + MLP$
$CONF_p \leftarrow CONF_p(\mathcal{D}^{labelled}_{train})$ using (MSE).
Evaluate: $MSE \leftarrow CONF_p$ on $\mathcal{D}_{test}$

RQ2: (Perception) Confidence and Stance Prediction In both our confidence prediction and stance prediction task, our goal is to predict a label of confidence/stance using our BERT text representations that include fillers. Formally, our confidence/stance predictor is obtained by adding a Multi-Layer Perceptron (MLP) on top of a BERT, which has been optionally fine-tuned using the MLM objective. Please refer to Figure 5.1 for a diagrammatic representation of our confidence/stance predictor with (W) and without (W/O) the MLM fine-tuning. The MLP is trained by minimising the mean squared error (MSE) loss. We keep the same token representation and pre-processing strategies from the previous section. Please refer to algorithm 2 for the procedure used for the confidence prediction task, which follows the same procedure for the stance prediction task.

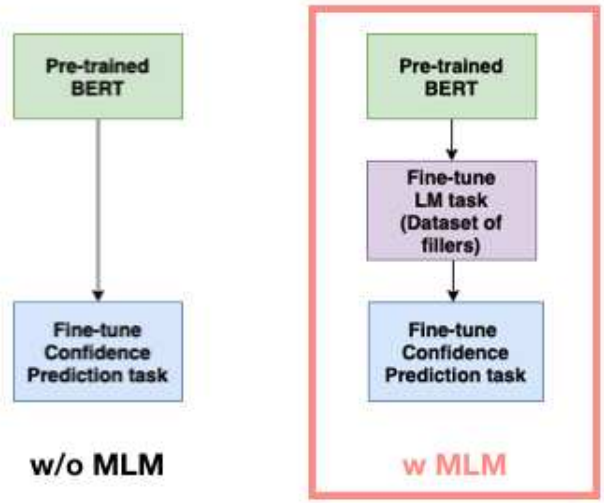


Figure 5.1: A diagrammatic representation of our confidence/stance predictor with (W) and without (W/O) the MLM fine-tuning

5.3 Results and Discussion

5.3.1 RQ1: (Production) Can the information contained by fillers be leveraged to model spoken language?

Language Modelling with fillers We compare the perplexity of the LM with different pre-processing strategies with a fixed token representation $\mathcal{T}_1$ (i.e no special treatment done to the fillers, in order for BERT to use its a priori knowledge of fillers). Firstly for $\mathcal{T}_1$, it is interesting to note that BERT provides embedding for "uh" and "um" despite being trained on written text (Wikipedia, BooksCorpus Zhu et al., 2015, Word Benchmark Chelba et al., 2014). Results are reported in Table 5.2(a). We compare $\mathcal{P}_{S1}$ (fillers removed in training and inference), $\mathcal{P}_{S2}$ (fillers kept in training, removed in inference), and $\mathcal{P}_{S3}$ (fillers kept, training and inference) with or without fine-tuning and observe that adding fillers, both during training and inference, leads to a model with lower perplexity and a perplexity reduction of at least 10%. Hence, fillers contain information that can be leveraged by BERT.

As shown, the fine-tuning procedure reduces the perplexity of the language model. Even without fine-tuning, we observe that $\mathcal{P}_{S3}$ outperforms $\mathcal{P}_{S1}/\mathcal{P}_{S2}$, as the perplexity reduces when adding fillers. This suggests that BERT has some a priori knowledge of spoken language, in terms of fillers,

and shows some potential for generalisation of the results on more complex data (such as dialogues). Hence, fillers can be leveraged to reduce uncertainty of BERT for SLM. This is not an expected result, as intuitively, one might think that the perplexity would reduce when fillers are excluded from both training and inference, due to the fact that the utterance is shorter and "simplified". The fact that $\mathcal{P}_{S3}$ outperforms the other pre-processing methods also suggests that the MLM procedure is an effective way to learn this information.

Our results on current SOTA LMs are a nod to Stolcke and Shriberg (1996), who found that the addition of fillers using *ngram* LMs leads a reduction in perplexity, due to the information fillers can provide regarding the neighbouring words to the right (with the language being English). From a psycholinguistic perspective, speakers can produce fillers to inform the listener about lexical choices that are made, which often are be less predictable words given the context (Beattie and Butterworth, 1979). Interestingly, listeners can also interpret fillers as such. For e.g., Corley, MacGregor, and Donaldson (2007) studied the effect of hesitation ("um") on the listener's comprehension using the $N400$ function of an Event-related potential (ERP), which they establish in predictable versus unpredictable words. The $N400$ effect can be observed during language comprehension, typically occurring 400 ms after the word onset; and exhibits a negative charge recorded at the scalp consequent to hearing an *unpredictable word*. In using hesitations preceding the unpredictable word, the $N400$ effect in listeners was *visibly* reduced. Indeed, our results using BERT are exciting in that speakers may produce fillers this way, and in terms of comprehension, fillers may be used as a signal for neighbouring words to the right.

Best Token representation We observe that $\mathcal{T}_1$ (no special treatment done to the fillers) outperforms the other representations in a fine-tuning setting, as shown in Table 5.2(b). Given the restricted size of our data and the dimension of the BERT embeddings (768), it is better to keep the existing representations (with $\mathcal{T}_1$), than adding and learning new representations from scratch. $\mathcal{T}_2$ ("uh" and "um" distinguished from other tokens and each other by a special token each) and $\mathcal{T}_3$ (fillers represented by a same, special token) perform the same. Results are mixed in terms of whether the two fillers could be distinguished. For e.g., Clark and Fox Tree (2002) hypothesised that "um" and "uh" are only distinguished in duration, i.e. that "uh" is used for a shorter pause in speech; which cannot not be reflected in text. However, other studies contradict this, for example even recent works in Le Grezause (2017) (and in subsequent chapters, we show that there is a distinction between the ways

the two fillers are used). Thus while Tran et al. (2019b) show that using contextualised embeddings pre-trained on large written corpora, can be fine-tuned on smaller spontaneous speech datasets, results suggest that it can be difficult to learn a different representation to distinguish the two fillers than an already existing one. Given these results, we fix $\mathcal{T}_1$ as the token representation strategy for the rest of the experiments.

Learnt Positional distribution of fillers: Given our findings in the previous chapter (i.e. speakers use a bulk of fillers sentence initially ($\approx 40-48\%$) compared to other locations in the review), we additionally test whether our model has learnt information about the placement of fillers. The aim is to see whether BERT can learn the positional distribution of fillers in an unsupervised manner, and this experiment can offer an insight into what the model learnt. Formally, we use fine-tuned BERT on $\mathcal{D}_{train}$ with fillers to see where the model estimates the most probable position of the fillers (which we call $\mathcal{LM}_{fillers}$) to be. Given a sentence S of length L, we insert after word j the mask token ('[MASK]') to obtain the corrupted sentence $\widetilde{S}$[2]. We compute the probability of the appearance of a filler in position $j+1$ according to the LM, which corresponds to $P([MASK] = filler|\widetilde{S})$, as illustrated by Figure 5.2. Formally, we plot the average of the probability of the masked word to be a filler given its position in the sentence, as shown in Figure 5.3. We observe that the fine-tuned BERT on $\mathcal{D}_{train}$ with fillers ($\mathcal{LM}_{fillers}$) predicts with high probability fillers occurring at the first position in the sentence (please refer to Table 5.3 for example sentences). This is consistent with the actual distribution of fillers in the dataset, as can be seen in Figure 5.3. The fine-tuned BERT on $\mathcal{D}_{train}$ without fillers ($\mathcal{LM}_{nofillers}$) predicts a constant low probability. Given the available segmentation of sentence boundaries (fine-grained discourse annotations are not available), it is interesting to note that our model was able to capture similar positional distribution of fillers that occur sentence initially. However, for sentence-medial fillers, which also can occur at natural discourse boundaries (for e.g. where the transcriber may put punctuation such as a comma), our model differs from the distribution in the dataset.

Thus in this section we show that although BERT uses contextualised word embeddings, the information contained in fillers can be leveraged to achieve a better modelling of spoken language.

[2]For clarity we abuse the notation and remove dependence in j.

Fine.	Setting	Token.	Ppl
w/o	$\mathcal{P}_{S1}$	$\mathcal{T}_1$	22
	$\mathcal{P}_{S2}$	$\mathcal{T}_1$	22
	$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**20**
w	$\mathcal{P}_{S1}$	$\mathcal{T}_1$	5.5
	$\mathcal{P}_{S2}$	$\mathcal{T}_1$	5.6
	$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**4.6**

(a)

Setting	Token.	Ppl
$\mathcal{P}_{S3}$	$\mathcal{T}_1$	**4.6**
$\mathcal{P}_{S3}$	$\mathcal{T}_2$	4.7
$\mathcal{P}_{S3}$	$\mathcal{T}_3$	4.7

(b)

Fine.	Model	Conf.	St.
w/o	$\mathcal{P}_{S1}$	1.47	1.98
	$\mathcal{P}_{S2}$	1.45	1.75
	$\mathcal{P}_{S3}$	**1.30**	**1.44**
w	$\mathcal{P}_{S1}$	1.32	1.39
	$\mathcal{P}_{S2}$	1.31	1.40
	$\mathcal{P}_{S3}$	**1.24**	**1.22**

(c)

Table 5.2: From left to right, the (a) LM Task, (b) Best token representation, (c) MSE of Confidence (Conf.) and the Stance (St.) prediction task. Wilcoxon test (10 runs with different seeds) has been performed. Highlighted results exhibit significant differences (p-value < 0.005). Data split is fixed according to Zadeh (*CMU-Multimodal SDK*) and results are given on the test set (see supplementary materials for for additional details).

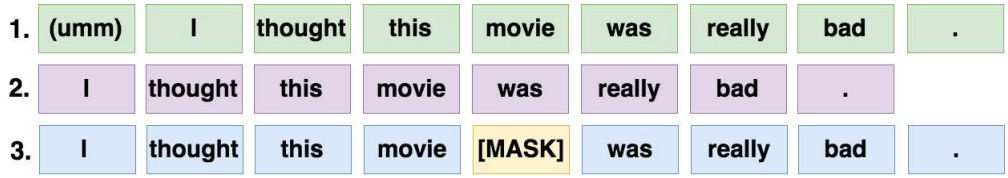


Figure 5.2: Predicting the probability of a filler, where 1. Raw input, 2. Pre-processed text with the filler removed, and 3. Illustrates the [MASK] procedure for predicting the probability of a filler at position 5.

5.3.2 RQ2: (Perception) Can fillers be a discriminate feature specifically in the prediction of confidence and in stance prediction?

We observe the impact that fillers have on two downstream tasks, the prediction of the listener's impression of the speaker's expressed confidence, and the prediction of stance. Psycholinguistic studies have observed the link between fillers and expressed confidence Smith and Clark, 1993; Brennan and Williams, 1995; Wollermann et al., 2013. Previous research on the link between fillers and their relation to a speaker's expressed confidence has been confined to a narrow range of QA tasks Schrank and Schuppler, 2015. Fillers have also been linked to stance prediction (Le Grezause, 2017), which we measure using sentiment labels provided in the dataset. We show that in a spontaneous speech corpus of spoken monologues, fillers can play a role in predicting both the perception of the speaker's expressed confidence and speaker's stance.

In Table 5.2(c) we observe that both with and without fine-tuning the $\mathcal{P}_{S3}$

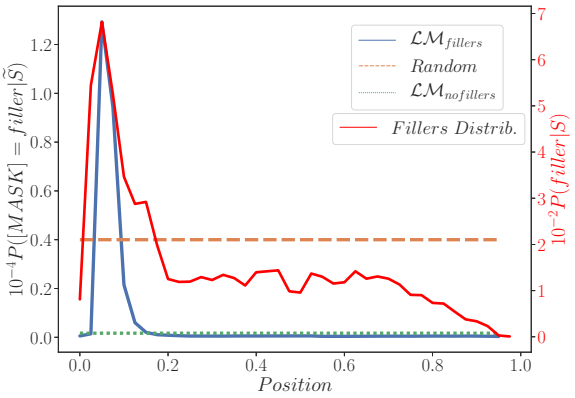


Figure 5.3: Predicting the position of fillers. *Fillers Distrib.* stands for the actual filler distribution in the dataset. *Random* stands for the random predictor which predicts $P([MASK] = filler|\widetilde{S}) = \frac{2}{|\mathcal{V}|}$ where $|\mathcal{V}|$ is the size of the vocabulary, and 2 represents both fillers.

(um) the title actually translates to The Brotherhood of War.
(um) The movie itself is a lot like Saving Private Ryan and Band of Brothers.
(um) You'll only like it if you're into kid of strange, bizarre humor.
It's just **(uh)** pretty obvious stuff you know.
But **(um)** a lot of the movie didn't really make sense.
(um) It's really funny, there there's (stutter) some really funny parts in it.
(um) But, I recommend watching this movie it's really good.
(um) The acting is only so-so.
(uh) Morgan Freeman is great in this movie, and **(uh)** so is Tim Robbins.
And so **(um)** I wouldn't really recommend it.
(um) Yeah, but that's it.

Table 5.3: Some samples from the dataset. As can be seen, many of the fillers occur sentence-initially.

decreases the MSE compared to $\mathcal{P}_{S1}$ and $\mathcal{P}_{S2}$. We observe that $\mathcal{P}_{S2}$ leads to higher MSE, possibly because of the discrepancy created between $\mathcal{D}_{train}^{labelled}$ and $\mathcal{D}_{test}^{labelled}$. Particularly of note, is that fillers can contribute to both to overall *global* and fine-grained *local* stance prediction – as in the previous chapter, we noticed in general, a stylistic tendency for fillers to be used when uttering weaker stance positions, which was also researched in Le Grezause (2017). This shows that fillers (with their *implicit* meanings) can be a discriminative feature in both the perception of the speaker's expressed confidence and stance prediction, apart from overt lexical cues (words that explicitly express

uncertainty/confidence, such as *maybe, I'm unsure* or sentiment, *amazing, disgusting*)).

Our hand-crafted filler features used as input to our linear models achieved lower MSE scores than BERT did in the downstream task of confidence prediction (with lowest MSE of 0.95 compared to lowest MSE of 1.24). The most plausible explanation for this is that the video input is given all at once to then predict the final label of confidence. That is, the model (bert-Base from the Hugging Face library) takes the CLS token (classification token) as input first, followed by a sequence of words as input (all at once). The *the maximum sequence length of the input here is 512 tokens*. This means that some video transcripts needed to be truncated before being used as input to bert-Base. As we discussed in the previous chapter, while fillers may be one discriminative feature of the listener's perception, it is *not the only discriminative feature*. The *length of the review* (r_len) is a strong indicator of the final rating of confidence, with HC speakers in general being more "informative" in their review. However, the overall goal of observing the informativeness of fillers in an unsupervised way, as well as tested the existing representations was achieved.

Does the addition of fillers always improve the results for downstream spoken language tasks? In the SSec. 5.3.1, we show that by including fillers, the MLM achieves a lower perplexity. An assumption one could make based on the work by Radford et al. (2019), is that with this model, the results for any further downstream task would be improved by the presence of fillers. However, we observe that to predict the persuasiveness of the speaker (using the high level attribute of persuasiveness annotated in the dataset Park et al., 2014), following the same procedure as outlined in algorithm 4, that fillers, in fact, are not a discriminative feature.

5.4 Conclusion

Thus in this chapter, we studied the representations of fillers using deep contextualised word embeddings. This is an important issue, as these models are pre-trained on massive amounts of written text, and require methodologies to further study the representations of spontaneous speech that are learnt during fine-tuning. To the best of our knowledge, **we are the first to *specifically* study the representations of fillers in deep contextualised word embeddings**. Indeed what was interesting to us, was the discovery that BERT already has an existing representation of fillers. We studied fillers both from a production standpoint (i.e., using a SLM task), and fillers from a comprehension standpoint (i.e. the the downstream tasks

of the prediction of perception of i) stance and ii) metacognitive state).

We develop methodologies without hand-crafted features, to show that fillers reduce the uncertainty of an LM in a downstream SLM task, showing that they do not need to blindly be removed as noise when modelling spoken language. To do so, we compared *which* token representation strategies and pre-processing strategies are best suited to model fillers on our dataset of spoken transcripts. Additionally, we compared the representations learnt of fillers on two downstream tasks; to show that fillers are a *discriminative* feature in both these tasks, without hand-crafted filler features. We showed that the better methodology for learning representations of fillers is by first doing MLM fine-tuning on the dataset of speech transcripts.

We also offered **suggestions for where improvement is needed in the representations of fillers**, based on our experimental results. In the previous chapter, we firstly showed that in terms of production, a bulk of fillers ($\approx 40-48\%$) occur sentence initially, regardless of listener ratings (and indeed, this is consistent with many other works that find the distribution of fillers commonly at discourse boundaries (Swerts, 1998)). When we plotted the positional distribution of fillers using our model that had learnt the representations of fillers, we found that the model can place sentence-initial fillers correctly, and then struggles with the placement of sentence-medial fillers. Furthermore, we showed that BERT was unable to distinguish between the two fillers "uh" and "um" despite our findings and others (Le Grezause, 2017) to show that they are used differently (as we find in subsequent chapters).

Thus, production wise, BERT may be able to take advantage of fillers to inform about context to the right in a SLM task. However, in a downstream task prediction of a speaker's metacognitive state, BERT may be unable to learn of certain nuances of the way fillers are used. An important finding from the previous chapter is that **not all fillers may contribute equally to the listener's perception of the speaker**. Using our hand-crafted crafted features, we observed that the the positional aspects of fillers could be more important in the final rating from the listener; particularly fillers that occur sentence medially compared to fillers that occur sentence initially. Listeners thus may find it useful when fillers are used sentence initially to give prosodic structure to the review. While this may be difficult to capture in speech transcripts, it is overall useful to understand the limitations of these models.

III | The Local Level Interaction of Fillers with the Primary Signal in Monologues

6 | Statistical Analysis: Fillers in the Process of Information Sharing in the Flow of the Discourse

6.1 Introduction and Background

In this chapter, we would like to explore the *contextual* use of fillers, specifically in relation to the *primary signal*, or what was said in essence. In this regard, we investigate how fillers interact with the primary signal at an utterance level, and whether this can help in understanding/ interpreting the perception of the listener at a discourse level. **We treat the message from the speaker as *an incoming source of information that builds up in the discourse*, and then we *observe the function of fillers in this process***. Thus, we computationally study at a micro-level, which fillers are informative in terms of the new information specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception. We base this work on the numerous psycholinguistic perspectives that state that the listener can treat fillers as informative signals with regard to the incoming message (e.g. Corley, MacGregor, and Donaldson (2007), Barr and Seyfeddinipur (2010), and Arnold et al. (2004)).

In Part II, we studied the different production contexts of fillers, for e.g. their positional distribution in sentences, their role in strong and weak stance utterances and so on. In this chapter, using the same dataset of monologues, we *explicitly focus on the use of fillers as a collateral signal to the primary signal*. Previously, most features we computed were at a macro-level. However, we need to develop methodologies that will allow us to consider the context of how utterances build to the discourse level, rather than considering utterances as if they occur in isolation.

According to Brennan and Williams (1995), the listener's interpretation of the speaker's utterance includes estimates about the speaker's commit-

ment to/ expressed confidence in what they are saying. When considering the comprehension of disfluent speech for e.g., research has linked fillers to the listener's assessment of a speaker's metacognitive state (Brennan and Williams, 1995). We showed previously that, indeed these results can apply to spontaneous speech datasets collected in real-life contexts (or non-QA datasets). We concluded that it is likely that there is a relationship between the average number of fillers that were uttered by the speaker, and the listener's impression of the speaker's metacognitive state (and at more extreme cases, the higher the fillers the lower the confidence and vice versa). We also accounted for overall nuances of this, for e.g. fillers that occur sentence-medially compared to sentence-initially seem to be associated more with the rating the listener gives the speaker. However, if the listener can perceive fillers to be informative signals with regard to the incoming message (including helping the listener estimate the overall level of knowledgeability the speaker holds), a question remains as to **how fillers used in the context of incoming information *can lead to* the listener's final estimation of the speaker's knowledgeability**. Existing works also do not focus on the connection between the granularities of discourse; i.e. how does this micro-level relate to the macro-level.

Thus, the aim of this work is to study how does a speaker's use of fillers relate to the incoming message from the speaker, and consequently, how does that relate to a listener's perception of the speaker. Our findings suggest that speakers stylistically do tend to use fillers in the incoming message, when introducing a new entity (to indicate new information), rather than an entity already introduced into the discourse. Our results also suggest that the occurrence of fillers specifically before new entities may not have an effect on the listener's perception of the speaker's expressed confidence, despite previous works that suggest the link between fillers and expressed confidence. This does not discount other possible metacognitive aspects, such as the listener may expect a speaker to use fillers typically when the speaker is introducing new information in the incoming message. We also find that the filler "um" seems to be used more in the context of incoming information, compared to the filler 'uh", suggesting that they have different roles in the discourse. The rest of the chapter is organised as follows: for the rest of this section, we briefly overview the theoretical foundations and research questions of our study[1], Sec. 6.2 describes the our research questions and methodology, Sec. 6.3, the results and discussions of the work, and Sec. 6.4, the conclusion.

[1]Please see Chapter 2 for a detailed discussion of this.

Metacognition and the listener's perspective When a speaker says an utterance, this articulation process includes an estimation of their commitment/ certainty about what they are saying. Research suggests that fillers and prosodic cues are linked to a speaker's metacognitive state, specifically; their *Feeling of Knowing (FOK)* or **expressed confidence** — a speaker's certainty or commitment to a statement (Smith and Clark, 1993). A speaker may *encode* meaning into their utterance using fillers, but the onus is on the listener to *decode* this information; making the interpretation of fillers contextual and dependent on the listener. Brennan and Williams (1995) observed that fillers and prosodic cues contribute to the listener's perception of the speaker's metacognitive state; which they refer to as the *Feeling of Another's Knowing (FOAK).*

Other studies also focus on the comprehension of disfluent speech, i.e. taking into account the listener's understanding of the speaker's disfluencies (Corley and Stewart, 2008), and not on why the disfluency itself was produced (Nicholson, 2007). For example, Vasilescu, Rosset, and Adda-Decker (2010a) observe that the French "euh" has both *disfluent* (signalling production difficulties of the speaker) and *fluent* (as a discourse marker – to bracket lexical units that may aid in listener comprehension) properties. However, *both* these uses of disfluencies are informative to the listener in different ways. Related to fillers serving as a *cue for incoming information*, research suggests that following fillers, listeners may expect a speaker to shift topics, as they carry information about larger topical units (Swerts, 1998), that the use of fillers biases listeners towards new referents rather than ones already introduced into the discourse (Arnold et al., 2004), relax listener's expectations when hearing an unpredictable word (Corley, MacGregor, and Donaldson, 2007), and that listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle) (Barr and Seyfeddinipur, 2010). In the present paper, we focus on the listener's comprehension of disfluencies. As Corley and Stewart (2008) state, "it is hard to determine the reason that a speaker is disfluent, especially if the investigation is carried out after the fact from a corpus of recorded speech". Thus we analyse the speaker's use of fillers from the incoming message and then observe what effect this may have on the listener's perception.

Limitations of our current findings In Part II, we found that in an unsupervised manner, that fillers can be a discriminative feature in the automatic prediction of a listener's impression of the speaker. These results empirically solidified an effect that was often *assumed* to be true (and in-

deed, fillers are sometimes interchangeably used with the term "hesitations" in certain works (Pickett, 2018; Corley and Stewart, 2008; Maclay and Osgood, 1959)). However, we focused on the overall impression the listener had of the speaker, i.e. the global, and did not account for more fine-grained information shared by the speaker. Furthermore, our methodology using deep contextualised word embeddings could not distinguish between the two fillers "uh" and "um". When we plotted the positional distribution of fillers, we saw the tendency of the model to predict with a higher probability that a filler would occur sentence-initially, but then struggled with placements of fillers sentence-medially. This is not completely unexpected, as fillers can have *unrestricted distribution* according to Meteer et al. (1995). It however, merits further investigation, as previous studies that used *ngram* language models discussed how fillers can inform about context to the right (Stolcke and Shriberg, 1996) – by informing models about upcoming words that have low contextual probability. Our work on SOTA deep contextualised word embeddings, showed also, that fillers can reduce the uncertainty of a LM in a SLM task. There is a strong unifying thread with all these studies (and psycholinguistic perspectives), that fillers provide some contextual cues to upcoming information, and listeners *perceive them as such.*

6.2 Research Questions and Methodology

We previously showed, both using hand-crafted features, and in an unsupervised way, that fillers can be a discriminative feature in the the *prediction* of a listener's perception. While the work from the previous chapters are important as preliminary analysis, they do not account for how fillers locally interact with the rest of the message in a *holistic* way.

Research Questions

Clark (1996) and Clark and Fox Tree (2002) proposed that speakers are able to utilise fillers as *collateral signals* in communication, in addition to the *primary signal* of the message. We colloquially refer to the primary signal of the message as *what* was said (in essence) and the collateral signal as *how* it was said. In Spoken Language Understanding (SLU), a similar phenomenon occurs of separating these two signals. However, in this context, reducing an input utterance into its primary signal (or *what* was said in essence) is standard practice (e.g. as seen in Tur and De Mori (2011), chapter 13. Speech Summarization). Indeed, in dialogue systems, the output transcripts of automatic speech recognisers are often cleaned of disfluencies such as fillers

in post-processing, despite the rich linguistic literature to suggest otherwise (Clark and Fox Tree, 2002). And yet, even recent work such as Barr and Seyfeddinipur (2010) support the collateral signal account, specifically that the listener is able to process fillers as a collateral signal (even if unclear whether the speaker (un)intentionally used them as such). This is an important finding, as it shows that perhaps the listener's attention is drawn to the cognitive state of the speaker. The problem then, as stated in Clark and Fox Tree (2002), remains about how to merge the two signals. Given the rapid advancements of dialogue systems, and growing interest in SLU, there is a need to move towards an automatic but holistic analysis of both together; if we hope to move towards better models and understanding of spontaneous speech. Thus the research questions are as follows:

RQ1: (Local effect of fillers): How does a speaker's use of fillers relate to the incoming information the speaker is sharing From the findings of Barr and Seyfeddinipur (2010) and Arnold et al. (2004) as discussed in Sec. 6.1, we would like to empirically analyse the role fillers play in a dataset of spontaneous speech, specifically related to new information from the incoming message of the speaker. Since the dataset we choose to study is a dataset of English monologue movie review videos (please refer to Chapter 3), we consider the speaker's mention of terms related to the movie annotated from metadata, such as actors and directors.

- **H1 Fillers are more likely to occur before the introduction of new and upcoming information in the review.**

RQ2: (Global effect of fillers): How does the speaker's production of fillers build to the listener's perception of the speaker? We would like to empirically analyse whether the speaker's production of fillers has an impact on the listener's overall impression of the speaker.

- **H2 From H1, the speaker's use of fillers preceding new information in the incoming message is associated with the listener's perception of the speaker's confidence.**

  Specifically, we hypothesise that when fillers are predominantly used in the context of preceding new information, listener's may judge the expressed confidence of the speaker as high, and listeners may only notice when fillers are used in other contexts (for e.g. as seen in Tottie (2014), listeners notice fillers when they are overused or used in the wrong context) which consequently will decrease the expressed confidence rating.

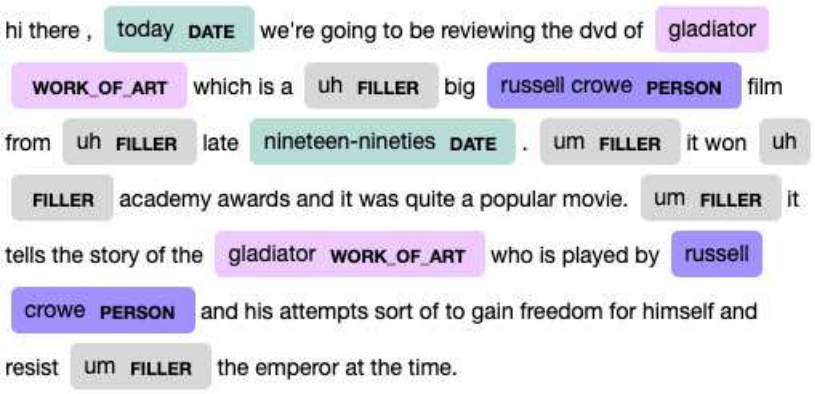


Figure 6.1: An example transcript that has annotated entities (in colour) using the EntityRuler. As shown, patterns from the metadata (e.g. "russell crowe") are added to the existing set (e.g. "nineteen-nineties"). Fillers are marked in grey. The first mention of "russell crowe" would be considered a new entity mentioned, while the second, an old one. Note, while the entity annotation is fairly reliable given the metadata, it is not exact. For e.g. the EntityRuler sometimes mislabels entities (the second mention of the word "gladiator".)

Dataset Please refer to Chapter 4 for the relevant points pertaining to the dataset used. We use the transcripts from the POM dataset as our textual input to the models, but this time, take the average score of the three annotators as our final label of confidence as described. For a full description of the dataset, please refer to Chapter 3.

Methodology

RQ1: (Local effect of fillers) How does a speaker's use of fillers relate to the incoming information the speaker is sharing We consider the speaker's mention of entities related to the movie, that we extract from metadata files[2]. These entities could be categorised into actor, director or title of the movie. We then add these custom entities to SpaCy's EntityRuler, a rule based named entity recogniser[3]. We prepossess the files (e.g. so that the filler annotations match the fillers in the existing model's vocabulary). We map the entities to match the existing patterns in the EntityRuler, for e.g "actor" is converted to "PERSON", by adding to the already existing entity

[2]The complete code and processed data will be made available online for reproducibility here `https://github.com/tdinkar/fillers_in_POM.git`

[3]`https://spacy.io/api/entityruler`

patterns (please refer to Figure 6.1). The tagging of entities follows the *BIO* format (beginning, inside and outside of an entity).

To investigate H1, we inspect for each transcript, the distribution of filler positions, in relation to the automatically annotated entities in the discourse (denoted by *Ent*). We split these entities into *Ent_new*; i.e. entities newly introduced in the discourse, to indicate new information in the incoming message, and *Ent_old* to indicate entities already introduced in the discourse. We specifically note the order of the tokens in the transcripts for the filler positions and the first token of the 1. *Ent_new* (the first occurrence of the *Ent*) and 2. *Ent_old* (the second and following occurrences of each *Ent*), using the *B* tag of the *Ent*. Then, we check whether the distributions of filler positions (by its token position in the transcript) are significantly different compared to the distributions of 1. *Ent_new* and 2. *Ent_old* positions (by its first token's position), by utilising a Kruskal-Wallis H test[4] and use the Benjamini-Hochberg procedure for multiple testing correction. We then estimate the effect size by computing Cliff's Delta δ[5]. Lastly, we compare the δ distributions of the two experiments, i.e. fillers with *Ent_new* versus fillers with *Ent_old* using a Wilcoxon signed-rank test, to see if they significantly differ.

RQ2: (Global effect of fillers): How does the speaker's production of fillers build to the listener's perception of the speaker? To investigate H2, we take the mean of the three confidence labels provided by the three annotators as the final rating of the speaker giving the review. We then consider reviews that are categorised as high-confidence (HC) and low-confidence (LC). Since confidence ratings are positively skewed[6] we take ratings of 3 (a little confident) and below to denote LC speakers, and 6 and above to denote HC speakers. The resulting size of the categories are 130 HC and 116 LC speakers. To calculate the percentage of fillers preceding new information (denoted by a new entity), we first consider the *Ent_new* labels that were automatically annotated in H1. We then count the number of fillers in the review that occur before (but not after) an *Ent_new*, constrained to a maximum distance of 1 token in between the filler and *Ent_new*. We

[4]We utilise this method according to the guidelines given in the scipy software (`https://scipy.org/`) where the test is only run if the samples for each category ≥ 5. We calculate Cliff's delta regardless of this criteria.

[5]Utilising effect size tools from `https://github.com/ACCLAB/DABEST-python/blob/master/dabest/_stats_tools/effsize.py`

[6]This is shown both in the annotation guidelines, and the ratings itself, as annotator's may have hesitated to rate the speaker 1 (not confident). and preferred instead to use the label 3 (a little confident).

Table 6.1: *OR* contingency table, where NE stands for the cumulative percentage of fillers that occur preceding an *Ent_new* for all HC (a) / LC (b) reviews, and OC the remaining cumulative percentage of fillers used in other contexts ((c) and (d) respectively).

		Outcome	
		HC	LC
Exposure	NE	a	b
	OC	c	d

normalise by dividing this count by the total number of fillers used in the review. From this, we obtain the percentage of fillers that occur before an *Ent_new* versus the percentage of fillers used in any other context that is not *Ent_new*. We then sum these two values for all HC and LC reviews, to get a cumulative percentage (please see Table 6.1).

We compute Odds Ratios (*OR*s) in order to investigate whether the use of fillers around new entities is associated with confidence. Odds ratios are an association measure that represents the odds that an outcome will occur given a particular exposure, compared to the odds that the outcome will occur in the absence of that exposure. Here, the odds denote the outcome of HC or LC, given the occurrence of fillers before new entities, compared to the occurrence of fillers that do not occur before new entities. We expect that the more fillers are used in the context of preceding new entities, the greater the odds of HC.

$$OR = \frac{odds_{HC}}{odds_{LC}}$$

where $odds_{HC} = a/c$ and similarly $odds_{LC} = b/d$ using Table 6.1 for reference. We then individually compute the *OR*s for each filler "uh" and "um", to see if there is a distinction in the way they are used in discourse.

6.3 Results and Discussion

6.3.1 RQ1: (Local effect of fillers) How does a speaker's use of fillers relate to the incoming information the speaker is sharing

H1 Fillers are more likely to occur before the introduction of new information in the review.

Table 6.2: Results of the Kruskal-Wallis H test, to compare the distributions of filler positions (by its token position in the transcript) compared to *Ent_new*/*Ent_old* positions, where "corrected" indicates the p-value after the Benjamini-Hochberg procedure. Note: Each cell indicates the number of reviews

	$p > .05$	$p \leq .05$
Ent_new	322	59
Ent_new corrected	381	0
Ent_old	477	70
Ent_old corrected	547	0

6 Results for H1 are given in Table 6.2 for the Kruskal-Wallis H test, to compare the distributions of filler positions compared to 1. *Ent_new* and 2. *Ent_old* positions. By Kruskal-Wallis H test the distributions are significantly different for $\approx 15 - 20\%$ of the reviews (where $p \leq .05$). However, after utilising the Benjamini-Hochberg procedure for multiple testing correction, the distributions using this method do not significantly differ. This test is calculated using the sum of the ranks of each distribution. Given that the average review length is short (≈ 256 tokens), and considering the close average median of fillers, *Ent_new* and *Ent_old* as given in Table 6.3, on reflection, this test may not capture nuances of the positional effects of fillers.

While significance testing focuses on a dichotomous result (i.e. significant versus not), we utilise Cliff's Delta δ to gain further insight into the magnitude of the effect. To interpret the results, Cliff's Delta δ ranges from -1 to 1, where 0 would indicate that the group distributions overlap completely; whereas values of -1 and 1 indicate a complete absence of overlap with the groups. For e.g. in H1 *Ent_new*, -1 indicates that all fillers in the review occur before new entities, and 1 indicates that all fillers in the review occur after new entities. This means that the smaller the effect size (close to zero) the larger the overlap, and the larger the effect size, the smaller the overlap. The magnitude of Cliff's Delta δ can be interpreted by using the thresholds from Romano et al. (2006), i.e. $|\delta| < 0.147$ "negligible", $|\delta| < 0.33$ "small", $|\delta| < 0.474$ "medium", and otherwise "large".

By computing δ to estimate effect sizes as given in Figure 6.2, we see that for most reviews, fillers do occur visibly before *Ent_new* (median $= -0.30$, $SD = 0.41$), but not before *Ent_old* (median $= 0.20$, $SD = 0.37$, given in Table 6.3), where the distributions of the δ values significantly differ ($Z = 27578.0$, $p < .05$ using Wilcoxon signed rank test). We see further evidence for this in Table 6.4, where majority of the reviews (451) have fillers

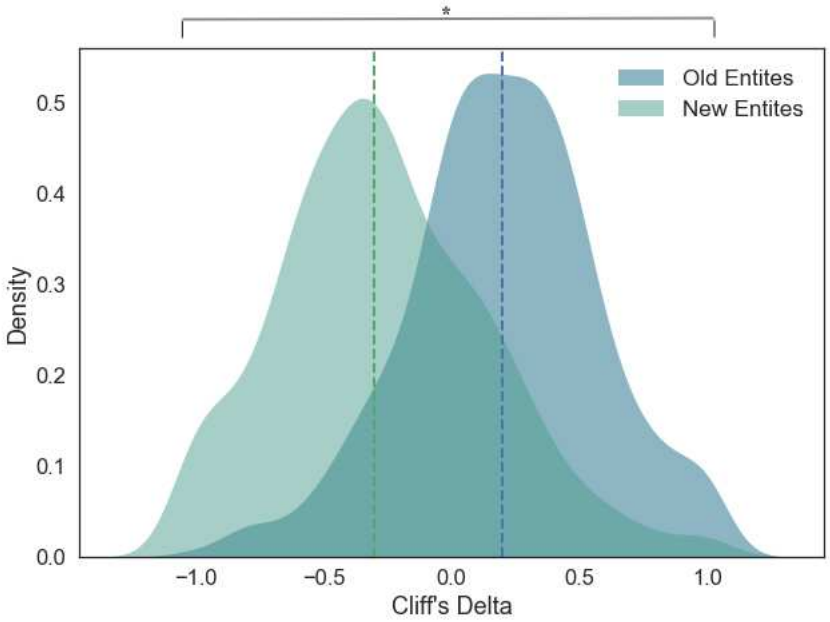


Figure 6.2: Distribution of Cliff's delta δ for fillers with *Ent_new* (New Entities) and fillers with *Ent_old* (Old Entites). Wilcoxon signed rank test has been performed to test whether the distributions significantly differ, with $p < .05$ given by $*$. The dotted line denotes the median (given in Table 6.3).

Table 6.3: Average median and *SD* for *Ent_new*, *Ent_old* (by first token position) and Fillers, and median and *SD* for effect size of the two δ distributions respectively.

	Avg. Median	Avg. SD
Ent_new	66.32	88.21
Ent_old	67.84	156.91
Fillers	66.05	125.95
δ*Ent_new*	-0.30	0.41
δ*Ent_old*	0.20	0.37

occurring before *Ent_new* (with larger overlap from Negligible to Medium effect size, i.e. of "nMedium" to "Negligible" δ sizes), compared to 277 reviews that had "nLarge" effect size (smaller overlap given the larger effect size, however still negative to indicate that fillers occur predominantly before *Ent_new*), and 139 reviews that had positive effect size (reviews that had fillers predominantly occurring after the introduction of new entities).

Table 6.4: Counts of Cliff's delta δ for fillers with *Ent_new* and fillers with *Ent_old* for all reviews, where the "n" or "p" before each row value indicates negative or positive values respectively.

	Ent_new	*Ent_old*
nLarge	277	36
nMedium	142	36
nSmall	146	66
Negligible	163	247
pSmall	62	156
pMedium	35	138
pLarge	42	189

We see the opposite δ effect sizes for *Ent_old*, where most of the reviews have fillers occurring after entities already introduced in the discourse (with predominately positive δ values as shown in Table 6.4), but not before. Fillers occurring after *Ent_old* is entirely plausible given that new entities can occur throughout the review, and not just at the start of one (as shown in Table 6.3, where the average median of *Ent_new* is roughly the same as *Ent_old*). Given the larger group with negligible effect size (247) for *Ent_old*, this does show that speakers may sometimes use fillers when repeating entities already introduced into the discourse.

In Chapter 5 we used a language model (LM) trained on our spontaneous speech dataset of monologues to observe the probability of a filler appearing at a certain position; and found that the learnt word distribution shows that the LM places fillers predominantly at the start of sentences, and then struggles with sentence-medial fillers. However, sentence boundary annotation is dependent on the perspective of the transcriber, which in turn is certainly based on the presence of prosodic cues and fillers itself. Our findings suggest that in this sentence boundary agnostic methodology (as this calculation is done word by word over the entire discourse), while the introduction of new entities can occur commonly at sentence boundaries transcribed in the dataset, there may be more subtle nuances to the way speakers produce fillers in terms of incoming information. **Therefore, regarding H1, stylistically speakers do tend to use fillers in the incoming message when introducing a new entity rather than one already introduced** (whether intentionally or not remains an open question), and the positions of fillers with respect to *Ent_new* significantly differ from positions of fillers with respect to *Ent_old*.

6.3.2 RQ2: (Global effect of fillers): How does the speaker's production of fillers build to the listener's perception of the speaker?

H2 From H1, the speaker's use of fillers preceding new information is associated with the listener's perception of the speaker's confidence.

To investigate the presence of fillers occurring before new information among confidence ratings, we computed *OR*s. To interpret the results, when $OR = 1$, the presence of the percentage of fillers that occur before new entities (exposure) does not affect the odds of neither HC nor LC (i.e. no association of the exposure with outcome). When $OR > 1$, the presence of the exposure is associated with higher odds of HC (positive association). When $OR < 1$, the presence of the exposure is associated with higher odds of LC (positive association with decrease of HC).

The results of the test show $OR = 0.72$ ($p < .001$, 95% CI : 0.6-0.8)[7]. While $OR < 1$ in this case, indicating that the presence of fillers occurring before new entities gives a higher odds of LC, it is closer to 1, showing that the presence of the stimulus on the outcome is small. Interestingly, these findings are the opposite of what was hypothesised, which was that the speaker's use of fillers preceding new information is associated with the listener's perception of confidence; i.e. the more fillers are used in this way, the greater the odds of HC. According to the results of the *OR*s test, fillers occurring before new entities do not have a great effect on the odds of HC (only 28% lower given the presence of new entities) of the rating that the listener gives the speaker. This is the consistent with the existing psycholinguistic literature on fillers as discussed in Sec. 6.1, that fillers can be used to inform about upcoming *new* information. Specifically, Arnold et al. (2004) for e.g. showed that fillers bias listeners towards new referents rather than ones already introduced into the discourse. Barr and Seyfeddinipur (2010) found that listener's expect the speaker to refer to something new following a filler (although they also found this to be specific to what was new for the speaker, and not only the listener), showing that listeners interpret fillers as delay signals, and infer plausible reasons for the delay by taking the speaker's perspective. While we cannot account for whether the annotator had rated the same speaker in multiple reviews, the annotator thus may expect the speaker to use fillers before new entities, or generally, before new expressions. In a study of the two fillers "um" and "uh" in American English, Tottie (2014) found that in natural conversation, listener's are not aware of the use of fillers, unless overused or used in the wrong context. Fillers used before new entites may

[7]Risk Ratio $RR = 0.826$ with $p = .001$, 95% CI : 0.7-0.9

not be considered usage in the "wrong context", and indeed, could simply indicate an increase in the number of entities in the review.

Looking at Figure 6.3 (taken from Chapter 4), to show the average rate of fillers in the review (given by the percentage of fillers used compared to tokens in the review), it is clear that the use of fillers differs between HC and LC rated speakers (The median filler rate is 4.82 and 3.049 for LC and HC respectively, with $U = 7329.5$ by two-sided Mann-Whitney U test with Bonferroni correction.). These results do not contradict Brennan and Williams (1995), i.e. there could be impressions formed by the listener about the speaker's expressed confidence based on fillers in spontaneous speech (as found in Part II). However, these results would suggest that the association may not be from fillers used in the context of introducing new information. This is an interesting finding; as fillers in these contexts may still have a metacognitive function as discussed above, but not necessarily related to the listener inferring that the speaker is facing *cognitive load in a way that negatively impacts perception*. Note, that with this strict distance criteria (maximally, 1 token of distance), a small percentage of fillers ($\approx 10\%$) overall occur *just before* new entities, while *generally* over the entire dataset, we see the tendency for fillers to occur before new entities. Thus when looking at Figure 6.3, there are still several other ways that fillers are utilised, as discussed in Part II. **While we cannot reject the null hypothesis, there is significant evidence using our methodology to suggest that the occurrence of fillers before new entities has a negligible association on confidence (both HC nor LC) (with $p < .001$ and $95\% CI$: 0.6-0.8).** Thus, these results suggest that fillers used in the context of introducing new entities in the discourse has negligible association on the listener's rating of confidence that they attribute to the speaker.

On the difference between "uh" and "um" The results of the test specific to the filler "um" show $OR = 0.70$ ($p < .001$, $95\% CI$: 0.57-0.85), and for the filler "uh" $OR = 1.045$ ($p = 0.58$, 95% CI : 0.89-1.22). We see that both fillers have an OR closer to 1, showing that the association of the stimulus on the outcome is small. However, we see that there is not sufficient evidence to conclude that the presence of the percentage of the filler "uh" that occurs before new entities (exposure) is not associated with the odds of confidence scores (i.e. no association of the exposure with outcome, given that the 95% CI crosses 1, indicating low level of precision). **This suggests there may not be enough of a distinction in the way the filler "uh" is used to introduce new and old information in HC/LC rated speakers.** However, when considering the filler "um", we see that the results are similar

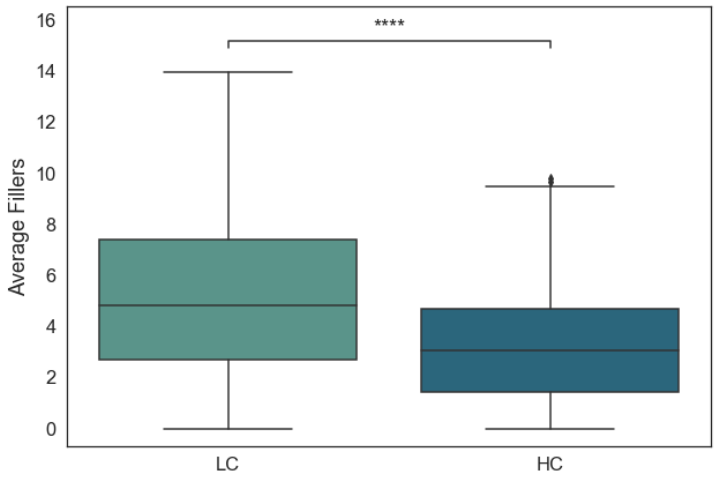


Figure 6.3: Box plots showing the speaker's average use of fillers *num* for HC/LC reviews, where $****$ denotes $p < .0001$. The median *num* is 4.82 and 3.049 for LC and HC respectively, with $U = 7329.5$ by two-sided Mann-Whitney U test with Bonferroni correction. (Taken from Chapter 4).

to when the experiment was conducted with both fillers together, $p < .001$ and 95% CI : 0.57-0.85, giving **strong evidence to support that the filler "um" occurring before new entities is not largely associated with confidence (only 30% lower chance given the presence of new entities)** (with $p < .001$ and 95% CI : 0.57-0.85). This is also interesting, considering that we previously saw that overall (considering all contexts) the filler "um" is *more important* in terms of the final rating the listener gives the speaker compared to the filler "uh". These findings support Tottie (2014)'s claims about the listener only noticing the speaker's use of fillers if overused or used in the wrong context; perhaps the listener *expects* (knowingly or unknowingly) the speaker to produce "um" in a more informative way than the filler "uh". Thus, the listener may not consider it when used correctly, but it may be perceptible if used incorrectly. This suggests that the filler "um" may be perceived more informative by the listener, and also, perceived differently than "uh", in terms of signals of information. Furthermore, this shows that the filler "um" may also be *produced* differently in discourse than the filler "uh".

6.4 Conclusion

The aim of this chapter was to investigate how fillers interact with the primary signal at an utterance level, and whether this can help in understanding/ interpreting the perception of the speaker at a discourse level. We treated the message from the speaker as *an incoming source of information that builds up in the discourse*, and then we *observe the function of fillers in this process*. Thus, we computationally studied at a micro-level, which fillers are informative in terms of the new information (measured by new entites) specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception. We based this work on the numerous psycholinguistic perspectives that state that the listener can treat fillers as informative signals with regard to the incoming message (e.g. Corley, MacGregor, and Donaldson (2007), Barr and Seyfeddinipur (2010), and Arnold et al. (2004)).

Our findings suggest that speakers generally do tend to use fillers in the incoming message, when introducing a new entity (to indicate new information), rather than an entity already introduced into the discourse. Our results also suggest that the occurrence of fillers specifically before new entities may not have an effect on the listener's perception of the speaker's expressed confidence, despite previous works that suggest the link between fillers and expressed confidence. These results were on the extremes of the dataset, i.e. speakers that were rated with low confidence versus speakers that were rated highly confident. **Thus, local hesitations need not always lead to global impressions of uncertainty.** This does not discount other possible metacognitive aspects, such as the listener may expect a speaker to use fillers typically when the speaker is introducing new information in the incoming message, and indeed, think that this "cognitive load" is expected (therefore it may not contribute to the perception of confidence). In the *perspective taking account* of language comprehension as discussed in Barr and Seyfeddinipur (2010); the listener might be drawn to the mind of the speaker and infer possible reasons for delays in speech. Our analysis shows the possibility of different metacognitive functions in this perspective taking account that are brought about by the use of fillers on the listener.

We also find that listener's may perceive the filler "um" as a more informative signal than the filler "uh"; and have expectations as such. These findings are based on our previous results, that show that *all* contexts of the filler "um" have more of an effect on the listener's rating of confidence than the filler "uh", but specific to the filler "um" introducing new information, there is very small association with the confidence rating. This also suggests

that they have different roles in the discourse (which we subsequently find in dialogues as well).

A limitation of this work, is when considering the use of fillers, an important aspect is the acoustic information – as fillers are ubiquitous to spontaneous speech. While our measures focus on the transcripts and use ranking, it loses this temporal information, for e.g. distances in time, durations of fillers etc. However, it is difficult to calculate H1 in terms of time (rather than position), due to the poor results of the forced alignment algorithms on this dataset. Since speaker's recorded themselves voluntarily and naturally using their own equipment, it is hardly surprising that the audio data is noisy. However, considering that SLU is often done on the output transcripts of ASR without considering acoustic information (except for the purposes of speech recognition), we consider these results as first steps towards the holistic study of how the collateral signal can contribute to the primary signal.

IV | Disfluencies in Communication:

Verbal and Behavioural Alignment in Situated Dialogues

7 | What Was Said? Fillers in Verbal Alignment

7.1 Introduction

In the previous part (Part III), our main goal was to develop methodologies to explore how fillers (a *collateral signal*), interact with the *primary/lexical signal*, or what was said in essence. We thus treated fillers as a source of incoming information in the speech stream. We investigated the local use of fillers at an utterance level, and observed whether this lead to an association with discourse level phenomena. We based the work on psycholinguistic perspectives that state that the listener can treat fillers as informative signals with regard to the incoming message (e.g. Corley, MacGregor, and Donaldson (2007), Barr and Seyfeddinipur (2010), and Arnold et al. (2004)).

However, so far, our work has only been based on monologues; where the speaker is conscious of an *unseen* listener. In the context of a dialogue, there is a more dynamic and iterative communication processes between the interlocutors. Particularly for our dataset of situated, task oriented dialogues (see Chapter 3, where the roles of IF and IG are constantly swapped), the interlocutors need to communicate with each other in order to make progress in the task. In this context, there are more *immediate, information sharing goals* between the dialogue participants, and broader goals such as *success in the task*. Thus, we need to account for this context of listener-speaker dynamic when considering the primary signal, compared to our previous approach.

With this in mind, **the aim of this chapter** is to build on our previous findings of fillers as informative cues for the primary signal, but in this context, focusing on dialogues. For this purpose, we automatically and empirically study the uses of fillers in a *situated task-oriented dialogue*, i.e. where the dialogue has an interdependence on the immediate environment, and thus, the formation of *expressions related to the task*. Focusing on expres-

sions related to the situated environment allows us to focus on information very specific to furthering the task. Furthermore, we account for the listener-speaker dynamic by developing algorithms to automatically annotate the first time this expression related to the task was uttered, to the time it was repeated and became part of a shared vocabulary between the interlocutors. This entire process is called *verbal/lexical alignment*. Concretely, we study:

verbal alignment

1. **At the utterance/local level:** The effect of fillers in relation to the primary signal, specifically verbal alignment. In this dialogue context, we only consider expressions that become part of a *shared vocabulary* amongst the interlocutors – i.e. only if the expression the speaker introduces is repeated by the listener, it is considered part of the shared vocabulary of the interlocutors.

2. **At the dialogue/global level:** Furthermore, we study whether this process can build to an association with the interlocutors' task success (measured by the team's *performance* in the task and their *learning* outcomes).

Specifically, we study a dataset of children's dialogues, where they are engaged in a *collaborative learning activity* (see Sec. 7.2), in which what they *say* is strongly tied to how they perform, and subsequently what they will ultimately learn from the activity.

From a methodological perspective, the automatic analysis of spontaneous speech in a collaborative learning activity and its connection to task success is extremely challenging for three reasons. Firstly, existing methods for studying collaboration rely on time-consuming manual annotations of conversations. Here, we want to provide automatic tools for studying spontaneous speech phenomena in the collaborative processes. Second, we tackle the problem of the analysis of spontaneous conversational speech of children, which has its own features, which could be more difficult to handle and less-structured than adults' speech. We propose automatic methods to investigate how children refer to their environment. Thirdly, alignment algorithms[1] are not specifically suited to study spontaneous speech phenomena like fillers. Thus we consider the children's *production* of fillers in the context of verbal alignment, and then ask whether this impacts their task success, i.e. task performance and learning outcomes. This part (Part IV) highlights the need to bridge the gap between the fields of collaborative learning and dialogue, and expand existing automatic measures for verbal behaviours in spontaneous speech, as they can have an impact on our understanding about the way children learn.

[1] or the "development of similar representations" (Pickering and Garrod, 2006)

The rest of the chapter is organised as follows: We describe the related work and terminology used in Sec. 7.2. In Sec. 7.3, we refer to the relevant sections in the thesis for a description of the dataset used, and outline our research questions (RQs) and hypotheses. In Sec. 7.4, we describe our methodology to adapt the alignment algorithm for situated dialogue, and then the methodology for the RQs and hypotheses. Sec. 7.5 gives the results and discussion, while Sec. 7.6 discusses the conclusion of the experiments.

7.2 Related Work

In this section, we discuss the research so far on verbal analysis in collaborative learning activities, and introduce terminology used in this chapter related to dialogue literature[2]. Then, we detail the kind of spontaneous speech phenomena behaviour we expect to see, based on previous literature and our findings. Lastly, we detail the rule-based alignment model that we adapt to study verbal alignment in situated dialogues. For a more detailed background specific to disfluencies, please refer to Part I.

Collaborative learning activities

Collaboration occurs in a situation in which individuals work together as members of a group in order to solve a problem (Roschelle and Teasley, 1995). In a collaborative activity, members build shared, abstract representations of the problem at hand (Schwartz, 1995). Collaborative activities have been analysed by focusing on the *verbal* (spoken communication) or the *non-verbal* (e.g. gaze, facial expressions, gestures) behaviours of the group members. Research has found that non-verbal cues could indicate the quality of collaboration, for example the link between high coupling of a pairs' focus of attention and high quality of collaboration (Jermann et al., 2011), the link between automatically annotated eye gaze coupling and interaction quality (Jermann and Nüssli, 2012), and the link between non-verbal cues like mutual gaze and laughter, and group cohesion (Bangalore Kantharaju et al., 2020). In terms of verbal behaviour, it was found that by automatically labelling silent and spoken episodes, the amount of speech is higher for pairs who had a higher quality of interaction (Jermann and Nüssli, 2012). Since a spoken episode could contain any kind of speech, this study was limited, and the authors conclude the need to analyse verbal behaviours in greater detail in the study of collaborative activities.

[2]We use the term dialogue literature to refer to relevant linguistic (e.g. psycholinguistics, discourse, conversation analysis) literature.

collaborative learning activities

When a collaborative activity is in an educational setting, it is termed as *collaborative learning*: where individuals 'learn' together. The group interactions in collaborative learning are expected to activate mechanisms that would bring about learning, although there is no guarantee that these beneficial interactions will happen (Dillenbourg, 1999). Collaborative learning involves group processes that are carried out interactively and hence it contains, but it is irreducible to, individual learning (Stahl, Koschmann, and Suthers, 2006). According to Dillenbourg et al. (1996), there is a need for new tools to analyse these interactions by treating them as group processes, to better understand different learning mechanisms. Since then, tools for collaborative learning have been developed and studied in different contexts, for e.g. to study the impact of interpersonal closeness between collaborators on learning outcomes (Madaio, Cassell, and Ogan, 2017b), how interpersonal closeness affects indirectness of feedback and instructions (Madaio, Cassell, and Ogan, 2017a) and even work on automatic interpersonal rapport prediction for peer tutoring (Madaio et al., 2017).

Drawbacks The verbal analysis of collaborative learning tends to be based on manual coding schemes that distinguish between various types of verbal interactions. For example, Visschers-Pleijers et al. (2006) labelled several types of verbal interactions, such as asking exploratory questions and handling conflicts about knowledge, and inquired into how they are distributed over time in a collaborative learning activity. Similarly, Yew and Schmidt (2009) used a more extensive coding scheme to qualitatively study the nature of verbal interactions. Yew and Schmidt (2012) also manually labelled and counted the number of relevant concepts that were verbalised, and related this to learning outcomes. However, we need automatic tools that analyse verbal interactions in a collaborative learning activity as, so far, qualitative analyses require extensive manual annotation.

In parallel, there have been studies that do not focus on educational goals and collaborative learning, but on how humans understand each other by analysing human-human dialogues. Similar to understanding group processes in collaborative learning, studying a dialogue is challenging; as rather than an individual effort, a dialogue is an interactive activity performed by two or more people, i.e. interlocutors (Clark and Wilkes-Gibbs, 1986). Interlocutors ideally take turns to try to reach a common or *mutual understanding* (Clark and Schaefer, 1989). Dissimilar to collaborative learning, works that study mutual understanding largely consider dialogues amongst adult interlocutors solving a problem together e.g. in (Garrod and Anderson, 1987; Fusaroli and Tylén, 2016), without the added dimension of a learning goal. However, task

performance is not always positively correlated with learning outcomes, as a student can fail in the task but learn from it (even learn from failure, as in "productive failure" (Kapur, 2008; Kapur and Bielaczyc, 2012)) and perform well in the task but not learn from it (i.e. unproductive performers (Kuhn, 2015)).

Thus, works on collaborative learning extensively study behaviours related to learning outcomes, but can lack in-depth dialogue analysis, and automatic tools are missing. On the other hand, several works on dialogue, particularly on mutual understanding, have extensively studied the dialogue, but without the added depth of learning outcomes. Thus while our preliminary focus is on disfluencies in the context of verbal alignment, a consequence of this is the development of methodologies that will *allow for* the automatic analysis of situated dialogues. This approach brings together works on collaborative learning to consider *completely spontaneous dialogues*. What we mean by this, is developing *holistic* methodologies to study the primary and collateral signal together.

Referring in a situated environment

In a collaborative activity, interlocutors often work together to achieve success in a task. When these activities are situated, the verbal communication has an additional dependence on the immediate environment in which the dialogue takes place (Kruijff et al., 2010). This entails interlocutors *refer*ring to the physical environment. Often, referring in such dialogues manifests in action, where in order to take an appropriate action, interlocutors must understand (to some degree) the other's way of referring to their environment (Clark and Brennan, 1991). Consider the following example of a situated dialogue, taken from the JUSThink Dialogue and Actions Corpus. Two interlocutors had to work together to construct a solution in a situated activity, by deciding to connect (or disconnect) mountains on a given map together.

situated dialogue

A: What about Mount Davos to Mount, Saint Gallen?
B: Because what if, you did if we could do it?
A: What about Mount um Davos to Mount Gallen?
B: Mount . . .
B: oh Mount Davos (7.1)
A: yeah to Mount Gallen.
B: to Mount Gallen yeah do that.
⟨*A* connects Mount Gallen to Mount Davos⟩
B: Okay my turn.

A's way of referring to a landmark in the environment (e.g. Davos – "Mount Davos") was understood by *B* as the same landmark (Davos – "oh Mount Davos"), as evidenced by repeating it. Thus, interlocutors form *referring expressions*, like "Mount Davos", which are pointers to objects, which we call *task-specific referent*s (here, the landmark Davos), that have been referred to within the situated environment. In the above Eg. 7.1, because *A* and *B* mutually understood the other's use of referring expressions; they make progress in drawing the route correctly. In this chapter, we focus on task-specific referents as our primary signal; i.e. what the interlocutors need at minimum to communicate with the other how to proceed in the task.

referring expressions

Grounding and alignment

Mutual understanding, at least to some degree can be attributed to a process called grounding (Clark and Schaefer, 1989; Clark and Brennan, 1991), which can be thought of as adding to what is already mutually understood. We have currently defined this notion imagining that interlocutors contribute to their common ground, i.e. build up a shared mutual understanding. However, Pickering and Garrod (2004) stress on the importance of an automatic *priming* mechanism, i.e. when a listener encounters an utterance from the speaker, it is more likely that subsequently the listener will produce an utterance by using this representation. Here, the listener need not explicitly assent-to/accept the speaker's contribution; instead, there can simply be an implicit common ground unless a misunderstanding requires changing the representation. Like this, *routines*, or referring expressions that are "fixed", become shared or *established* amongst interlocutors. Thus rather than grounding, they define *alignment* of representations, to mean the "development of similar representations in the interlocutors" (Pickering and Garrod, 2006). Therefore the interlocutors succeed in understanding each other when there is *alignment* between them, or a shared representation, at different linguistic levels. Pickering and Garrod (2004) and Pickering and Garrod (2006) in a maze task found that at first glance, the dialogue seemed unstructured, but then found underlying alignment structure that supported collaboration amongst the interlocutors.

routines

alignment

Medium of communication and spontaneous speech

Situated dialogues can happen through constrained mediums of communication, for example chatting via a text environment e.g. in (Stahl, 2007; Dillenbourg and Traum, 2006). We focus on face-to-face communication, due to the co-presence of the interlocutors, visibility and audibility in the

medium, which can dramatically affect the process of mutual understanding (Dillenbourg and Traum, 2006; Clark and Brennan, 1991). Since we focus on face-to-face communication, spontaneous speech phenomena becomes a relevant dimension of analysis in verbal behaviour.

In the study of verbal alignment in a situated face-to-face environment, we aim to study shared lexical representations developing between interlocutors, and also the interaction with surrounding spontaneous speech phenomena. Clark (1996) and Clark and Fox Tree (2002) proposed that speakers are able to utilise fillers as *collateral signals* in communication, in addition to the *primary signal* of the message. We colloquially refer to the primary signal of the message as *what* was said (in essence) and the collateral signal as *how* it was said. In traditional SLU, a similar phenomenon occurs of separating these two signals. However, in this context, reducing an input utterance into its primary signal (or *what* was said in essence) is standard practice (e.g. as seen in (Tur and De Mori, 2011, Ch. 13: Speech Summarisation)). However, these expressions do not occur in a vacuum, they are part of a spoken dialogue amongst interlocutors, which is 'messy' (disfluent).

The relevance of a pedagogical goal in the study of fillers The analysis of spontaneous speech studying both primary and collateral signal together is often neglected, but could be illuminating on a pedagogical level. There are many works that discuss the information sharing properties of fillers. Barr and Seyfeddinipur (2010) found that listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle). Vasilescu, Rosset, and Adda-Decker (2010a) investigate the potential of the French "euh" (which could be considered a filler, which sometimes can be used in a discourse marker role) to bracket lexical information with the aim of improving QA systems. Works such as Barr and Seyfeddinipur (2010), (and to some extent, Vasilescu, Rosset, and Adda-Decker (2010a), whonobserve that certain discursive properties of the French "euh" may benefit listener comprehension) support the collateral signal account from Clark and Fox Tree (2002), specifically that the listener is able to process fillers as a collateral signal (even if unclear whether the speaker (un)intentionally used them as such). This is an important finding, as it shows that perhaps the listener's attention is drawn to the cognitive state of the speaker, and infers plausible reasons for the speaker to use fillers. Fillers have also been linked to signs of hesitation and uncertainty (Pickett, 2018; Smith and Clark, 1993; Brennan and Williams, 1995), and our previous analysis shows that listeners are perceptive to this. We believe that by focusing both on the primary signal (verbal alignment; i.e.

the priming and establishment of shared expressions related to the task), and the collateral signal (the production of fillers surrounding these expressions) we will get further insight into the way fillers are used to share information in a dialogue, and also, the way children refer to their situated environment and thus learn.

Rule-based methods for extracting structures of alignment

Dubuisson Duplessis, Clavel, and Landragin (2017) and Dubuisson Duplessis et al. (2021) proposed automatic and generic measures to extract *lexical structures of alignment* (which they also refer to as *verbal alignment*) in a task-oriented dialogue. The proposed method works on alignment based on surface matching of text and does not focus on other levels of linguistic alignment as envisioned in Pickering and Garrod (2004). However, it is done with the aim of automatically finding these text patterns in the dialogue, by sequentially processing a transcript in a way that does not require manual annotation. In this chapter, we provide a new tool/framework for studying situated dialogue building on the automatic and generic methodology by Dubuisson Duplessis, Clavel, and Landragin (2017) and Dubuisson Duplessis et al. (2021): this implies to model how the interlocutors refer to their environment and to extend the tool based on this model. We expand on this in Sec. 7.4.

7.3 Research Questions

Dataset Please see Chapter 3 for a detailed description of the JUSThink Dialogue and Actons Corpus used in this chapter, particularly for i) **At an utterance level:** why this dataset was suited to study alignment, ii) **At dialogue level:** a background on how the task success variables of performance (*error*) and learning outcomes (*learn*) are calculated, and iii) A detailed description of the transcription campaign of this dataset.

task-specific referents

This chapter focuses on fillers in relation to the *task-specific referents* that interlocutors minimally require to succeed in the task. Therefore, we restrict the possible referring expressions to ones that contain task-specific referents, in particular, to the objects that the interlocutors are explicitly given on the map (see Figure 3.3). Interlocutors only need this terminology with certain function words to progress in the task (e.g. "Montreux to Basel"). We believe that this design choice is particularly suited to study alignment in this type of activity, as the frequent swapping of views encourages the interlocutors to

communicate with the other their intents using these referents. This allows us to focus on spontaneous speech phenomena that are explicitly linked to the 'situatedness' of the task and it's relation to a dialogue measure of task success. While there are certainly other referring expressions to consider (e.g. "That mountain there") not containing task-specific referents, it would require some degree of manual annotation. It is similar for other communicative expressions related to the task (e.g. "I didn't understand"). We thus focus on task/domain specific referents that can be automatically extracted.

With this in mind, RQ1 considers the *production* of fillers in relation to task-specific referents, while RQ2 considers the relation between fillers and task success. Thus RQ1 and RQ2 focus on the surface-level utterances, i.e. "What/How did the interlocutors say", and "how did this impact their task success".

RQ1: (Local effect of fillers) How do the interlocutor's use of fillers relate to their verbal alignment? In RQ1, we specifically consider the link between the use of task-specific referents and surrounding fillers. We choose fillers specifically as they can be used by the interlocutor to inform the listener about upcoming new information, or even production difficulties that they are facing. For e.g. particular to the establishment of referring expressions, research has shown that disfluency (studied with the filler "uh") biases listeners towards new referents (Arnold et al., 2004) rather than ones already introduced into the discourse, helps listeners resolve reference ambiguities (Arnold, Kam, and Tanenhaus, 2007), and that listeners expect the speaker to refer to something new following the filler "um", compared to noise of the same duration (such as a cough or sniffle) (Barr and Seyfeddinipur, 2010). Specifically, we hypothesise:

- **H1: Fillers tend to occur in the vicinity[3] of priming and establishment of task-specific referents.** We expect that verbal alignment in the dialogue, via priming (the speaker first introduces the referent) and establishment (the listener utilises this referent for the first time) of routine expressions are associated with fillers. Thus the interlocutor can use a filler to inform the other about the introduction of a new contribution into the dialogue, or inform the other about the acceptance of this contribution (by repetition of the expression for the first time).

priming

establishment

[3]We use the term vicinity here; i.e either before or after, because the process of routinisation is for the entire dialogue, i.e. an expression may be primed at some point and established at some other point in the dialogue – even the end.

RQ2: (Global effect of fillers) How do the interlocutor's use of fillers relate to their task success? Specifically, we hypothesise:

- **H2: Fillers that precede the priming and establishment of task-specific referents are associated with task success.** In the *perspective taking account* of language comprehension as discussed in Barr and Seyfeddinipur (2010); the listener might be drawn to the mind of the speaker and infer plausible reasons for delays in speech – i.e. use of fillers. Our analysis on monologues shows the possibility of different metacognitive functions in this perspective taking account that are brought about by the use of fillers on the listener, i.e., that fillers used to introduce new information may not have an effect on the listener's perception of the speaker's confidence. Following this, we expect that each interlocutor is able to distinguish between fillers used in the formation of a routine (both priming and establishment), and this might be associated with a greater chance of task success, particularly the learning process.

7.4 Methodology

Adapting rule based alignment models to study verbal alignment We formally define a *routine* expression (adapted from Dubuisson Duplessis, Clavel, and Landragin (2017), Dubuisson Duplessis et al. (2021), and Pickering and Garrod (2004)) as a referring expression shared by two interlocutors if i) the referring expression is produced by both interlocutors, and ii) it is produced at least once without being part of a larger routine. A routine is based on the reuse of a referring expression, but is specific to the exact matching of token sequences in two utterance strings. Using the above terminology, 'Montreux' is a task-specific referent, and an interlocutor might *prime* the referring expression "Mount Montreux". If the other interlocutor *reuses* this referring expression, it thus becomes *routine*. In particular, we define the utterance at which a referring expression becomes routine as the *establishment* of that routine. We extract routines and the utterances at which the routines are primed and established from the transcripts as in (Dubuisson Duplessis, Clavel, and Landragin, 2017; Dubuisson Duplessis et al., 2021). Then, we filter for the routines that contain a task-specific referent[4].

[4]We process the transcripts (folder *transcripts/* in the dataset at DOI: 10.5281/zenodo.4627104) with a script (*tools/4_extract_routines_from_transcripts.ipynb* in the tools at DOI: 10.5281/zenodo.4675070) to extract routines (in tables in folder *processed_data/routines/*, available with the tools). We use the `dialign` package v1.0, avail-

Restraining the set of expressions allows us to focus on information specific to furthering the task. When looking at Eg. 7.1, we see in the first utterance that *A* says, "What about Mount Davos to Mount, Saint Gallen?". In the alignment model proposed by Dubuisson Duplessis, Clavel, and Landragin (2017) and Dubuisson Duplessis et al. (2021), *A*'s next utterance "What about Mount um Davos to Mount Gallen?" would contain a partial match of the surface pattern "Davos to Mount" (as an example of priming the expression once more). The goal of Dubuisson Duplessis, Clavel, and Landragin (2017) and Dubuisson Duplessis et al. (2021) is different, in that they want to automatically assess the percentage of shared vocabulary that develops amongst interlocutors in the solving of a task. However, in our method, filtering only for task-specific referents allows to specifically track when key expression required to further the task was introduced into the dialogue (primed) by the speaker, and when it was established by the listener as part of the shared vocabulary.

RQ1: (Local effect of fillers) How do the interlocutor's use of fillers relate to the priming and establishment of shared expressions? To investigate H1, we inspect the distribution of filler times, in relation to the i) priming times and ii) establishment time of routine expressions. This is done to account for when a speaker introduces a new task-specific referent into the dialogue, and when the listener makes this expression routine for the first time. In particular, we note the order of the tokens in the dialogue for the filler positions and the first token of the priming/establishment instance. Then, we check whether the distributions of filler times (by its token position) with priming/establishment times are significantly different (by its first token's position), by utilising a Mann-Whitney U Test, and estimate the effect size by computing Cliff's Delta.

RQ2: (Global effect of fillers) How do the interlocutor's use of fillers relate to their task success? To investigate H2, we divide the teams into high performing (HP) and low performing (LP) teams, and high learning (HL) and low (LL) teams respectively. We do so by taking the teams with the highest and lowest performance and learning scores respectively, and removing teams that ranked in the middle of task success. For performance (see Figure 3.4 for reference), this means HP are teams $17, 28, 20$ and LP are teams $8, 47, 18$[5]. Similarly, for HL teams, we consider $10, 11$, and teams

able at `https://github.com/GuillaumeDD/dialign`, by (Dubuisson Duplessis, Clavel, and Landragin, 2017; Dubuisson Duplessis et al., 2021).

[5]While teams $8, 47, 7, 10$ all have the same *error* score, in order to make the team numbers balanced for the ORs test, we rank these teams based on *learn*, and take team

Table 7.1: OR contingency table, where PE stands for the cumulative percentage of fillers that occur preceding a priming/established expression for all HP/HL (a) LP/LL (b) teams, and OC the remaining cumulative percentage of fillers used in other contexts ((c) and (d) respectively).

		Outcome	
		HP/HL	LP/LL
Exposure	PE	a	b
	OC	c	d

20, 18 as LL teams. The resulting size of the categories are 3 each of HP and LP, and 2 each of HL and LL teams respectively. We then consider the percentage of fillers preceding the routine formation of an expression (either priming or establishment) for each team by counting the number of fillers used by the team that occur before (but not after) a routine, constrained to a maximum distance of 1 token before the filler and the routine[6]. We normalise by dividing this count by the total number of fillers used by each team. We then sum these two values for all HP/HL teams and then for LP/LL teams, to get a cumulative percentage (please see Table 7.1).

We compute Odds Ratios (ORs) in order to investigate whether the use of fillers around routines is *associated* with performance or learning. Odds ratios are an association measure that represents the odds that an outcome will occur given a particular exposure, compared to the odds that the outcome will occur in the absence of that exposure. Here, the odds denote the outcome of HP/HL or LP/LL, given the occurrence of fillers before routine (either priming or establishment times), compared to the occurrence of fillers that occur elsewhere. We expect that the more fillers are used in the context of preceding the formation of routine expression, the greater the odds of task success (HP/HL).

7.5 Results and Discussion

7.5.1 RQ1: (Local effect of fillers) How do the interlocutor's use of fillers relate to their verbal alignment?

H1 Fillers tend to occur in the vicinity of priming and establishment of task-specific referents.

8, 47, with lower *learn* scores.

[6]we combine the number of fillers preceding the primed expression and established expression, due to otherwise sparsity of fillers in this category.

We investigate how are the fillers distributed as compared to the priming and establishment of the routines. The results for Mann-Whitney U test and effect size as estimated by Cliff's Delta (δ) for both fillers "uh" and "um" are given in Table 7.2. δ ranges from -1 to 1, where 0 would indicate that the group distributions overlap completely; whereas values of -1 and 1 indicate a complete absence of overlap with the groups. For example, -1 indicates that all fillers occur before priming times, and 1 indicates that all fillers occur after priming.

We observe that filler and **priming times** differ significantly for most of the teams (for eight out of ten teams by Mann-Whitney U test $p < .05$). We see positive δ values, except for one team. Most teams thus have fillers that occur after priming times. We see that the magnitude of δ for most teams is large except for Team 7 (excluding Teams 20 and 8 that are not statistically significant)[7]. We interpret that most fillers do occur visibly after priming times (with little overlap from the large effect sizes). We observe that filler and **establishment times** differ significantly for most of the teams (for seven out of ten teams by Mann-Whitney U test $p < .05$). We see δ ranging from -0.67 for Team 7, to 0.33 for Team 11 (though most are negative). However, we see that the magnitude of δ for most teams is small, with the exception of two teams. This means that the distributions of fillers differ with a small effect size compared to the distributions of establishment times, especially considering that the token number values do not overlap; i.e. we do not expect an effect size of 0. We thus interpret most fillers do occur around (from positive and negative δ values) establishment times (with larger overlap from small effect sizes).

These results indicate that in the process of routine formation, in between the priming and the establishment of the expression, there seems to be a period in which the interlocutors use fillers. This placement of fillers is of interest due to the potentially unfamiliar vocabulary of the task-specific referents that the interlocutors had to utilise in the situated environment. The results demonstrate a lack of fillers at the start of the formation of a routine; i.e. they occur visibly after the priming of expressions that contain task-specific referents. The large effect size shows that this is predominantly the case for most teams. In addition to this, fillers were found to occur around establishment times. We suggest that a part of establishment is often a clarification request, as shown by the following example. Teal indicates the routine expression, while blue indicates the filler.

[7]The magnitude of Cliff's Delta (δ) can be interpreted by using the thresholds from Romano et al. (2006), i.e. $|\delta| < 0.147$ "negligible", $|\delta| < 0.33$ "small", $|\delta| < 0.474$ "medium", and otherwise "large".

Table 7.2: The results for the Mann-Whitney U test that compares the distribution of filler times (both "uh" and "um") with establishment and priming times for H1. The effect size is estimated by Cliff's Delta (δ). Teams are sorted by decreasing task performance i.e. increasing *error*. The horizontal line separates well-performing teams (that found a correct solution) from badly-performing teams. U is the U statistic, and p is the 'two-sided' p-value of a Mann-Whitney U test (without continuity correction as there can be no ties, via our unique token number assignment).

	priming			establishment		
Team	U	p	δ	U	p	δ
28	1600.5	$< .05$	0.45	828.5	$< .05$	-0.25
17	823.0	$< .05$	0.49	431.0	.12	-0.22
20	1746.5	.15	0.16	1139.5	$< .05$	-0.24
11	2229.0	$< .05$	0.82	1624.0	$< .05$	0.33
9	8805.5	$< .05$	0.66	6561.5	$< .05$	0.24
47	3189.0	$< .05$	0.54	1627.0	$< .05$	-0.21
10	982.0	$< .05$	0.51	660.0	.92	0.02
8	883.0	.48	0.10	461.0	$< .05$	-0.43
7	562.0	$< .05$	-0.29	259.0	$< .05$	-0.67
18	2136.0	$< .05$	0.44	1326.0	.34	-0.11

***A*:** *Mount Zurich to Mount Bern*.
. . . . . .
***B*:** {*F uh*} isn't already *Mount Zurich to Mount Bern* isn't it connected?
. . . . . .
***A*:** So if we erase Mount *Zurich* or Mount Ber- to Mount Bern or Mount *Zurich* to Mount Gallen?
***B*:** Wait {*F uh*} *Zurich* to Mount . . . ? (7.2)

Speakers are able to prime these expressions without using fillers, but this does not guarantee that the primed expression was fully understood by the listener, as shown by the results of the Mann-Whitney U test. Following the idea of IF and IG pairs in the dialogue, the IG (speaker) tends to be in the abstract view, and can see the minimally represented names of gold mines. They need only concentrate on the gold mines that have the lowest cost to connect. Perhaps the reason why the IG does not use (by our measures) fillers as much is because they can see the task-specific referents available to them written down on the map. Expressions that contain task-specific

Table 7.3: Summary statistics for fillers and the established routines of teams, sorted by decreasing *error*.

	count		median (%)		
Team	filler	routine	filler	priming	establishment
28	38	58	11.6	3.5	15.2
17	27	41	14.0	6.2	17.1
20	59	51	20.7	18.1	22.8
11	35	70	34.5	3.5	24.4
9	81	131	46.3	11.4	37.4
47	56	74	19.5	6.7	21.9
10	20	65	20.1	8.7	20.4
8	26	62	19.8	15.7	36.0
7	27	59	24.9	25.9	37.8
18	52	57	10.8	4.3	11.8

referents could successfully become part of the IG's expression lexicon when given these new referents. The IG may also feel at ease to read out these expressions. The IF (listener) must follow the instructions with actions, and since they can not see the cost of adding and removing edges, they must search for the specific gold mine names given by the IG. This could create an uncertainty in the IG, and bring about a need to clarify. Thus for **H1, overall, we see that for most of the teams, the fillers tend to occur visibly after priming times, and around establishment times.**

Are the two fillers used differently in the process of routine formation of task-specific referents? The results for the filler "um" are given in Table 7.4. We observe that for the filler "um", **priming times** differ significantly for six out of ten teams (by Mann-Whitney U test $p < .05$). Like both fillers together, we see positive δ values, except for one team. Most teams thus have the filler "um" occurring after priming times. We see that the magnitude of δ now for the teams that are statistically significant ranges from small to large effect sizes. This means that while the filler "um" predominately occurs after priming times (positive δ values), the range of δ effect sizes shows that the filler "um" has little to larger overlap with priming times. We observe that the filler "um" and **establishment times** differ significantly for some of the teams (for six out of ten teams by Mann-Whitney U test $p < .05$). We still see that the δ values are predominantly negative (for eight out of ten teams), and that that the magnitude of δ for most teams is small, with the exception of two teams. This means that the distributions of the filler "um"

Table 7.4: The results for the Mann-Whitney U test that compares the distribution of the filler "um" time with establishment and priming times for H1. Please see Table 7.2 for a full description.

	priming			establishment		
Team	U	p	δ	U	p	δ
28	484.0	$< .05$	0.52	227.0	.13	-0.29
17	236.0	.19	0.28	135.0	.21	-0.27
20	1173.5	.14	0.18	713.5	$< .05$	-0.28
11	1339.0	$< .05$	0.91	1022.0	$< .05$	0.46
9	4384.0	$< .05$	0.67	3317.0	$< .05$	0.27
47	1694.0	$< .05$	0.43	834.0	$< .05$	-0.30
10	733.0	$< .05$	0.41	474.0	.59	-0.09
8	766.0	.23	0.18	414.0	$< .05$	-0.36
7	88.0	.40	-0.25	42.0	$< .05$	-0.64
18	1708.0	$< .05$	0.36	981.0	.06	-0.22

differs with a small effect size compared to the distributions of establishment times, especially considering that the token number values do not overlap; i.e. we do not expect an effect size of 0. We thus interpret the filler "um" does occur around (from positive and negative δ values) establishment times (with larger overlap from small effect sizes).

The results for the filler "uh" are given in Table 7.5. We observe that for the filler "uh", **priming times** differ significantly for most teams (eight out of ten teams by Mann-Whitney U test $p < .05$). The filler "uh" compared to the filler "um" tends to differ in distribution compared to the distribution of priming times (exhibited by the significance of p for majority of the teams compared to six of the teams for filler "um"). Like both fillers together, we see positive δ values, except for one team. Most teams thus have the filler "uh" occurring after priming times. We see that the magnitude of δ now for the teams that are statistically significant ranges from small to large effect sizes, but predominantly large effect sizes. This means the filler "uh" occurs visibly after priming times, indicated by the large range of δ (with little overlap from the large effect size). Thus we see that **the filler "uh" is used slightly differently in terms of priming times compared to the filler "um", with stronger evidence to support that the filler "uh" occurs visibly after priming times compared to the filler "um"** (which may have a general tendency to occur after priming times, but may range from little to large overlap).

We observe that the filler "uh" and **establishment times** differ signif-

Table 7.5: The results for the Mann-Whitney U test that compares the distribution of the filler "uh" time with establishment and priming times for H1. Please see Table 7.2 for a full description..

	priming			establishment		
Team	U	p	δ	U	p	δ
28	1058.5	$< .05$	0.40	550.5	$< .05$	-0.27
17	587.0	$< .05$	0.59	296.0	.23	-0.20
20	522.0	.62	0.08	375.0	.15	-0.23
11	854.0	$< .05$	0.74	591.0	.23	0.21
9	3791.5	$< .05$	0.65	2762.5	.06	0.21
47	1495.0	$< .05$	0.68	793.0	.43	-0.11
10	249.0	$< .05$	0.92	186.0	.15	0.43
8	96.0	.93	0.03	42.0	.11	-0.55
7	460.0	$< .05$	-0.29	215.0	$< .05$	-0.67
18	428.0	$< .05$	0.88	345.0	$< .05$	0.51

icantly for only a few of the teams (for three out of ten teams by Mann-Whitney U test $p < .05$). This means that we cannot conclude that the distributions of the filler "uh" compared to the establishment times significantly differs, like we could for the filler "uh" with priming times. We see that the δ values are negative for some of the teams (negative values for six out of ten teams). We still see mostly negligible-small effect sizes, indicating that the placement of the filler "uh" still occurs around the vicinity of establishment times (however, by this method, did not differ in distribution significantly). Thus **there is stronger evidence to support that compared to the filler "uh", the filler "um" differs significantly in distribution with an established expression, but with larger overlap given the negligible-small effect size**. Thus individually, the filler "uh" tends to occur visibly after priming times, while both fillers tend to occur around establishment times.

When we look at Table 7.6 compared to Table 7.7, we do not see a tendency for all teams to use one filler more than the other (i.e. some teams use the filler "um' more compared to "uh" and vice versa). However, the results as discussed inspecting Table 7.4 and Table 7.5, suggest that the **while the two fillers have a differing rate of use within teams (accounting for stylistic differences), they may have a different function in the overall process of a routine formation**. That is, considering all teams as a whole, there is stronger evidence to support that the filler "uh" differs in distribution compared to priming times, and tends to occur visibly after

Table 7.6: Summary statistics for the filler "um" and the established routines of teams, sorted by decreasing *error*.

	count		median (%)		
Team	"um"	routine	"um"	priming	establishment
28	11	58	12.3	3.5	15.2
17	9	41	7.5	6.2	17.1
20	39	51	18.6	18.1	22.8
11	20	70	37.1	3.5	24.4
9	40	131	51.7	11.4	37.4
47	32	74	16.7	6.7	21.9
10	16	65	17.9	8.7	20.4
8	21	62	21.3	15.7	36.0
7	4	59	21.7	25.9	37.8
18	44	57	8.9	4.3	11.8

priming times. There is stronger evidence to support the that the distribution of the filler "um" differs compared to the distribution of establishment times, though both fillers tend to occur in the vicinity, given the larger overlap from small effect sizes.

The filler "um" might be used more in the priming and establishment of a routine expression compared to the filler "uh". From the results, we further constrain the distance between each individual filler by maximally 1 token before either priming or establishment times, to observe if there are specific differences in the way the two fillers are used in this context. We see a that the filler "uh" is not present before the primed expression for six out the ten teams, compared to "um", which did not have a filler present before a primed expression for only three teams. In Table 7.8, we see that when we increase the distance in tokens, the mean percentage of "um" used increases before the primed expression the most (i.e. when the speaker introduces a new expression into the shared vocabulary space for the first time), compared to an increase in either filler before an established expression. Furthermore, we see greater evidence to support that the filler "um" is used in the establishment of routine expressions, with the filler "uh" not used at all before the established expression for seven out of the ten teams. When we calculate the mean percentage of each filler used before a routine expression is formed (see Table 7.8), it is $uh = 7.33\%$ and $um = 13.80\%$ respectively (with priming $uh = 4.26\%$ and $um = 6.95\%$ and establishment $uh = 3.06\%$ and $um = 6.84\%$). While this partially may be accounted for in speaker style, all teams still use *both fillers* (with half the teams less "um"

Table 7.7: Summary statistics for the filler "uh" and the established routines of teams, sorted by decreasing *error*.

	count		median (%)		
Team	"uh"	routine	"uh"	priming	establishment
28	26	58	9.7	3.5	15.2
17	18	41	14.6	6.2	17.1
20	19	51	22.1	18.1	22.8
11	14	70	26.6	3.5	24.4
9	35	131	40.5	11.4	37.4
47	24	74	23.4	6.7	21.9
10	4	65	26.8	8.7	20.4
8	3	62	18.3	15.7	36.0
7	22	59	28.0	25.9	37.8
18	8	57	21.6	4.3	11.8

to a similar rate of both fillers, and half the teams using noticeably more "um"s), suggesting that they are used differently in discourse.

Summary Regarding H1, overall, we see that for most of the teams, the fillers tend to occur visibly after priming times, and around establishment times. When we look at the use of fillers individually, there is greater evidence to support that overall, that the filler "uh" occurs after priming times (with little overlap from large effect sizes), and that both fillers occur around establishment times (however, with larger overlap given the smaller effect size). These results are for all fillers. However, when we specifically consider fillers constrained to a certain distance *before* a primed/established expression, we see that (while stylistically one team may use more of one filler than the other), the filler "um" is used more than "uh" in the specific role of forming routine expressions, particularly with primed expressions – i.e. when the speaker introduces the new expression into the shared vocabulary space. Similar to monologues, what is fascinating to us is thus that the filler **"um" in dialogues is utilised more than the filler "uh" in the context of information sharing specific to the primary signal**. There are however, several other functions of fillers outside this specific role of the development of a shared vocabulary in dialogue, such as in the role of holding the speaker turn, interrupting and so on. This is evidenced also by the mean percentage of fillers used in other contexts.

Filler	distance	routine	priming	establishment
um	0	8.41	3.62	4.78
	+1	13.80	6.95	6.84
	+2	19.43	11.13	8.29
	+3	22.89	12.98	9.91
uh	0	6.94	3.49	3.04
	+1	7.33	4.26	3.06
	+2	13.17	7.08	6.09
	+3	16.06	7.50	8.56

Table 7.8: The mean percentage of the filler "um" and "uh" that occurs from 0 to +3 tokens before either priming or establishment times (routine). Teams are sorted by decreasing *error*.

7.5.2 RQ2: (Global effect of fillers) How do the interlocutor's use of fillers relate to their task success?

H2 Fillers that precede the priming and establishment of task-specific referents are associated with task success.

To investigate the association of fillers occurring before routines among performance and learning scores, we computed ORs. To interpret the results, when $OR = 1$[8], the presence of the percentage of fillers that occur before routines (exposure) is not associated with the odds of either HP/HL, nor LP/LL (i.e. no association of the expo-sure with outcome). When $OR > 1$, the presence of the exposure is associated with higher odds of HP/HL (positive association). When $OR < 1$, the presence of the exposure is associated with higher odds of LP/LL (positive association with decrease of HP/HL).

For task performance The results of the test show $OR = 1.37$ ($p > .05$, $95\%\, CI : 0.86\text{-}2.11$). While the OR is > 1, we cannot conclude on whether the presence of the percentage of fillers that occur before routines (exposure) affects the odds of neither HP nor LP (large variability with lower bound $CI < 1$, and upper bound $CI > 1$, accounting for the null hypothesis). This may be in part due to there not being a straightforward way to split the data based on performance. When observing the high performing teams in Table 7.3, we see that only 3 out of the 10 teams were high performers and found an optimal solution (the percentage of successful transcribed teams is 30% compared to 21% of successful teams in the the whole dataset)). Out of these teams, team 20 performed well. However, they did not learn (in fact,

[8]null hypothesis

Table 7.9: The percentage of fillers for that occur maximally 1 token before priming or establishment times (routine), compared to the percentage of fillers occurring in other contexts. Teams are sorted by decreasing *error*. The average percentage of fillers that occur before routines is = 12.21%, priming times = 5.33% and establishment times = 6.88% respectively.

Team	routine	priming	establishment	other
28	23.68	21.05	2.63	76.32
17	7.41	3.70	3.70	92.59
20	20.34	10.17	10.17	79.66
11	8.57	0.00	8.57	91.43
9	8.64	3.70	4.94	91.36
47	7.14	5.36	1.79	92.86
10	10.00	0.00	10.00	90.00
8	15.38	15.38	0.00	84.62
7	3.70	3.70	0.00	96.30
18	17.31	5.77	11.54	82.69

"unlearnt", as shown with a negative learning outcome Figure 3.4). In the metadata, if we look at the the progression of costs for submitted solutions, team 20 abruptly went from a high cost solution to an optimal solution (described further in Norman et al. (2021)). This informs us that they "got lucky" in finding a solution, and are not necessarily a high performing team. Similarly, with the low performing teams, it is difficult to categorise the teams as such, as the task was not specifically designed for the the teams to find an optimal solution (as is evidenced by the skewed *error* scores towards "low performance"). Thus due to the lack of demarcation between high and low performing teams, it is difficult to draw inferences on their use of fillers.

For learning outcomes The results of the test show $OR = 0.43$ ($p < .05$, 95% CI : 0.24-0.80). In this case $OR < 1$, indicating that the presence of fillers occurring before routines gives a higher odds of LL. According to the results of the *OR*s test, fillers occurring before routines reduce the odds of HL (67% lower given the presence of a filler before a routine) of the team. Interestingly, these findings similar to our findings in monologues, which was that the use of fillers preceding routines is associated with lower *learn* scores but the impact is smaller; i.e. the more fillers are used in this way, the greater the odds of LL.

While in Table 7.2, we saw the general trend of fillers occurring visibly after priming times, and around establishment times. When we are stricter

with the distance between filler and priming/establishment tokens as given in Table 7.9, we see that i) the filler "um" is used more than "uh" in the specific role of forming routine expressions, particularly with primed expressions – i.e. when the speaker introduces the new expression into the shared vocabulary space and that ii) there are other ways fillers are used in a dialogue context, given our strict distance criteria. Since our main aim was to propose methodologies to efficiently compute verbal alignment contexts (what we consider, the primary signal), and then study the interaction of fillers with verbal alignment, this does not discount for when fillers are used *in other types of communicative scenarios*, such as holding the speaker turn.

We are modest in our conclusions given the small sample sizes of teams with HL/LL (while the analysis in RQ1 was on all 20 speakers and $\approx$ 4 hours of audio). Given the results of H1, it would suggest that fillers used in the establishment of routines highlight some hesitation of the interlocutors in establishing a shared vocabulary, which leads to greater odds of LL. It is interesting, given that there is research to suggest that clarification requests highlight possible miscommunication that interlocutors may have been unaware of otherwise – leading participants to perform better (Mills, 2014), suggesting these results need to be verified with a larger sample size. Our results differ from Mills (2014) in that i) our data uses completely spontaneous speech transcript, rather than a chat (text) tool, and thus the clarification requests may not have contained fillers in this context and ii) performance on a task does not guarantee learning (for example, there can be unproductive performers (Kuhn, 2015)).

Thus regarding H2, we can not conclude that the usage of fillers that occur before the priming or establishments times are for better performing teams. However, using this analysis, there is evidence to suggest that fillers used in the development of a shared vocabulary (or task specific referents) may have an negative association on the outcome of learning – highlighting some hesitation of the interlocutors.

7.6 Conclusion

Our broad objective of this chapter was to develop methodologies to explore how fillers (a *collateral signal*), interact with the *primary/lexical signal*, or what was said in essence. However, so far, our previous work had only been based on monologues; where the speaker is conscious of an *unseen* listener, but unlike dialogue, there is not a dynamic and iterative communication processes between the interlocutors. Thus, we needed to account for this context of listener-speaker dynamic when considering the primary signal, compared

to our previous approach. To study the interaction of these two signals in dialogues, we adapted a rule-based algorithm specifically to study *verbal alignment* in situated activities. Alignment itself is the process of interlocutors forming shared representations (in this case, shared lexical representations) in a dialogue. It has been shown that a dialogue is successful when there is alignment between the interlocutors, at different linguistic levels. However, research on alignment in dialogue shows a lack of automatic measures suited the the medium of communication; i.e spontaneous speech. Thus to investigate this relationship between fillers and verbal alignment, we considered the corpus of data (dialogue transcripts from audio files and action logs) generated by teams of two children engaged in a collaborative learning activity, called JUSThink, which aims at providing an intuitive understanding of graphs and spanning trees.

Collaborative learning activities are a particularly interesting type of collaborative task, due to their "multi-layered goals", typically *immediate, information sharing goals* between the (dialogue) participants, and broader dialogue level goals such as *success in the task*, which includes performance goals (e.g., finding the solution to a math problem) and deeper, learning goals (e.g., understanding the notion of equation[9]). Thus, in this chapter our second objective is to study how this interaction of the primary and collateral signal could lead to dialogue level task success.

Our methodological contribution is to first propose an adaptation on existing rule based verbal (or lexical) alignment algorithms for situated dialogues, to study the alignment of expressions specifically related to the situated activity. Our results show that overall, **fillers tend to occur visibly after the first time an expression is introduced into the discourse** (priming) by one interlocutor, **and around the time it is repeated by the other interlocutor**, thus becoming part of the *shared vocabulary* (establishment). Fillers in general here are used to clarify the instructions verbalised by the interlocutor. When we specifically consider fillers constrained to a certain distance *before* a primed/established expression, we see that (while stylistically teams may use more of one filler than the other), the filler "um" is used more than "uh" in the specific role of verbal alignment, particularly with primed expressions – i.e. when the speaker first introduces the new expression into the shared vocabulary space. An important contribution about these results, is that we found *both in monologues and dialogues* that the filler "um" is utilised more in the introduction of the new information (which

[9]In Chapter 3, we also contribute a dataset of anonymised, transcribed children dialogues, event logs of their task progression, and code, which are all made publicly a available to either reproduce our results or to use for further research.

in both cases we consider as new referents related specifically to the topic of discourse) than the filler "uh". Thus based on our previous work in monologues, we could conclude two fillers are used differently (in agreement with Le Grezause (2017)), and **the filler "um" carries more information sharing properties than the filler "uh"; as is evidenced by their function in relation to verbal alignment in dialogues**. In terms of alignment contexts leading to task success, our results would suggest that fillers used in the establishment of routines highlight some hesitation of the interlocutors in establishing a shared vocabulary, which is associated with greater odds of lower learning. It was also difficult to find a demarcation between high and low performers, given that the activity was not designed specifically for interlocutors to find a solution, and $\approx 20\%$ of the entire dataset "performed well".

This chapter concludes our work on fillers. We have developed methodologies to show that for both monologues and dialogues, fillers locally are informative signals of communication, and that globally, they also contribute to discourse level phenomena, such as the listener's impression of the speaker or variables of task success. In both our works on monologues and dialogues, we try to introduce nuance and contextual usage of fillers, rather than extrapolating one characteristic of fillers as ground truth. We also found both fillers individually used differently in the context of information sharing, showing that they have different functions in discourse. Often fillers in dialogues are studied for their turn-taking properties, choosing this aspect to focus on the interaction between the listener and the speaker. We introduce a another layer of analysis (and indeed, we saw that the fillers we specifically study in this way can still account for fillers used in other ways), that is specifically fillers used in the context of one interlocutors introducing new expressions, to fillers used in the context of accepting these expressions as part of their shared representation space.

8 | What Was Done? "Oh" in Behavioural Alignment

8.1 Introduction and Background

In the previous chapter, our main goal was to study how the primary signal (what was said, in essence) interacts with the collateral signal (in this case, fillers) in a situated, task oriented dialogue. In this dialogue context, there is a more dynamic and iterative communication process between the interlocutors participating in the activity; a back-and-forth between the interlocutors in the aim to reach *common-ground, or mutual understanding*. Thus to study this interaction, we adapted a rule based algorithm for situated dialogues, to focus on the *verbal (lexical) alignment* of expressions related to the task; from the first time a new expression is introduced into the vocabulary by one interlocutor, to the time it becomes part of a *shared vocabulary* amongst interlocutors. We then studied the use of fillers in the context of this verbal alignment process, to find that fillers tend to occur around the time the interlocutor *accepts* the expression as part of the shared vocabulary, and that the filler "um" is utilised more in the process of alignment than the filler "uh" on this dataset. Given the results from Part II, our work studying the two fillers in both monologues and dialogues points to the filler "um" being used more in the context of information sharing, and thus the two fillers are used differently in discourse.

Our second goal, was to understand how these *local alignment contexts* build into a function of dialogue level phenomena; i.e. success in the task. In this regard, we found that fillers used in the development of a shared vocabulary may have a negative association with the learning outcomes of the task (i.e. how much the children learnt from the activity), highlighting some possible hesitation/ cognitive burden the interlocutors verbalised during the task.

As we discussed in the previous chapter, situated activities have an *in-*

terdependence on the physical environment; the interlocutors need to form expressions related to the situated activity, and take actions in the activity to further the task. Hence, a crucial part of a situated dialogue, is not simply what was *said* by the interlocutors, but the corresponding actions taken as a result – i.e what was *done* by the interlocutors. Thus, the main goal of this chapter is to study the role of disfluencies in *behavioural alignment*; a new alignment context we propose to mean when *instructions* verbalised by one interlocutor are *followed* with physical actions by the other interlocutor. In this chapter, we consider the primary signal of behavioural alignment, i.e. what was instructed leading to what was done *in response*. To do so, we propose an action-by-action, rule based algorithm, to automatically annotate the *follow-up* actions taken after (automatically inferred) instructions were verbalised. Thus while the previous chapter focused solely on the surface-level utterances, i.e. "What did the interlocutors say", this chapter builds on this with actions, i.e. "What did the interlocutors do *afterwards*". Since the nature of the activity is situated, we look into whether the interlocutors are stuck *saying* task specific referents, or whether the use of these referents spurs the interlocutors into action; i.e. into *doing*. Then, we study the role of the information marker in relation to the follow up actions taken in the physical environment from the situated activity. Like the previous chapter, our secondary goal is to observe how these local behavioural alignment contexts can build to dialogue level task success.

behavioural alignment

Concretely the **aim of this chapter** is study the interlocutor's use of the information marker "oh" in relation to the follow up actions taken in the situated activity. We consider the use of "oh" as an information management marker, because its verbalisation marks the *focus of speaker's attention*, which then also *becomes a candidate for the listener's attention*. This creation of a joint focus of attention allows for *transitions in the information state* (Schiffrin, 1987). We are interested in what is commonly known by the interlocutors regarding the task (as an information state). Thus instructions and corresponding actions that contain "oh" in their surrounding context may be used by the interlocutors to inform the other about some *change* in their information state. To the best of our knowledge, the use of "oh" has been studied more in the context of integrating information at the verbal level, but not a behavioural one. For this purpose, we firstly propose a rule-based algorithm automatically annotate the instructions specific to the situated activity. Then, we match using the metadata whether these instructions were followed by an action. We study:

1. **At the local level:** The role of the information marker "oh" in the behavioural alignment (instructions-to-actions) process. For this, we

automatically annotate the instructions of an interlocutor, as the verbalised instructions one interlocutor gives to the other, which we extract through their use of expressions specific to the task. Then we see what were the follow up actions of these instructions.

2. **At the dialogue/global level:** We study whether this process can build to the interlocutors' task success (measured by the team's *learning* outcomes)[1].

For a detailed background, please refer to the previous chapter, Sec. 7.2, where we discuss research on collaborative learning activities, and introduce terminology used in both chapters related to dialogue literature[2]. The rest of the chapter is organised as follows: In Sec. 8.2, we refer to the relevant sections in the thesis for a description of the dataset used, and outline our research questions (RQs) and hypotheses. In Sec. 8.3, we describe our methodology for our behavioural alignment model to infer instructions-to-actions, and methodology for the specific RQs and hypotheses. Sec. 8.4 gives the results and discussion, while Sec. 8.5 discusses the conclusion of the experiments.

8.2 Research Questions

Dataset Please see Chapter 3 for a detailed description of the JUSThink Dialogue and Actons Corpus used in this chapter, particularly for i) **At an utterance level:** why this dataset was suited to study alignment of actions, ii) **At dialogue level:** a background on how the task success variable of learning outcomes (*learn*) is calculated, and iii) A detailed description of the transcription campaign of this dataset. This chapter thus investigates the role of the information marker "oh" with i) the use of instructions manifesting in the interlocutors' actions within the task, and ii) whether this builds to an association with their task success.

Corresponding and different actions Previous research would suggest that the earlier interlocutors align with each other in terms of verbalised instructions and corresponding follow-up actions, the better they progress in the task (i.e., interlocutors being in alignment in a successful dialogue, see

[1]In this chapter, we do not use the labels of performance to measure task success, as there is not a clear demarcation between good performing and bad performing teams.

[2]We use the term dialogue literature to refer to relevant linguistic (e.g. psycholinguistics, discourse, conversation analysis) literature.

Sec. 7.2) thus the greater chance of task success. However, work on collaborative learning suggests that individual cognitive development (in our case, positive learning outcomes) happens via *socio-cognitive conflict* (Mugny and Doise, 1978; Doise and Mugny, 1984), and its regulation (Butera, Sommet, and Darnon, 2019). This means a verbalised instruction could be followed by a corresponding or a different action; as a different action could result in collaboratively resolving conflicts and together building a solution – resulting in task success. It is thus the required effort to construct the shared understanding together that results in collaborative learning (Dillenbourg, Järvelä, and Fischer, 2009). For behavioural alignment purposes, we focus on *corresponding* actions, but also follow-up actions that were *different* from the verbalised instructions, to gain a clearer picture into the alignment processes of the interlocutors.

instructions

For this purpose, we consider the *instruction*s of an interlocutor, as the verbalised instructions one interlocutor gives to the other, which we extract through their use of task-specific referents. A physical manifestation of this instruction could result in a *corresponding* edit action, or a *different* edit action. There also need not be an instruction that precedes an action, i.e. an interlocutor in the visual view (i.e. the one responsible for taking actions, but not able to see the cost of connecting nodes) can take less-informed actions by themselves, see Figure 3.2.

"Oh" as a consequence of actions in the dialogue We consider the use of "oh" as an information management marker, to mark a focus of a speaker's attention, which then also becomes a candidate for the listener's attention. This creation of a joint focus of attention allows for transitions in the information state (Schiffrin, 1987). In general, changes in the information state – or what is commonly known about the task – of the participating interlocutors should be due to physical actions related to the task[3], particularly because the situated activity has an *interdependence* on the physical environment.

"Oh" as a marker for new information has been studied in various scenarios. Aijmer (1987) for example, did a corpus analysis of "oh" to identify the contexts in which the marker is used, starting from a base general description of its usage as *a mental reaction to a stimulus*; e.g. "Oh, flowers!". Among the specific contexts identified, the most relevant to this work is "oh" used as a marker to verbalise reactions to *surprising information* (which for our purposes, we consider a change in the information state). Fox Tree and

[3]Indeed, specifically for this task, as stated in Chapter 3, the robot uses minimal terminology to explain the task to the interlocutors, leaving them to explore and figure out for themselves the best way to approach the task.

Schrock (1999) also studied the use of "oh" in online comprehension experiments to find that it is used by listener's to help them *integrate information* in spontaneous speech. They found for e.g. that recognition of words was faster after listener's heard "oh". It is to note, that Fox Tree and Schrock (1999) describe "oh" as a *discourse marker*, which they say are expressions that differ spontaneous talk from planned talk ("well", "I mean", "like"). We elaborate further on terminology in Chapter 2. We are interested in the use of "oh" in the context of what is commonly known by the interlocutors regarding the task. Thus, we consider the verbalisation of "oh", as an update in the interlocutor's knowledge regarding how to solve the situated activity, resulting from actions taken. To the best of our knowledge, the use of 'oh" has been studied more in the context of integrating information at the verbal level, but not a behavioural one.

RQ1: (Local effects of "oh") How do the interlocutors' use of information management markers relate to their behavioural alignment? In RQ1, we expect that in a situated task, when interlocutors verbalise the information management marker "oh", it is a direct consequence of some physical action that occurs in the task. Instructions that contain "oh" in their surrounding context could thus be used by the speaker to inform the listener about some change in their information state related to some physical manifestation in the situated activity. This can either be done in speaker turn utterance ("Oh, this is how we do the task") or as a backchannel (typically by the listener: "Oh okay, oh yeah . . . "). Thus we hypothesise that:

- **H1:** The information management marker "oh" should be present in the vicinity[4] of when instructions are followed by a corresponding or a different action.

RQ2: (Global effects of "oh") How does the interlocutors' use of information management markers relate to their task success? We expect that interlocutors that better verbalise their level of understanding the task to the other interlocutor (measured using the information management marker "oh") should have a greater association with task success. Specifically:

- **H2:** The use of "oh" in the alignment of actions is associated with task success.

[4]We use the term vicinity here; i.e either before or after here, because there is a circular process of interlocutors taking action based on what is known, which in turn updates what is known about the task.

8.3 Methodology

(mis)matched actions

If an instruction is verbalised (for e.g. to connect "Mount Basel to Montreux") by an interlocutor (IG), it could result in an action of connecting the two. We hence say an instruction *match*es an action when the instruction is executed by the other interlocutor (IF) via an action in the situated environment, within the period of a turn of views before they are swapped. We study the discrepancy created when the IF does not follow the IG, which we call *mismatch* of instructions-to-actions. In Eg. 8.1, the instruction (to connect Gallen to Davos) matches the action (connecting these two)[5]. In Eg. 8.2, we illustrate a dialogue excerpt that results in a mismatch [6]:

***A*:** What about Mount Davos to Mount, Saint Gallen?
***B*:** Because what if, you did if we could do it?
***A*:** What about Mount um Davos to Mount Gallen?
***B*:** Mount ...
***B*:** oh Mount Davos (8.1)
***A*:** yeah to Mount Gallen.
***B*:** to Mount Gallen yeah do that.
⟨*A* connects Mount Gallen to Mount Davos⟩
***B*:** Okay my turn. *[Match!]*

***B*:** Go to Mount Basel. *[Instruction to add an edge to Basel.]*
***A*:** That's, it's expensive.
***B*:** Just do it.
***A*:** You can't, you can't, I can't because there's a mountain there ... (8.2)
***A*:** So I'm going, so I'm going here.
⟨*A* connects Mount Interlaken to Mount Bern⟩ *[Mismatch!]*

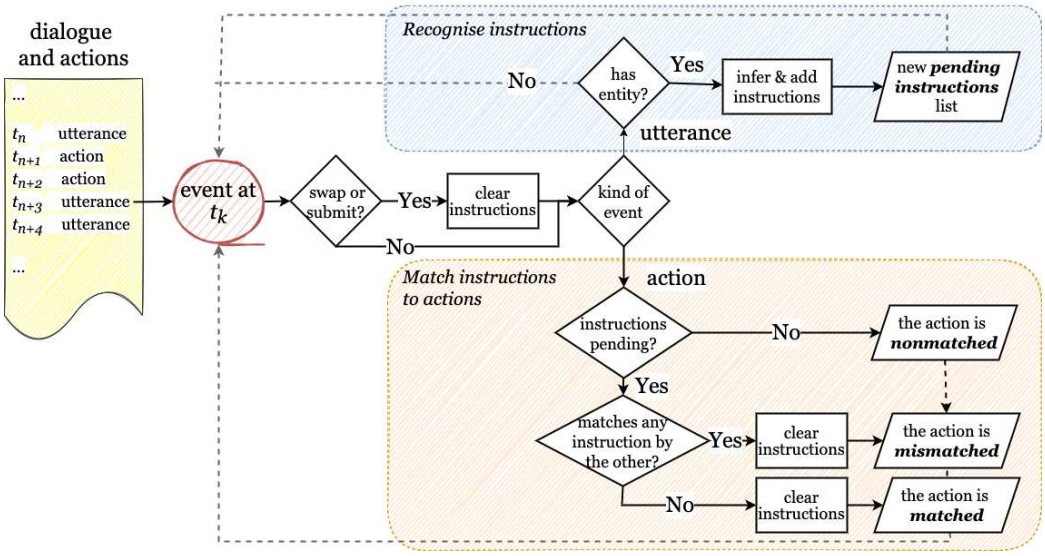


Figure 8.1: Representation of the schema in a flowchart (e.g. the parallelogram shows the annotated output), to annotate behavioural alignment between the interlocutors.

Adapting the situated alignment model to infer instructions-to-actions

Extracting Instructions We extract instructions from the utterances through the interlocutors' use of task-specific referents. In the schema as shown (Figure 8.1), we first have a process of checking if there is an input utterance ("is Input Utterance"), then checking whether the input utterance contains entity patterns. To check these entity patters ("Contains Entities"), we employ named-entity recognition (NER) feature of the Python library spaCy that performs this entity recognition. We add the node names of the mountains (e.g. "Montreux"), and also "verbs"; i.e. "add", "subtract" .... Then, if the input utterance contains our custom entity patters, then we automatically infer instructions from the utterance by joining these entities together (e.g. the result may be Add(Node_1,Node_2), because the interlocutor explicitly said the verb "add" and also the names of the mountains). When

[5]The complete code is available in the tools (*tools/6_ recognise_ instructions_ detect_ follow-up_ actions.ipynb*) that generates an annotated corpus (*processed_ data/annotated_ corpus*).

[6]From Team 8, extracted/detected by using our algorithm 4 and algorithm 5. The whole dialogue with the annotations is available in the annotated corpus (*processed_ data/annotated_ corpus*), available with the tools.

complete algorithm in Sec. 8.6, (supplementary materials).

Finding (mis)matched of instructions-to-actions To find (mis)matches of instructions-to-actions, we then follow the logic as given in the schema to determine whether there is an action at the time an instruction was inferred ("isAction"), and then checking whether the action matches the inferred instruction.

In Eg. 8.1, Gallen-Davos was the result of a negotiation, rather than a complete given instruction by the IG. B says in one utterance "oh Mount Davos" and then in another "to Mount Gallen yeah do that", resulting in two cached inferred instructions; (Davos,?) and (Gallen,?) respectively. This accounts for some amount of multiple speaker turn[7] negotiations, and possible other ways of referring to task-specific referents (e.g "Now go from **there** to Davos"). Though an instruction could be carried out after the views swap again, i.e. in the following turn, in our algorithm, the pending instructions are cleared at every swap, resulting sometimes in a *nonmatch*, or when an action occurred but there was no inferred instruction (which is different from a mismatch of instructions to actions). This process is shown in the schema as "Add to Pending Instructions" (to cache instructions), and "Clear Instructions" (to clear instructions). A full and concrete example of the added annotations of our automatically inferred instructions-to-actions is given in Table 8.1.

nonmatched action

[7]Note, we specifically use "speaker turn" to distinguish from a turn in the collaborative activity; i.e. every swapping of views. Please see Sec. 3.2 for further details.

Table 8.1: An example from the output of recognising instructions and detecting follow-up actions (by RECOGNISE-INSTRUCTIONS (algorithm 4) and MATCH-INSTRUCTIONS-TO-ACTIONS (algorithm 5)), from Team 10. *View* denotes which view the interlocutor is in (refer to Figure 3.2), either *abstract* (Ab) or *visual* (V). *Annotations* denotes the automatically inferred instructions and follow-up actions in the activity. For example, $Instruct_A$ indicates that interlocutor A has given an instruction to add two nodes (inferred from referents), which can be partially recognised *(Gallen,?)*. As shown, the algorithm builds up (or "caches") instructions until an edit action is performed ('-' in Utt.). Note, since B is in the visual view, their inferred instruction is deliberately not matched.

	Utt.	View	Verb	Utterance	Annotations
A	198	Ab	says	Maybe we start from, Mount Zermatt ?	Instruct_A(Add(Zermatt,?))
B	199	V	says	No lets do Mount Davos to, where do you wanna go?	Instruct_A(Add(Zermatt,?), Instruct_B(Add(Davos,?))
A	200	Ab	says	. . . to Mount, St Gallen.	Instruct_A(Add(Zermatt,?), Instruct_B(Add(Davos,?), Instruct_A(Add(Gallen,?))
B	201	V	says	Okay.	*As previous*
B	-	V	adds	Gallen-Davos	Instruct_A(Add(Zermatt,?), Match_B (Instruct_A(Add(Gallen,?)))

Table 8.2: OR contingency table, where MA stands for the cumulative percentage of "oh" that occurs around a (mis)matched action for all HL (a) LL (b) teams, and OC the remaining cumulative percentage of "oh" used in other contexts ((c) and (d) respectively).

		Outcome	
		HL	LL
Exposure	MA	a	b
	OC	c	d

Research Questions

RQ1: (Local effects of "oh") How do the interlocutors' use of information management markers relate to the alignment of their instructions-to-actions? To investigate H1, we check if the distributions of (mis)matches and 'oh' marker times are significantly different by Mann-Whitney U test, and estimate the effect size by computing Cliff's Delta. This calculation differs from the previous chapter, as actions are not part of an utterance (compared to establishment time for example); hence we cannot use the order of tokens as was previously used. Given that we have *utterance level timing*[8] we instead compare the end times of the utterance that contains the marker with the (mis)matched action times.

RQ2: (Global effects of "oh") How does the interlocutors' use of information management markers relate to their behavioural alignment? To investigate H2, we use Odds Ratios (ORs) to investigate whether the use of "oh" around (mis)matched actions is associated with task success, as was done in the previous chapter (see Sec. 7.4). Firstly, we use the same split for High Learning (HL) and Low Learning (LL) teams. Based on the previous chapter, since there is no clear demarcation for teams for performance, we only use the measure *learn* to observe task success. In this case, we consider the **timing** of "oh" and actions (compared to token distance as was used previously). We take the percentage of "oh"s that occur around the scaled time of a (mis)matched action, and the percentage of "oh"s that occur in other times. We then calculate the odds of HL or LL (outcome), given the occurrence of "oh"s (exposure) around (mis)matched actions, compared to the occurrence of "oh"s that occur at another time. Please see Table 8.2 for a contingency table.

[8]Forced alignment algorithms and ASR did not work well on this dataset.

8.4 Results

8.4.1 RQ1: (Local effects of "oh") How do the interlocutors' use of information management markers relate to their behavioural alignment?

H1 The information management marker "oh" should be present in the vicinity of when instructions are followed by a corresponding or a different action.

To investigate H1, we consider the distribution of the information management marker "oh" through time, in relation to the distribution of matches and mismatches. Firstly, we note that the "oh" marker occurs 333 times in the transcripts (average per transcript = 33.3, SD = 20.3). See Table 8.3 for the number of utterances that contain one or more "oh"s for each team. In particular, we check if the distributions of "oh" times and (mis)matches are significantly different by Mann-Whitney U test, and estimate the effect size by computing Cliff's Delta (δ). The results of the tests for each team are given in Table 8.4. Here, $\delta = -1$ would mean that all 'oh's occur earlier than (mis)match times, and 1; that all "oh"s occur later[9].

Table 8.3: Summary statistics for the "oh" and (mis)matched instructions-to-actions, sorted by decreasing *error*. Count is given as the number of utterances that contained (mis)matched actions.

	count				median (%)		
Team	oh	match	mismatch	match/mism.	oh	match	mismatch
28	24	19	13	1.5	61.7	62.7	73.6
17	20	23	10	2.3	61.1	61.1	69.4
20	65	24	6	4.0	35.0	58.0	42.3
11	15	34	18	1.8	57.3	59.5	38.5
9	29	28	9	3.1	42.9	59.9	70.2
47	29	28	21	1.3	53.4	56.8	75.7
10	18	36	15	2.4	47.6	67.3	65.3
8	55	30	11	2.7	36.3	65.3	86.3
7	51	43	13	3.3	49.7	65.5	79.7
18	12	25	7	3.6	69.6	69.5	75.0

[9]As in the previous chapter, δ ranges from -1 to 1, where 0 would mean that the group distributions overlap completely; whereas values of -1 and 1 indicate a complete absence of overlap between the groups.

For all teams except one, the value of δ is negative, indicating that the distribution of "oh" tends to occur earlier than (mis)match times. We observe that "oh" and (mis)match times do not significantly differ for half of the teams (Mann-Whitney U test $p < .05$). We see δ ranging from -0.54 for Team 8, to 0.09 for Team 11: this indicates that "oh" occurs differently compared to the (mis)match times within the teams. For the teams that had significantly different distributions, the effect size ranges from negligible ($|\delta| < 0.147$) to medium ($|\delta| < 0.474$), with the exception of Team 8, that has large effect size ($\delta = -0.54$). For all teams, we see predominantly (seven out of ten teams) negligible to small effect sizes, indicating a larger overlap given the smaller effect size. Thus, regarding H1, the results are mixed; we see that **the distribution of "oh" and (mis)matched actions differs significantly for half the teams**, and the predominantly negligible to small effect sizes indicate a larger overlap of the distributions. Thus, for half of the teams, by this measure, we could not find a significant difference between the distribution of "oh" in the dialogue, with either matched or mismatched actions – suggesting that they do tend to occur around the same time.

Table 8.4: The results for Mann-Whitney U tests that compare the distribution of information marker (i.e. "oh") times, with (mis)match times for H1. The effect size is estimated by Cliff's Delta (δ). Teams are sorted by decreasing task performance i.e. increasing *error*. The horizontal line separates well-performing teams (that found a correct solution) from badly-performing teams. U is the U statistic, and p is the 'two-sided' p-value of a Mann-Whitney U test (without continuity correction as there can be no ties, via our unique token number assignment).

	(mis)match		
Team	U	p	δ
28	344.0	.51	-0.10
17	318.0	.83	-0.04
20	715.0	$< .05$	-0.27
11	427.0	.58	0.09
9	429.0	.16	-0.20
47	488.0	$< .05$	-0.31
10	264.0	$< .05$	-0.42
8	522.0	$< .05$	-0.54
7	841.0	$< .05$	-0.41
18	157.0	.36	-0.18

Is "oh" used differently for matched and mismatched actions? We compare individually the distribution of "oh" with matched action times (i.e. automatically inferred instructions that had corresponding actions), or mismatched action times (i.e. automatically inferred instructions that had different actions) using the same methodology as described in Table 8.3. The results are given in Table 8.5 (for matched actions) and Table 8.6 (for mismatched actions) respectively. For matched actions, we see that only four out of ten teams differ in distribution with "oh" times, and for mismatched actions, three out of ten teams. The teams that have a significant difference in their distributions are teams that did not perform well (which may include team 20 for matched actions, which simply lucked out in finding a solution – as discussed in Chapter 7). Furthermore, for matched actions, we see a gradual rise in the effect size as performance decreases (although not perfectly linear), indicating that **the better performing teams had "oh" occurring more around matched instruction-to-action times (with small effect size to indicate larger overlap of the groups), compared to badly performing teams (with larger effect size and smaller overlap)**. However, since there is no binary distinction between good performers and bad (the task was not designed for the interlocutors specifically to complete the task, with only 20% of the dataset consisting of "good performers"), we consider these findings as qualitative.

Given the mean of the absolute δ values (*mean* $|0.22|$ and $|0.33|$ for matched and mismatched teams), the results of the distributions would indicate that "oh" is used slightly more in the context of matched actions than mismatched actions. Perhaps "oh" may be used more when the IF wants to explicitly signal to the IG that they are following the IG's instructions (which would result in matched actions) compared to when they do not (resulting in a mismatched action). The only negative effect sizes would indicate that "oh" times predominantly occur before action times, for both matched and mismatched actions; i.e. the IF follower could inform the IG using the marker before taking action. However, we also see in Table 8.3 that matched actions occur far more (according to our automatically inferred instructions-to-actions) than mismatched actions and thus cannot conclude whether "oh" is used more for matched actions. In general, using our methodology, we thus infer more matched actions than mismatched ones, and there may be variations with the way "oh" is used for matched and mismatched actions, though this requires further investigation (and indeed, this brings up a fascinating question of whether "oh" is strategically used by the IF to inform the IG that they are following instructions compared to when they do not follow instructions).

Table 8.5: The results for Mann-Whitney U tests that compare the distribution of information marker (i.e. "oh") times, with only matched action times for H1. Please see Table 8.4 for a full description.

	match		
Team	U	p	δ
28	221.0	.86	-0.03
17	229.0	.98	0.00
20	533.5	$< .05$	-0.32
11	241.5	.77	-0.05
9	329.5	.22	-0.19
47	316.5	.15	-0.22
10	176.0	$< .05$	-0.46
8	423.0	$< .05$	-0.49
7	670.0	$< .05$	-0.39
18	134.0	.60	-0.11

Table 8.6: The results for Mann-Whitney U tests that compare the distribution of information marker (i.e. "oh") times, with only mismatched action times for H1. Please see Table 8.4 for a full description.

	mismatch		
Team	U	p	δ
28	125.5	.33	-0.20
17	86.5	.55	-0.14
20	184.0	.82	-0.06
11	185.5	.07	0.37
9	97.0	.25	-0.26
47	173.0	$< .05$	-0.43
10	87.5	.09	-0.35
8	99.0	$< .05$	-0.67
7	170.5	$< .05$	-0.49
18	25.5	.16	-0.39

Table 8.7: The results for Mann-Whitney U tests that compare the distribution of information marker (i.e. "oh") times with nonmatch times. Please see Table 8.4 for a full description.

	nonmatch		
Team	U	p	δ
28	470.0	.38	-0.13
17	568.0	$< .05$	0.32
20	1088.0	.14	0.20
11	780.0	.10	0.27
9	349.0	.14	-0.22
47	1020.0	.34	-0.12
10	317.0	$< .05$	-0.38
8	1177.0	$< .05$	-0.35
7	4439.0	.09	0.16
18	189.0	.13	0.31

The distribution of "oh" with nonmatched action times In order to see whether the information management marker "oh" occurs in the vicinity of nonmatched actions (i.e., when actions taken by the interlocutors were not the result of an automatically inferred instruction), we use the same methodology as stated in Table 8.3. This is to see if the marker "oh" occurs generally in the vicinity of actions – including ones that our method did not associate with an instruction, or did not have explicit associated verbalised instructions. The results for the distribution of "oh" times with nonmatched actions are as given in Table 8.7. Only three out of seven teams had the distribution of "oh" and nonmatched actions times significantly differ.

Please follow Table 8.9 for an example of our automatically annotated nonmatched actions. What is interesting (as shown in the table), is that we see that nonmatched actions (given as $\text{Nonmatch}_{\text{Interlocutor}}$ in the table) frequently occur when the IF follows their own intentions, or act without explicit verbalised instructions of the IG in the abstract view. For e.g., as seen after Utterance 10, $\text{Nonmatch}_{\text{A}}(\text{Do}_{\text{A}}(\text{Add}))$, where A connects two nodes. **For seven teams, using this method we did not find significant differences in the distribution of "oh" with nonmatched actions, indicating that "oh" may occur around nonmatched action times**. The predominantly negligible to small δ values further indicate that this is the case, with larger overlap given the smaller effect size. Integrating such information (i.e. what was said by the interlocutor) with real, physical actions occurring in the situated activity gives more insight into the way

the task is progressing and the information state of the interlocutors, and indeed could be useful to integrate in the automatic and online processing of situated activities. **Thus, these results highlight the usefulness of the information marker "oh", in that indeed, in a situated activity, when interlocutors use the marker, it may be associated with actions taken in the activity and consequently, updates in the information state of what is known about the activity.**[10]

8.4.2 RQ2: (Global effects of "oh") How does the interlocutors' use of information management markers relate to their task success?

H2 The use of "oh" in the alignment of actions is associated with task success.

To investigate the association of "oh" in the alignment of actions with task success, we computed ORs. To interpret the results, when $OR = 1$[11], the presence of the percentage of "oh"s that occur around the time of (mis)mstched actions is not associated with the odds of either HL, nor LL (i.e. no association of the exposure with outcome). When $OR > 1$, the presence of the exposure is associated with higher odds of HL (positive association). When $OR < 1$, the presence of the exposure is associated with higher odds of LL (positive association with decrease of HL).

For learning outcomes The results of the test show $OR = 1.97$ ($p < .001$, $95\%\ CI$: 1.32-2.96). In this case $OR > 1$, indicating that the presence of "oh" occurring around the same time of (mis)matched actions gives a higher odds of HL. This means that **the interlocutors' verbalisation of "oh" around (mis)matched action times have a positive association with better learning outcomes.** These results indicate that when the marker "oh" is verbalised in the context of *(mis)matched actions*, following from previous work, it may be used in the context of the interlocutor expressing some change in their information state. What is of important however, is the distinction between *the right kind of information gain required to move*

[10] According to Schiffrin (1987) the information management marker "oh" marks the focus of a speaker's attention, that then becomes a candidate for the listener's attention. While this may be the case in the situated activity, we cannot speak for what is *commonly* or *individually* known about the activity itself, i.e. one interlocutor may use the marker, but need not inform the other interlocutor *what* was the actual update of their information state. Linking the points in time that the marker "oh" occurs with some task progression scores may give more insights on this aspect.

[11] null hypothesis

further in the task ("oh" used when (mis)matched actions occur), and other information gain, as it has a greater association with the odds of higher learning outcomes.

Discounting only the context of behavioural alignment, we would like to see if "oh" in the context of *any action* undertaken in the situated activity has a higher association with learning. For nonmatched actions on learning outcomes the results of the test show $OR = 4.94$ ($p < .001$, $95\% CI : 3.20\text{-}7.61$), showing an even greater odds of HL compared to (mis)matched actions. This means that **the interlocutors' verbalisation of "oh" around nonmatched action times have a greater association with better learning outcomes, even greater than "oh" around (mis)matched action times.** In RQ1 (SSec. 8.4.1), we discussed how nonmatched actions often result from the interlocutors working in isolation, or at least without explicit verbalised instructions inferred by our rule based behavioural alignment algorithm. What is interesting is that the simple verbalisation of "oh" of a speaker, even when interlocutors are working in isolation, at minimum may draw the attention of the listener and inform them that some change of information state regarding the activity has occurred.

In Table 8.8 and Table 8.9 we give some additional examples of "oh" in the context of nonmatched actions, for a team for high learning and low learning teams respectively. In terms of qualitative results, as shown, in Table 8.9 (a team that did not learn), there are examples of "oh" occurring when there is no context of an action (lines 13-15, lines 461-462); and "oh" occurring when there is the context of an action (i.e. both (mis)matched and nonmatched) (line 56). However, our results would indicate that "oh" occurring as a specific reaction to an action undertaken (i.e. a reaction to a stimulus) is associated with better learning outcomes. This suggests that there is a potential for the information marker "oh" to distinguish when specific kinds of information gain related to the situated activity are occurring. It would be interesting in future experiments to explore if this information could be used to decide *when a robot should intervene* in a collaborative learning activity; after for e.g. identifying unproductive periods in the task (a verbalisation of "oh" not specific to the situated activity) versus when a robot should allow the interlocutors to explore the task for themselves.

8.5 Conclusion

When working with situated activities, there is an *interdependence on the physical environment*; i.e. interlocutors need to form expressions related to the situated activity, and take actions in the activity to further the task.

Table 8.8: Excerpts that contain information management marker 'oh' from Team 17, who had high task success (i.e. performed and learnt well). Utt. stands for the utterance number, which is not applicable for the edit actions. See the caption of Table 8.9 for further details.

	Utt.	Verb	Utterance	Annotations
...	...	...	...	...
I	4	says	So you only build from something that is already connected.	-
B	5	says	Oh.	-
A	6	says	Oh okay.	-
B	-	adds	Zermatt-Davos	$\text{Nonmatch}_\text{B}(\text{Do}_\text{B}(\text{Add}))$
B	-	adds	Gallen-Davos	$\text{Nonmatch}_\text{B}(\text{Do}_\text{B}(\text{Add}))$
A	-	adds	Zurich-Davos	$\text{Nonmatch}_\text{A}(\text{Do}_\text{A}(\text{Add}))$
...	...	...	...	...
B	-	adds	Basel-Bern	$\text{Match}_\text{B}(\text{Instruct}_\text{A}(\text{Add(Basel,?)}))$
A	61	says	Yeah, and then go to Mount Zurich.	$\text{Instruct}_\text{A}(\text{Add(Zurich,?)})$
B	-	adds	Basel-Zurich	$\text{Match}_\text{B}(\text{Instruct}_\text{A}(\text{Add(Zurich,?)}))$
A	62	says	Yeah.	-
	63		Oh no that costs more.	
	64		uh ...	
B	65	says	We should erase it.	-
...	...	...	...	...
A	266	says	Then do Mount Bern to Mount Zermatt.	$\text{Instruct}_\text{A}(\text{Add(Bern,Zermatt)})$
A	267	says	Maybe that's better.	-
B	268	says	You can't do that.	-
A	269	says	Oh.	-
A	270	says	Then do ...	-
B	271	says	Mount Bern to Mount Interlaken?	$\text{Instruct}_\text{B}(\text{Add(Bern,Interlaken)})$
A	272	says	Yeah.	-
A	273	says	I think that's 4 though.	-
B	-	adds	Interlaken-Bern	$\text{Mismatch}_\text{B}(\text{Instruct}_\text{A}(\text{Add(Bern,Zermatt)}))$
A	274	says	So don't do that.	-
B	275	says	Is that 4?	-
A	276	says	Oh yeah it's, it is 4.	-
...	...	...	...	...

Table 8.9: Excerpts that contain information management marker "oh" from Team 20, who performed well but did not learn. Annotations could have the pending instructions, and (mis)matched or nonmatched actions. For brevity, nodes can be inferred from the Utterance column for (mis)matched or nonmatched actions, unless partial (mis)match.

	Utt.	Verb	Utterance	Annotations
...	...	...	...	...
B	10	says	I'm just gonna ...	-
A	-	adds	Luzern-Zermatt	$\text{Nonmatch}_\text{A}(\text{Do}_\text{A}(\text{Add}))$
A	11	says	Uh ...	-
A	12	says	Uh ...	-
B	13	says	Oh there.	-
A	14	says	Oh two three.	-
B	15	says	Oh that's what you've been doing this all time.	-
...	...	...	...	...
B	56	says	Oh I think we have to connect all of them.	$\text{Instruct}_\text{A}(\text{Add}(\text{Gallen},?))$
B	-	adds	Luzern-Interlaken	$\text{Mismatch}_\text{B}(\text{Instruct}_\text{A}(\text{Add}(\text{Gallen},?)))$
B	-	adds	Luzern-Zurich	$\text{Nonmatch}_\text{B}(\text{Do}_\text{B}(\text{Add}))$
A	57	says	Oh.	-
B	58	says	Okay I did some	-
A	59		okay for me.	
A	-	adds	Luzern-Davos	$\text{Nonmatch}_\text{A}(\text{Do}_\text{A}(\text{Add}))$
A	60	says	Oh no.	-
B	61	says	I think we are doing terrible.	-
...	...	...	...	...
A	450	says	Oh.	-
B	451	says	3.	-
B	452		What?	
A	453	says	Let me ...	-
A	-	removes	Luzern-Zermatt	$\text{Nonmatch}_\text{A}(\text{Do}_\text{A}(\text{Remove}))$
A	454	says	There you go.	-
A	455		What no.	
B	456	says	You are erasing my mistake.	-
B	457		How dare you.	
A	458	says	I know.	-
	459		Wait, what?	
	460		Can I get pencil again?	
	461		Oh oh.	
	462		Oh.	
	463		Okay.	
B	464	says	We messed up again didn't we?	-
...	...	...	...	...

Thus a crucial aspect of a situated dialogue, is not simply what was *said* by the interlocutors, but the corresponding actions taken as a result – i.e what was *done* by the interlocutors. So far, methodologies have not been developed to study how the verbal level might interact with the behavioural one. Secondly, methodologies do not consider how important the information marker "oh" could be in these contexts, particularly in a collaborative learning activity where an additional dimension of analysis is simply not performance in the task, but also, how much did interlocutors learn from the task and the interaction. "Oh" is called an information management marker because its verbalisation marks the focus of a speaker's attention to stimuli, and then becomes a candidate for the listener's attention – becoming part of the shared information state. Thus, the marker "oh" is particularly suitable to study in the context of both situated activities (where updates in one's knowledge about the task are due to to actions undertaken in the task), and in a task oriented dialogue where there may be additional pedagogical goals.

Thus, the main goal of this chapter is to study the role of "oh" in *behavioural alignment*; a new alignment context we propose to mean when instructions verbalised by one interlocutor are followed with physical actions by the other interlocutor. In this chapter, we consider the primary signal of behavioural alignment, i.e. what was instructed leading to what was done in response. To do so, we propose an action-by-action, rule based algorithm, to automatically annotate the follow-up actions taken after (automatically inferred) instructions were verbalised. Thus while the previous chapter focused solely on the surface-level utterances, i.e. "What did the interlocutors say", this chapter builds on this with actions, i.e. "What did the interlocutors do afterwards". Then, we studied the role of the information marker in relation to the follow up actions taken in the physical environment from the situated activity. Like the previous chapter, our secondary goal is to observe how these local behavioural alignment contexts can build to dialogue level task success.

Our methodological contributions are to **propose an *action-by-action, rule based algorithm* to study behavioural alignment**; i.e. to automatically infer in a timely manner when instructions *given* by an interlocutor (the instruction giver) are *followed* with actions in the physical environment by another interlocutor (the instruction follower). This algorithm builds up (cache) instructions, and depending on the action taken, clears these instructions – thus adding a temporal consideration to the processing of the dialogue. An important contribution of this chapter is firstly, **the methodologies we developed that link *the verbal modality with the physical modality*** which is particularly relevant for situated activities, and indeed, one rarely considered. While we proposed this method specific to our task (where the

role of who *gives* the instructions and who *follows* the instructions are frequently swapped), this methodology could be used in tasks where the role of instruction giver and follower are fixed.

Our experimental results strongly suggest that the verbalisation of "oh" occurs around the times of (mis)matched actions (instructions that were given by the interlocutor either leading to a corresponding or a different action) and nonmatched actions (actions occurring when there is no explicit verbalisation of an instruction) in the activity. Results indicate that the case is stronger for the occurrence of "oh" with nonmatched action times, and we found it to be the case for 70% of the teams. These results highlight the usefulness of the information marker "oh", in that indeed, in a situated activity, when interlocutors use the marker, it may be associated with actions taken in the activity and consequently, updates in the information state of what is known about the activity.

Qualitatively, we observed that better performing teams have "oh" occurring more around *matched* instruction-to-action times, compared to badly performing teams (with larger overlaps given small effect sizes for the well performing teams, and vice versa for the badly performing team). Since we observed from the previous chapter that there is no binary way to split high and low performance, we do not then see if the production of "oh" in local contexts can build to dialogue level task performance.

However, we do investigate whether "oh" used in the context of both mismatched and nonmatched actions can be associated with higher learning outcomes. Our experimental results suggest that the interlocutors' verbalisation of "oh" around (mis)matched action times have a positive association with better learning outcomes. Additionally the interlocutors' verbalisation of "oh" around nonmatched action times have a greater association with better learning outcomes, even greater than "oh" around (mis)matched action times. These results are important, as they highlight that while "oh" is generally used as a reaction to a stimulus (the general definition that (Aijmer, 1987) began with), in collaborative learning activities, there may be a *distinction* between information gain; i.e "oh" associated with updates in information state regarding the situated activity, versus "oh" used in other contexts. Indeed, the simple verbalisation of "oh" of a speaker with nonmatched actions, even when interlocutors are working in isolation (i.e. nonmatched actions compared to (mis)matched actions), at minimum may draw the attention of the listener and inform them that some change of information state regarding the activity has occurred. In collaborative learning activities, there is the potential for the information marker "oh" to distinguish when specific kinds of information gain related to the situated activity are occurring. Future work could consider if this information could be used to decide *when a robot should*

intervene in a collaborative learning activity; after for e.g. identifying unproductive periods in the task (a verbalisation of "oh" not specific to the situated activity/ not associated with actions in the task) versus when a robot should allow the interlocutors to explore the task for themselves (production of "oh" coinciding with actions in the task).

8.6 Supplementary: Algorithms

RECOGNISE-INSTRUCTIONS, described in algorithm 4, uses RECOGNISE-ENTITIES (algorithm 3, which identifies task specific referents) to automatically infer a sequence of instructions from an input utterance, via a rule-based algorithm). We then design MATCH-INSTRUCTIONS-TO-ACTIONS, described in algorithm 5. MATCH-INSTRUCTIONS-TO-ACTIONS pairs instructions with actions as matches or mismatches, for a verbal and physical actions list A (please see the next paragraph). Note that we also allow the inference of partial instructions i.e. that contain one node name only. MATCH-INSTRUCTIONS-TO-ACTIONS (algorithm 5), uses CHECK-MATCH (algorithm 6) to check if an instruction matches with an action, and accounts for partial matches of nodes, i.e. an IG need not explicitly say both node names. MATCH-INSTRUCTIONS-TO-ACTIONS builds up ("caches") pending instructions that have not been followed, clearing them once the views are swapped/every turn. The cache of instructions is also cleared if after an edit action, a match or mismatch of instructions-to-actions is detected[12]. This

[12]In our implementation, we combine the transcripts and the logs (from the dataset at DOI: 10.5281/zenodo.4627104) via a script (*tools/5_construct_the_corpus_by_combining_transcripts_with_logs.ipynb* in the tools at DOI: 10.5281/zenodo.4675070) to generate a combined corpus (*processed_data/corpus/* available with the tools). Then, we process this corpus via a script (*tools/6_recognise_instructions_detect_follow-up_actions.ipynb* in the tools) to generate an annotated corpus (*processed_data/annotated_corpus*, available with the tools). The annotated corpus contains the inferred instructions, matches, mismatches, and nonmatches if there were no instructions to be matched.).

allows for matching of negotiated nodes.

Algorithm 3: RECOGNISE-ENTITIES finds the edit entities in an utterance via a simple rule-based named entity recognition procedure. It is implemented in a script available with the tools, for which we employ named-entity recognition (NER) feature of the Python library spaCy that performs this entity recognition.

Input: A sequence of tokens $U = \langle t_1, t_2, \ldots, t_n \rangle$ that make up an utterance U

Output: A sequence of entities $E = \langle e_1, e_2, \ldots, e_m \rangle$

$N \leftarrow \langle$'Montreux', 'Bern', ..., 'Basel'$\rangle$ `// all node names`
$A \leftarrow \langle$'add', 'remove', 'build', 'connect', 'do', 'go', 'put'$\rangle$ `// add verbs`
$R \leftarrow \langle$'away', 'cut', 'delete', 'erase', 'remove', 'rub'$\rangle$ `// remove verbs`
$E \leftarrow$ an empty sequence `// for inferred entities`
foreach *token* $t \in U$ **do**
    label $\leftarrow$ *null*
    if $t \in N$ **then** *label* $\leftarrow$ 'Node'
    else if $t \in A$ **then** *label* $\leftarrow$ 'Add'
    else if $t \in R$ **then** *label* $\leftarrow$ 'Remove'
    if *label* $\neq$ *null* **then**
        $e \leftarrow$ a new entity object ; e.token $\leftarrow t$; e.label $\leftarrow$ *label*
        insert e into E
return E

Preprocessing (to obtain a verbal and physical actions list A) We combine the transcript and edit actions in a subject-verb-object(-turn-attempt) format.

- Each utterance in the transcript is added as an action with the verb 'says'.
- Each edit action from the logs is added with the verb 'adds' or 'removes', according to whether it is an add action or a remove action, respectively.

Each action $a \in A$ has fields:

- a.subject $\in$ {'A', 'B'}, the two learners that are collaborating to solve the given problem.
- a.verb $\in$ {'says', 'adds', 'removes'}, the edit actions and utterance action for matching instructions with edit actions.
- a.object $\in$ {Utterances} $\cup$ $\{(u, v) : (u, v) \in$ Edges$\}$.
- a.turn $\in \{1, 2, \ldots, n\}$ indicating the turn number of the period the action belongs to (where for utterances, the start time of the utterance

Algorithm 4: RECOGNISE-INSTRUCTIONS finds the edit instructions in an utterance. It is implemented in a script available with the tools), where the *instructions* gives the list of instructions that are inferred for an utterance.

Input: A sequence of tokens $U = \langle t_1, t_2, \ldots, t_n \rangle$ that make up an utterance U

Output: A sequence of instructions $I = \langle i_1, i_2, \ldots, i_k \rangle$

```
1  E ← RECOGNISE-ENTITIES(U)
2  I ← an empty sequence                    // for inferred instructions
3  i ← a new instruction object ; i.verb ← null ; i.u ← null; i.v ← null
     // u is the first and v is the second node by mention
4  foreach entity e ∈ E do
5      if e.label = 'Add' or e.label = 'Remove' then
6          if i.verb ≠ null then            // already inferring an instruction
7              if i.u ≠ null then           // save the partial instruction
8                  insert i into I
9              i.u ← null; i.v ← null       // clear node 1 and 2
10             i.verb ← e.label
11     else if i.u = null then              // that is, e.label = 'Node'
12         i.u ← e.token
13     else if i.v = e.token then
14         if i.u ≠ e.token then            // if not repeating node name
15             i.v ← e.token
16         if i.verb = null then            // default to a verb if not detected
17             if I.length = 0 then         // no previous instruction: default
                 to 'Add'
18                 i.verb ← 'Add'
19             else                         // default to previous instruction's verb if exists
20                 i.verb ← I[I.length − 1].verb
21             insert i into I
22             i ← a new instruction object (i.verb ← null, i.u ← null,
                 i.v ← null)
23     return I
```

belongs to). After every two edits, the turn number incremented by 1.
- a.attempt $\in \{1, 2, \ldots, m\}$ indicating the attempt number of the period the action belongs to (where for utterances, the start time of the utterance belongs to). After every submission, the attempt number is incremented by 1.

8.7 Supplementary: Accuracy of Algorithms

For studying Verbal Alignment (Chapter 7) The algorithm extracts routine expressions by the *exact matching of token sequences*: thus, the accuracy of the inference depends only on having accurate transcripts. We have gold-standard transcriptions with standardised variations of pronunciation (the details of the transcription are given Chapter 3). Thus the extraction of routine expressions, and subsequently determining the priming and establishment times are not sources of error.

However, this exact matching of token sequences is exhaustive. The original work by Dubuisson Duplessis et al. (2021) is intended to measure alignment by enumerating all existing matches (for example, even if interlocutors primed and established the token "what", this would be counted towards a routine expression formed). This is why we filter for referents that are specific to the task. Therefore, we would like to highlight that verbal alignment is *lexical*, based on a surface level matching of token sequences. Furthermore, we keep in mind the issues of transcribing disfluent speech (Le Grezause, 2017; Zayats et al., 2019a), and thus use transcription software (PRAAT) to

Algorithm 5: MATCH-INSTRUCTIONS-TO-ACTIONS pairs a list of pending instructions with actions as matches or mismatches. It is implemented in a script available with the tools.

Begin
  Input: A sequence of verbal and physical actions $A = \langle a_1, a_2, \ldots, a_k \rangle$
  Output: A sequence of $M = \langle m_1, m_2, \ldots, m_k \rangle$ holding (mis)match info m_i for each a_i
  $P \leftarrow$ an empty sequence for pending instructions to be matched
  $M \leftarrow$ an empty sequence for match/mismatch for each action in A
  attempt $\leftarrow 1$ `// submission no for clearing the pending instructions list`
  turn $\leftarrow 1$ `// turn no for clearing the pending instructions list`

Algorithm 5: MATCH-INSTRUCTIONS-TO-ACTIONS contd.

```
5
6   foreach action a ∈ A do
        ; // clear pending instructions if a new turn (or attempt
          i.e. submission)
7       if a.turn = turn + 1 then
8           clear P          // remove all items in the sequence P ;
9           turn ← a.turn    // update the current episode (i.e. new turn)
10      else if a.attempt = attempt + 1 then
11          clear P          // remove all items in the sequence P ;
12          attempt ← a.attempt    // update the current episode (i.e. new
              attempt)

        ; // for say action, recognise instructions and update the
          pending instructions list
13      if a.verb = 'says' then
14          I ← RECOGNISE-INSTRUCTIONS(a.object)          // a.object is the
              utterance ;
15          foreach instruction i ∈ I do
16              i.agent ← a.subject        // set the instructing agent ;
17              insert i into P            // update the pending instructions

        ; // for do action, try to match with a pending instruction
18      else if a.verb = 'does' then
19          I' ← {i : i ∈ I and i.agent ≠ a.subject}        // filter for the other
              interlocutor's instructions ;
20          m ← a new matching object ; m.match ← null ;
21          if I'.length > 0 then          // there is an instruction that may
              (mis)match
                ; // try to match a pending instruction with the
                  current action
22              foreach instruction i ∈ I' do
23                  if CHECK-MATCH(i, a) then
24                      m.match ← True; m.instruction ← i; m.action ← a
25              if m..match = null then       // no matches, hence a mismatch
26                  i ← I'[I'.length − 1]   // get the last instruction by the
                      other ;
27                  m.match ← False; m.instruction ← i; m.action ← a
                ; // process the match (if matched or mismatched)
28              if m..match ≠ null then // match: True or mismatch: False
29                  M[i] ← m    // add match object to list of matches ;
                    ; // remove matching instructions from pending
                      instructions sequence
30                  foreach instruction i ∈ P do
31                      if CHECK-MATCH(i, a) then remove i from P ;
32  return M
```

Algorithm 6: CHECK-MATCH checks if an instruction matches with the action. It allows partial matching for partially inferred instructions (i.e. only one of the node names is mentioned).

Input: An instruction i and an action a
Output: True if the intended action in i and action a match, False otherwise

```
1  if i.action ≠ a.verb then
2  |  return False
3  u ← a.object.u            // first node in the edited edge
4  v ← a.object.v        // second node in the edited edge, sorted
5  if i.v ≠ null then          // instruction is partially inferred i.e.
     contains i.u only
6  |  if i.u = u or i.u = v then                  // if one node matches
7  |  |  return True
8  |  else
9  |  |  return False
10 else if (i.u = u  or  i.u = v)  and  ( i.v = u  or  i.v = v) then
     // both match
11 |  return True          // note that i.v ≠ i.u by its way of inference
12 else
13 |  return False
```

ensure no unnecessary insertion, substitution or deletion of disfluencies.

For studying behavioural alignment There are two possible sources of error: while i) inferring instructions, and then while ii) matching them with actions. To infer instructions, we allow for a build up of instructions over a period of time (caching instructions), and then clear these instructions. This ensures that an instruction given at the start of the interaction is not matched with an action at the end of the interaction, i.e. there is a *temporal constraint* of when an instruction is a valid instruction. The main source of error in inferring instructions could be in *anaphora resolution*, e.g. if an interlocutor says "Connect that node *there*". This is why we allow partially inferred instructions (i.e. only one of the node names is mentioned) and (partial) matching of these instructions with actions. For example, from the utterance "Maybe we start from Mount Zermatt.", with only one node being explicitly stated, we infer the partial instruction *Add(Zermatt,?)*. Then, we consider the follow-up action as matched, if it is an add action, with 'Zermatt' as one of the nodes of the added connection. Thus, the inferring of instructions could be considered *greedy*, and suffer from inferences at each iteration without considering broader context.

With regards to matching instructions to actions, problems arise when considering the *formulaic definition of behavioural alignment*. The algorithm is explicit in considering which interlocutor is in which view (i.e. always characterising interlocutors into IF and IG). Matching instructions-to-actions are defined in a computationally strict way, and may not consider matched instructions resulting from negotiations, or consequently the build up of matches over larger periods of time (as the instructions are cleared frequently). Furthermore, given this formulaic definition, this may be a *difficult task to manually annotate* by human annotator, i.e. constantly considering who is in the position to give instructions versus follow (which changes throughout the interaction).

The annotation of high-level constructs, such as engagement, emotion and in our work, behavioural alignment, all have a perceptual component that is intrinsically hard to define, and thus guide for annotation purposes (e.g. see Nasir et al. 2021). Thus given the way behavioural alignment is defined, our measures *may not capture the whole picture*, of what an annotator might *perceptually* label the interaction to be.

8.8 Supplementary: Automatic Speech Recognition (ASR) Comparison

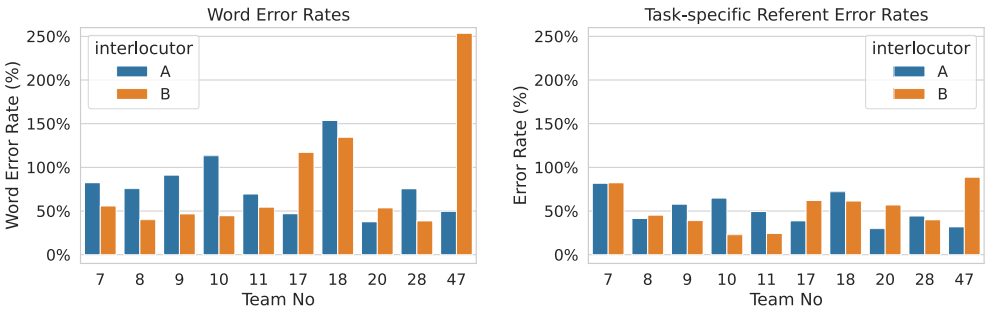
An accurate transcription of the task-specific referents is crucial for both our verbal and behavioural alignment measures (discussed in the previous Chapter 7 and this one). To evaluate how well automatic speech recognition would perform in obtaining transcripts for our analyses, we used the Google Cloud Speech-to-Text services.

To configure the recognition system, we extracted a 15 seconds long audio sample that contains task-specific referents and a filler 'um', see the reference transcript at Table 8.10. We need to assist the system towards improving its accuracy for the task-specific referents, as these are context specific words that do not occur frequently and thus would be difficult to recognise. To do so, we supplied the task-specific referents as 'hints' to the ASR system. Then, we used the extra boost feature of the system to increase the probability that these specific phrases will be recognised as described in the system's documentation (boost = 200). Furthermore, we set the system to use the enhanced "video" model that is particularly suitable "for audio that was recorded with a high-quality microphone or that has lots of background noise". This is the case for our data that contains background music of the game and some audio spill where one interlocutor's microphone could pick up sound from the other interlocutor.

For the audio sample, Table 8.10 presents our manually obtained reference transcript, the automatic transcript with the default model, and the transcript with a model adapted to our dataset. The transcription with the adapted model seems promising, as the task-specific referents are correctly transcribed. With the same adapted configuration, we automatically transcribed the complete audio files, for the subset of data. We evaluate using traditional word error rate (WER). Figure 8.2 presents the error rates for the automatic transcripts. We see that the word error rates are very high (median = 62.6%), and varied between interlocutors of the same team (ranging from = 37.7% to = 253.4%). When we filter for the task-specific referents, we see that the referent error rates are high as well (median = 47.3%). Thus it is infeasible to use automatic transcripts for this work, and therefore use gold-standard transcripts. We note that the filler 'uh' is not recognised.

type	Int.	Utterance
G	A	what about Mount Davos to Mount , Saint Gallen ?
	B	because what if , you did if we could do it ?
	A	what about Mount um Davos to Mount Gallen ?
D	A	What about Mount Davis to Mount Saint Helen ?
	B	Cuz what if you did , if we could do it .
	A	What if you did ?
	A	Could you it ? What about Mount Davis to mount gallon ?
	B	Mount .
B	A	What about Mount Davos to mount st . Gallen ?
	B	Cuz what if you did , if we could do it .
	A	What about Mount Davos to mount Gallen ?

Table 8.10: On an audio sample, where G gives the gold standard transcription, D indicates the default ASR without boosting, and B indicates the adapted/boosted model.



(a) Word Error Rates

(b) Task-specific Referent Error Rates

Figure 8.2: (left) Word and (right) task-specific referent error rates per interlocutor for the transcripts obtained by automatic speech recognition.

Conclusions

9 | Conclusions and Perspectives

9.1 Conclusions

Our overall goal of computationally studying disfluencies in SLU using psycholinguistic perspectives lead us to implement a range of strategies and methodologies. Given the range of methodological approaches, our topic has benefited the feedback and interest from many different communities.

Trends in SLU and the impact on our work

With the increasing popularity of voice assistant technologies and dialogue systems, the lines between tasks and the kind of data used are ever-shifting. For e.g., NLU research has now been expanded to overlap more frequently with SLU, to consider the *textual processing of speech transcripts* (as even discussed in Ruder (2020) as "Speech first-NLP"). Previously researchers in NLU may have been aware of disfluencies in the context of a disfluency detection task, because the task developed to predominantly use perfect text transcriptions, and "are almost *exclusively* conducted on pre-segmented utterances of the Switchboard corpus of telephone conversations" (Rohanian and Hough, 2021). Before the emergence of disfluency detection as a task, researchers were working on disfluencies with the aim that ASR systems should be more robust to them. The shift from being a mainstream concern in ASR (e.g. in Shriberg (2005)'s work titled "Spontaneous speech: How people really talk and why engineers should care"), to being more of an academic text processing type task may have occurred because task oriented dialogue became so widely and commercially used. With this came the assumption that one *specifically planned intent* is expressed every one input utterance, reducing the possible number of disfluencies that occur in sentences.

Now however, with hybrid conversations, where user interactions switch between task-oriented and open domain requests (Kim et al., 2021), this topic is starting to come to the forefront once more. However, now we have more sophisticated means and visibility regarding interdisciplinary knowledge from

both linguistic and SLU communities. We thus can study where disfluencies may be informative in SLU, rather than only considering how they should be removed from the input utterance as noise. Thus, while not a "benchmark" topic as such, this topic has benefited from interest of a wide variety of fields, such as NLU, and HRI. This is also why, we were able to methodologically make contributions in more SOTA NLP tasks (i.e. studying the representations of fillers in deep contextualised word embeddings), pedagogical HRI (studying the role of disfluencies in verbal and behavioural alignment and its impact on learning), and broader SLU tasks (prediction of the listener's impression of the speaker's metacognitive state, both using linear and neural models). There is now the general awareness that spoken speech is not like written text, and increasing studies (e.g. Kim et al. (2021)) work on how to learn representations of spontaneous speech given that architectures are usually developed on large amounts of written data.

Review of our objectives and contributions

The broad objective set out at the beginning of the thesis, was to computationally study disfluencies in SLU, motivated by psycholinguistic perspectives on the production and comprehension of speech. Specifically for us, when we spoke about *computationally studying disfluencies*, we meant to develop **methodologies** to automatically process disfluencies to **empirically observe** i) their *impact on* the production and comprehension of speech, and ii) how they *interact with* the primary signal (the lexical, or what was said in essence). Since we use psycholinguistic perspectives as inspiration, here processing is not for the purposes of removal (like the computational perspectives discussed), but rather, for the purpose of understanding their informativeness.

We began our work in Part I with a broad survey of the two perspectives of disfluencies; a psycholinguistic perspective and a computational perspective. In the psycholinguistic perspectives, we distinguished between studies that focus on the speaker's production of disfluencies, and then the impact of disfluencies on listener comprehension. We later used these distinctions throughout the thesis, of what the speaker produced, versus how this may have affected listener perception (Part II and Part III). We also spoke about the dynamic and iterative listener-speaker context, and took inspiration from this for our study of verbal and behavioural alignment between interlocutors in dialogues (Part IV).

From the computational perspectives, we first defined specific terminology that is predominantly used when studying disfluencies in SLU; particularly

for disfluency detection. We highlighted also that a lack of common terminology leads to a lack of interdisciplinary sharing on the findings. This section was in part, motivated by the overwhelming amount of research we encountered when first embarking on the thesis topic.

We then discussed the general computational perspectives of disfluencies in SLU. While many reviews have been done in linguistics (Lickley, 1994; Nicholson, 2007) and also on computational perspectives (Ginzburg et al., 2014; Purver, Hough, and Howes, 2018), the focus is more on reviewing linguistic literature, or works on automatically identifying disfluencies. We found that an existing gap in the literature was in the consideration of disfluencies from a broader SLU perspective (where there is overlap with higher dimensions of affect/personality computing), as well as a generation/TTS perspective. By making distinctions between broader SLU, and SLU as it may be understood for SDS, we point out a missing component of research, that so far we think has not been considered in existing literature reviews. We also discussed the drawbacks of these works, i.e. the tendency to not focus on the contextual use of disfluencies, there may be some that are informative, some that are uninformative, some that clarify and some that confuse. Throughout the thesis, we focus on the different contexts and functions of disfluencies; and make distinctions between general patterns of disfluencies we found in the dataset, and findings that may be more qualitative.

Lastly, we described why we chose to predominantly focus on the tokens "uh", "um" and "oh"; because of the countless psycholinguistic perspectives that emphasise their capabilities in information sharing, their high frequency occurrence in datasets, as well as even feasibility of automatic transcription from open-source speech recognisers.

In Part II and Part III We first focused on our broad objective to computationally study the production and comprehension context of fillers, and then we focused on methodologies to observe the nuances of how fillers interact with the primary signal. We studied this in a dataset of natural and voluntarily recorded monologues, where the listener was not told to pay attention to the specific use of the fillers in their annotation of high level attributes of the speaker. To summarise our methodological experiments:

- In Chapter 4, we conducted a preliminary analysis from a *production and comprehension* perspective to study the relationship between fillers, and the listener's estimation of the speaker's metacognitive state. This included i) designing a set of hand-crafted filler features to study different *production contexts*, based on psycholinguistic literature, ii) using these features to do a detailed statistical analysis and study

whether these features may affect the listener's impression differently, and iii) results across linear models to establish whether fillers (through our filler features) can help in the prediction of a listener's *perception* of the speaker's metacognitive state.

- In Chapter 5, we do an analysis from a *production* perspective in an unsupervised task (spoken language modelling (SLM)), to study the representation of fillers in deep contextualised word embeddings. We then analyse from a *comprehension* perspective the information offered from fillers without hand-crafted features on two supervised tasks; the prediction of a listener's estimation of i) a speaker's metacognitive state and ii) a speaker's stance.

- In Chapter 6, we focus more on the *interaction of fillers with the primary signal*. We proposed computational methods based on statistical analysis in order to study the interaction of disfluencies on the primary signal, from local to global. Specifically, we computationally study at a micro-level, which fillers are informative in terms of the information new and specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception.

In Part IV We focused on how fillers and the disfluency "oh" interact with the primary signal. This time, we utilised a dataset of children's dialogues, which was designed to encourage collaboration as they engaged in a collaborative learning activity. In this regard, while local utterance level information also builds to global, discourse phenomena (here, task success), there is a much more *dynamic, interactive context to consider between interlocutors* in dialogues. Here, we consider the primary signal as alignment between interlocutors, or a shared linguistic representation. To summarise our methodological experiments:

- In Chapter 7, we proposed rule-based algorithms to specifically extract verbal alignment contexts (or lexical alignment) in situated dialogues. Then, we did a i) statistical analysis to study local patterns of fillers in the context of verbal alignment between interlocutors; from the first time an interlocutor introduces an expression to the time an expression becomes part of a *shared vocabulary* (by repetition from the other interlocutor), and ii) statistical analysis to show the role of fillers used in the context of verbal alignment in relation to global variables of task success.

- In Chapter 8, novel incremental, rule based algorithms to study behavioural alignment[1]. Then, a statistical analysis to study i) the local patterns of the discourse marker "oh" in the context of behavioural alignment between interlocutors and ii) the role of "oh" used in the context of behavioural alignment, and its relation to variables of task success.

To illustrate the progress made in computationally studying disfluencies in SLU, let us consider **our contributions** from the point of view of our broad objectives of our work, that is to develop **methodologies** to automatically process disfluencies and **empirically observe** i) their *impact on* the production and comprehension of speech, and ii) how they *interact with* the primary signal (the lexical, or what was said in essence). Thus, we divide the contributions to our objectives into methodological ones, or empirical findings.

$\mathcal{O}_1$ The impact of disfluencies on the performance of SLU systems: the production and comprehension of fillers

In terms of methodological contributions, to the best of our knowledge, we are the first to develop methodologies to study the information offered uniquely by fillers in the performance of SLU systems; both using hand-crafted and unsupervised methods. We developed a set of filler features based on psycholinguistic findings of fillers (particularly, related to cognitive load). These features could be useful as a general contribution for works that utilise speech transcripts, for e.g. to aid in interpretability.

Our main methodological contribution, is the study of the representations of fillers using deep contextualised word embeddings. This is an important issue, as these models are pre-trained on massive amounts of written text, and require methodologies to further study the representations of spontaneous speech that are learnt during fine-tuning. Thus, **we are the first to *specifically* study the representations of fillers in deep contextualised word embeddings**. We develop methodologies without hand-crafted features, to show that fillers reduce the uncertainty of an LM in a downstream SLM task, showing that they do not need to blindly be removed as noise when modelling spoken language. To do so, we studied *which* token representation strategies and pre-processing strategies are best suited to model fillers on our dataset of spoken transcripts. We compared the representations learnt of fillers on two downstream tasks; the prediction of the perception of i)

[1]Which we defined as when instructions given by one interlocutor are followed with an action by another interlocutor, in a timely manner.

stance and ii) metacognitive state. We showed that the better methodology for learning representations of fillers is by first doing MLM fine-tuning on the dataset of speech transcripts.

From the results of our experiments, we empirically show that there is a relationship between fillers that were produced by the speaker, and the listener's perception of the speaker's metacognitive state. **Fillers can thus be a discriminative feature in the prediction of a listener's perception, and *thus aid SLU performance***. We established this on a large dataset that is not limited to QA contexts, using both hand-crafted features and unsupervised methods. Indeed the way fillers are used can build to an *overall* impression the listener has of the speaker, and is not only a signal that may momentarily affect the listener's impression – as was studied in previous research. Here, **we empirically found this link, which is often, simply *assumed* to be true**.

Similarly, we also found that speaker's produce more fillers when uttering opinions with a weaker stance compared to a stronger stance. We believe this finding still could have metacognitive affects; for e.g. here fillers may be used to *tone down the force of an assertion*. This then confirmed **the potential of fillers in the prediction of the perception of stance** (without hand-crafted features) at a discourse level, and not simply at a micro-level as was studied in Le Grezause (2017). This is not as straightforward as one might think, as reviews may contain *mixes of positive and negative stances*.

A common misconception may be that fillers contribute to disfluency so therefore they are not desirable in speech. Many studies also lack contextual analysis; *individual utterances build into a function of discourse*, and fillers produced may be both informative and uninformative depending on the context. An important finding is that **not all fillers may contribute equally to the listener's perception of the speaker**. Using our hand-crafted crafted features, we observed that the the positional aspects of fillers could be more important in the final rating from the listener; particularly fillers that occur sentence medially compared to fillers that occur sentence initially. Listeners thus may find it useful when fillers are used sentence initially to give prosodic structure to the review. Lastly, we offered **suggestions for where improvement is needed in the representations of fillers in deep contextualised word embeddings**, based on our experimental results. We firstly showed that in terms of production, a bulk of fillers ($\approx 40-48\%$) occur sentence initially, regardless of listener ratings (and indeed, this is consistent with many other works that find the distribution of fillers commonly at discourse boundaries (Swerts, 1998)). When we plotted

the positional distribution of fillers using our model (that had learnt representations of fillers in an unsupervised way), we found that the model can place sentence-initial fillers correctly, and then struggles with the placement of sentence-medial fillers. Furthermore, we showed that BERT was unable to distinguish between the two fillers "uh" and "um" despite our findings and others (Le Grezause, 2017) showing that they are used differently (as subsequently discussed).

$\mathcal{O}_2$ The comprehension of disfluencies: the interaction of fillers with the primary signal

Methodological contributions While in this objective, our empirical contributions are prominent, we do want to point out the general strategy of treating the message from the speaker as *an incoming source of information that builds up in the discourse*, and then *observing the function of fillers in this process*. We thus developed a methodology to computationally study at a micro-level, which fillers are informative in terms of the new information (measured by new entites) specific the discourse, and consequently, at a macro-level, what is the impact of these fillers on the listener's perception.

Empirical results of our experiments suggest that speakers generally do tend to use fillers in the incoming message, when introducing new information, rather than information already introduced into the discourse. These are important findings, as we observed this general tendency on the entire dataset of 500 speakers (while exceptions exist, as we discussed). Our results also suggest that the occurrence of *fillers specifically before new entities may not have an effect on the listener's perception of the speaker's expressed confidence*, despite previous works that suggest the link between fillers and expressed confidence. These results were on the extremes of the dataset, i.e. speakers that were rated with low confidence versus speakers that were rated highly confident. **Local hesitations need not always lead to global impressions of uncertainty.**

We also find that listener's may perceive the filler "um" as a more informative signal than the filler "uh"; and may have expectations as such. These findings are based on our previous results, that show that *all* contexts of the filler "um" have more of an effect on the listener's rating of confidence than the filler "uh", but specific to the filler "um" introducing new information, there is very small association with the confidence rating. This also suggests that they have different roles in the discourse, and is also confirmed in our analysis of dialogues.

$\mathcal{O}_3$ Alignment in the listener-speaker dynamic: the interaction of disfluencies with the primary signal

Methodological contributions We[2] first propose an adaptation on existing rule based verbal (or lexical) alignment algorithms for situated dialogues, to study the alignment of expressions specifically related to the situated activity. Then, we **propose a *novel, incremental, rule based algorithm* to study behavioural alignment**; i.e. to automatically infer in a timely manner when instructions *given* by an interlocutor (the instruction giver) are *followed* with actions in the physical environment by another interlocutor (the instruction follower). While humans process sentences *incrementally*, modern architectures often have the input of the entire utterance/dialogue, because the attention mechanisms build forward and backwards representations. While our methodology is not incremental in a word by word fashion (rather, action by action), we still build up (cache) instructions, and depending on the action taken, clear these instructions – adding a temporal consideration to the dialogue. Thus an important contribution is **the methodologies we developed that link *the verbal modality with the physical modality* (i.e. what was said with what was done)**, which is particularly relevant for situated activities, and indeed, one rarely considered. While we proposed this method specific to our task (where the role of who *gives* the instructions and who *follows* the instructions are frequently swapped), this methodology could be used in tasks where the role of instruction giver and follower are fixed.

Empirical results We studied the function of disfluencies in verbal and behavioural alignment using statistical methods, after automatically inferring these alignment contexts using the we methodologies proposed.

Overall, we see that in terms of verbal alignment, **fillers tend to occur visibly after the first time an expression is introduced into the discourse** (priming) by one interlocutor, **and around the time it is repeated by the other interlocutor**, thus becoming part of the *shared vocabulary* (establishment). Fillers in general here are used to clarify the instructions verbalised by the interlocutor.

[2]When we speak about the "communication of the primary signal", we specifically mean alignment of *what was said in essence* (verbal alignment) and *what was done in essence* (behavioural alignment). Then, we study what are the functions of disfluencies (i.e. *how it was said*) in these two contexts. Overall, the study of disfluencies in the multi-level alignment theory of Pickering and Garrod (2004) is one rarely considered in existing literature (let alone, one computationally studied).

When we specifically consider fillers constrained to a certain distance *before* a primed/established expression, we see that (while stylistically speakers may use more of one filler than the other), the filler "um" is used more than "uh" in the specific role of verbal alignment, particularly with primed expressions – i.e. when the speaker first introduces the new expression into the shared vocabulary space. This shows that the two fillers are used differently (in agreement with Le Grezause (2017)), and **the filler "um" carries more information sharing properties than the filler "uh"; as is evidenced by their function in relation to verbal alignment**. Thus an important contribution about these results, is that we found *both in monologues and dialogues* that the filler "um" is utilised more in the introduction of the new information (which in both cases we consider as new referents related specifically to the topic of discourse) than the filler "uh". Thus the results of our experiments show that for both monologues and dialogues, fillers locally can be informative signals of communication, and that globally, they also contribute to discourse level phenomena, such as the listener's impression of the speaker or variables of task success.

We then studied the information marker "oh"[3] in the context of behavioural alignment. Our experimental results strongly suggest that the verbalisation of "oh" occurs around the times of (mis)matched actions (instructions that were given by the interlocutor either leading to a corresponding or a different action) and nonmatched actions (actions occurring when there is no explicit verbalisation of an instruction) in the activity. Results indicate that the case is stronger for the occurrence of "oh" with nonmatched action times, and we found it to be the case for 70% of the teams. These results highlight the usefulness of the information marker "oh", in that indeed, in a situated activity, when interlocutors use the marker, it may be associated with actions taken in the activity and consequently, updates in the information state of what is known about the activity.

What is even more interesting, is the finding that the presence of **"oh" occurring around the same time as actions is associated more with task success, specifically learning**. We found that "oh" used in the context of both mismatched and nonmatched actions can be associated with higher learning outcomes. Additionally the interlocutors' verbalisation of "oh" around nonmatched action times have a greater association with better learning outcomes, even greater than "oh" around (mis)matched action times. These results are important, as they highlight that **while "oh" is generally used as a reaction to a stimulus (the general definition that (Ai-**

[3]According to Schiffrin (1987) the information management marker "oh" marks the focus of a speaker's attention, that then becomes a candidate for the listener's attention.

jmer, 1987) began with), in situated learning activities, there may be a *distinction* between information gain; i.e "oh" associated with updates in information state regarding the situated activity, versus "oh" used in other contexts. Indeed, the simple verbalisation of "oh" of a speaker with nonmatched actions, even when interlocutors are working in isolation (i.e. nonmatched actions compared to (mis)matched actions), at minimum may draw the attention of the listener and inform them that some change of information state regarding the activity has occurred.

Our results thus show that in collaborative learning activities, there is the **potential for the information marker "oh" to distinguish when specific kinds of information gain related to the situated activity are occurring**. Future work could consider if this information could be used to decide *when a robot should intervene* in a collaborative learning activity; after for e.g. identifying unproductive periods in the task (a verbalisation of "oh" not specific to the situated activity/ not associated with actions in the task) versus when a robot should allow the interlocutors to explore the task for themselves (production of "oh" coinciding with actions in the task).

9.2 Perspectives

Overall this thesis represents one of the many possible first steps that could be taken to computationally study disfluencies in SLU. We discuss the possible future directions and some drawbacks of our work.

According to Shannon (1948)'s Information theory, information transmission using language as a means of communication is done through a noisy channel; thus containing a mixture of information and noise. Unfortunately, disfluencies have been relegated to the category of noise, when there are numerous works (including this thesis) that show their informativeness as (un)intentionally produced signals of communication that the listener perceives. While works like Gupta et al. (2021) show the role of disfluencies in the *confusion* of systems in a QA context, we need to work more on the role of disfluencies in both the *confusion and clarification* of pragmatic intent; otherwise, we are not capturing important contextual information transmitted across the speech channel.

A motivating goal in this thesis, was the lack of contextual analysis surrounding disfluencies in existing research, leading to a general impression of how disfluencies can be linked to a wide variety of phenomena. Since disfluencies are highly contextual (with even some disfluencies like fillers *not* containing any semantic content), ideally trying to account for as much context as possible is desirable. A problem arises, in that there are many ways

in which one could integrate context; as disfluencies can depend on speaker style, accent, comfort level of language/topic, type of speech ...

Rather than a sociolinguistic perspective, we tried adopt the psycholinguistic perspectives of disfluencies used as (un)intentional cues to signal information, and also being perceived as such. With this in mind, we tried to integrate context in specific ways which we believed were applicable to SLU. A first way, was importantly in treating utterances as if they built up to a discourse, accounting for local level use of disfluencies, and effects that this might have on discourse level phenomena. This allowed us to make connections with existing work, and test for example, the *generalisability* of previous results on new and larger datasets using ML models, and on *different granularities* of analysis. In parallel, hierarchical architectures that may take into account context at a word level, then an utterance level and so on are very popular in SLU. Another integration of context, was in the consideration of how instructions verbalised in the dialogue lead to physical actions in the environment; considering a modality that is often neglected.

However, when discussing spontaneous speech phenomena, a missing dimension in our work is certainly the acoustic/ prosodic context. Disfluencies are an inherent characteristic of speech, and there have been numerous works highlighting their acoustic properties. Acoustic data could be considered in many ways, including simply to observe the feasibility of some of the methodologies proposed *without* carefully curated transcripts. This type of analysis has gained popularity recently, for e.g. testing the robustness of disfluency detection on output transcripts of ASR (Rohanian and Hough, 2021), or the same with dialogue act segmentation and classification (Tran, 2020). Indeed, this is a general problem in SLU, with many works focusing on the *gold standard transcripts* provided by human transcribers, pre-segmented utterances and so on. While we did test the accuracy of transcripts using ASR systems, we found the transcription for the disfluencies studied in the thesis lacking. The output transcripts of ASR on our dataset of children's dialogues yielded poor results. We believe that a reason for this, is that the dialogues itself contain several out of vocabulary words (gold mine names based on Swiss mountains), and the children also varied in their pronunciation of these words. Thus while we develop automatic methods to annotate alignment contexts in situated dialogues, a drawback in our work is that we still need to consider that it requires the availability of good quality text transcripts.

Furthermore, another acoustic direction could be to consider alignment at a prosodic level, particularly since disfluencies are highly correlated to prosody. It greatly surprised us that "paralinguistic alignment", i.e. the alignment of interlocutors' prosodic characteristics like pitch, loudness etc... is an

entirely separate field, that does not make connections with alignment/mutual understanding as it is studied in the NLU community. Unfortunately, our paralinguistic alignment model yielded poor results, plausibly because of the strong variations of pronunciations. Tran (2020) propose methods to automatically learn prosodic cues that is independent of speaker style, so there are methods emerging to learn generic prosodic representations. Meta AI also introduced their Generative Spoken Language Model (GSLM), that leverages representation learning to allow it to work directly from only raw audio signals, thus recognising that text can sometimes have representations of speech that are not well captured without prosodic context.

When considering context, one must also consider the ever shifting sands of communication. A descriptivist perspective is that language is constantly evolving, yet we still in SLU largely use datasets like the SWBD dataset collected thirty years ago. The idea of integration of other modalities to account for changing methods of communication, such as a visual one is exciting even in the context of disfluencies. While psycholinguistic work integrated in their methodologies hard to describe images and objects (to elicit disfluent contexts), we see this phenomena now occurring even in the automatic generation of image descriptions (Takmaz et al., 2020). In this work, the automatically generated captions *contained disfluencies (fillers)* that were elicited from humans when they were uncertain about how to describe objects. Thus disfluencies could be studied in many contexts, but we believe that it is beneficial to particularly study their relation with other modalities, as they often (as we have shown) act as signals for incoming information across modalities.

On a different note, in terms of taking stock of where we are today, aside from the distinction we discussed regarding form and meaning (from Bender and Koller (2020)), we would like to highlight Shriberg (2005)'s work titled "Spontaneous speech: How people really talk and why engineers should care". In this work, there were four fundamental challenges identified for spoken language applications; recovering hidden punctuation, coping with disfluencies, allowing for realistic turn-taking, and hearing more than words (e.g. affect). While Shriberg (2005) may have been speaking in an ASR context, what we could be added to this is, is that unlike texts, dialogues may be locally very coherent but globally not so coherent. However, this work is as relevant today as it was when it first came out. Thus while Shriberg (2005) foresaw this issue in 2005, this problem has become a reality today with hybrid systems (and indeed, with the same issues as described in the paper). Even recently, work by Kim et al. (2021) point out that dialogue modelling is based on written conversations mostly because of existing data set. From the psycholinguistic perspective, (Clark, 1997; Bailey and Ferreira,

2003) stated, that works developed on "linguistic materials that would have been produced by a writer rather than a true speaker". It seems then that there are unexpected and surprising parallels in both fields, with the problem in both being that theory is based on idealised versions of dialogues. We thus hope this thesis contributes to the study of more realistic discourse in both fields.

Bibliography

Aijmer, Karin (1987). "Oh and ah in English conversation". In: *Corpus linguistics and beyond*. Brill, pp. 61–86.

Anderson, Anne H. et al. (1991). "The Hcrc Map Task Corpus". In: *Language and Speech* 34.4, pp. 351–366. ISSN: 00238309. DOI: 10.1177/002383099103400404.

Arnold, Jennifer E, Carla L Hudson Kam, and Michael K Tanenhaus (2007). "If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33.5, p. 914. DOI: 10.1037/0278-7393.33.5.914.

Arnold, Jennifer E et al. (2004). "The old and thee, uh, new: Disfluency and reference resolution". In: *Psychological science* 15.9, pp. 578–582. DOI: 10.1111/j.0956-7976.2004.00723.x.

Bailey, Karl GD and Fernanda Ferreira (2003). "Disfluencies affect the parsing of garden-path sentences". In: *Journal of Memory and Language* 49.2, pp. 183–200.

— (2007). "The processing of filled pause disfluencies in the visual world". In: *Eye movements*. Elsevier, pp. 487–502.

Bangalore Kantharaju, Reshmashree et al. (May 2020). "Multimodal Analysis of Cohesion in Multi-party Interactions". In: *Language Resources and Evaluation*, pp. 498–507. (Visited on 03/13/2021).

Bard, Ellen G, Robin J Lickley, and Matthew P Aylett (2001). "Is disfluency just difficulty?" In: *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.

Barr, Dale J (2001). "Trouble in mind: Paralinguistic indices of effort and uncertainty in communication". In: *Oralité et gestualité: Interactions et comportements multimodaux dans la communication*, pp. 597–600.

Barr, Dale J and Mandana Seyfeddinipur (2010). "The role of fillers in listener attributions for speaker disfluency". In: *Language and Cognitive Processes* 25.4, pp. 441–455.

Barriere, Valentin, Chloé Clavel, and Slim Essid (Aug. 2017). "Opinion Dynamics Modeling for Movie Review Transcripts Classification with Hidden Conditional Random Fields". In: *Proceedings of Interspeech 2017*. Stockholm, Sweden. URL: https://hal.telecom-paris.fr/hal-02287607.

Bartunov, Sergey et al. (2015). "Breaking Sticks and Ambiguities with Adaptive Skip-Gram". In: *CoRR* abs/1502.07257. arXiv: 1502.07257. URL: http://arxiv.org/abs/1502.07257.

Bear, John, John Dowding, and Elizabeth Shriberg (1992). "Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog". In: *30th Annual Meeting of the Association for Computational Linguistics*, pp. 56–63.

Beattie, Geoffrey W and Brian L Butterworth (1979). "Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech". In: *Language and speech* 22.3, pp. 201–211.

Bell, Alan et al. (2003). "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation". In: *The Journal of the Acoustical Society of America* 113.2, pp. 1001–1024.

Bender, Emily M and Alexander Koller (2020). "Climbing towards NLU: On meaning, form, and understanding in the age of data". In: *Proc. of ACL*.

Boyd, Ryan L and H Andrew Schwartz (2021). "Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field". In: *Journal of Language and Social Psychology* 40.1, pp. 21–41.

Brennan, Susan E and Michael F Schober (2001). "How listeners compensate for disfluencies in spontaneous speech". In: *Journal of Memory and Language* 44.2, pp. 274–296.

Brennan, Susan E and Maurice Williams (1995). "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers". In: *Journal of memory and language* 34.3, pp. 383–398.

Butera, Fabrizio, Nicolas Sommet, and Céline Darnon (2019). "Sociocognitive conflict regulation: How to make sense of diverging ideas". In: *Current Directions in Psychological Science* 28.2, pp. 145–151. DOI: 10.1177/0963721418813986.

Calhoun, Sasha et al. (2010). "The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue". In: *Language resources and evaluation* 44.4, pp. 387–419.

Candea, Maria, Ioana Vasilescu, and Martine Adda-Decker (2005). "Inter-and intra-language acoustic analysis of autonomous fillers". In: *Disfluency in Spontaneous Speech*.

Cervone, Alessandra (2020). "Computational models of coherence for open-domain dialogue". In: *Computer Science*.

Chelba, Ciprian et al. (2014). "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling". In: *Proceedings of Interspeech 2014*. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.8781&rep=rep1&type=pdf.

Clark, Herbert H (1996). *Using language*. Cambridge university press.

— (1997). "Dogmas of understanding". In: *Discourse Processes* 23.3, pp. 567–598.

Clark, Herbert H. and Susan E. Brennan (1991). "Grounding in communication". In: *Perspectives on socially shared cognition*. Washington, DC, US: American Psychological Association, pp. 127–149. ISBN: 978-1-55798-121-9. DOI: 10.1037/10096-006.

Clark, Herbert H and Jean E Fox Tree (2002). "Using uh and um in spontaneous speaking". In: *Cognition* 84.1, pp. 73–111.

Clark, Herbert H. and Edward F. Schaefer (1989). "Contributing to discourse". In: *Cognitive Science* 13.2, pp. 259–294. ISSN: 03640213. DOI: 10.1016/0364-0213(89)90008-6.

Clark, Herbert H. and Deanna Wilkes-Gibbs (1986). "Referring as a collaborative process". In: *Cognition* 22.1. ISBN: 0010-0277, pp. 1–39. ISSN: 00100277. DOI: 10.1016/0010-0277(86)90010-7.

Colman, Marcus and Patrick Healey (2011). "The distribution of repair in dialogue". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 33. 33.

Corley, Martin, Lucy J MacGregor, and David I Donaldson (2007). "It's the way that you, er, say it: Hesitations in speech affect language comprehension". In: *Cognition* 105.3, pp. 658–668.

Corley, Martin and Oliver W Stewart (2008). "Hesitation disfluencies in spontaneous speech: The meaning of um". In: *Language and Linguistics Compass* 2.4, pp. 589–602.

Crible, Ludivine (2018). "Discourse Markers and (Dis) fluency: Forms and Functions Across Languages and Registers". In:

Crible, Ludivine et al. (2015). "" Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs v. 1.0". In:

Degand, Liesbeth, Bert Cornillie, and Paola Pietrandrea (2013). "Discourse Markers and Modal Particles. Categorization and description." In:

Degand, Liesbeth and Anne Catherine Simon (2009). "On identifying basic discourse units in speech: theoretical and empirical issues". In: *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 4.

Deshmukh, Neeraj et al. (1998). "Resegmentation of SWITCHBOARD". In: *Fifth international conference on spoken language processing*.

Devlin, Jacob et al. (June 2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Dillenbourg, Pierre (1999). "What do you mean by collaborative learning?" In: *Collaborative-learning: Cognitive and Computational Approaches*. Ed. by Pierre Dillenbourg. Oxford: Elsevier, pp. 1–19.

Dillenbourg, Pierre, Sanna Järvelä, and Frank Fischer (2009). "The Evolution of Research on Computer-Supported Collaborative Learning". In: *Technology-Enhanced Learning: Principles and Products*. Ed. by Nicolas Balacheff et al. Dordrecht: Springer Netherlands, pp. 3–19. ISBN: 978-1-4020-9827-7. DOI: 10.1007/978-1-4020-9827-7_1.

Dillenbourg, Pierre and David Traum (2006). "Sharing Solutions: Persistence and Grounding in Multimodal Collaborative Problem Solving". In: *Journal of the Learning Sciences* 15.1. Publisher: Routledge, pp. 121–151. DOI: 10.1207/s15327809jls1501_9.

Dillenbourg, Pierre et al. (1996). "The evolution of research on collaborative learning". In: *Technology-Enhanced Learning: Principles and Products*. ISBN: 9781402098260, pp. 3–19. ISSN: 1098-6596.

Doise, Willem and Gabriel Mugny (1984). *The social development of the intellect*. Vol. 10. International series in experimental social psychology. Pergamon Press.

Dral, Jeroen, Dirk Heylen, and Rieks op den Akker (2011). "Detecting uncertainty in spoken dialogues: an exploratory research for the automatic detection of speaker uncertainty by using prosodic markers". In: *Affective Computing and Sentiment Analysis*. Springer, pp. 67–77.

Dubuisson Duplessis, Guillaume, Chloé Clavel, and Frédéric Landragin (2017). "Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction". In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, pp. 71–81. DOI: 10.18653/v1/w17-5510.

Dubuisson Duplessis, Guillaume et al. (2021). "Towards alignment strategies in human-agent interactions based on measures of lexical repetitions". In: *Language Resources and Evaluation*. ISSN: 1574-0218. DOI: 10.1007/s10579-021-09532-w.

Dutrey, Camille et al. (2014). "A CRF-based approach to automatic disfluency detection in a French call-centre corpus". In: *Fifteenth Annual Conference of the International Speech Communication Association*.

Ekman, Paul et al. (1980). "Relative importance of face, body, and speech in judgments of personality and affect." In: *Journal of personality and social psychology* 38.2, p. 270.

Ettinger, Allyson (2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48. DOI: 10.1162/tacl_a_00298. URL: https://www.aclweb.org/anthology/2020.tacl-1.3.

Fox Tree, Jean E (1995). "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech". In: *Journal of memory and language* 34.6, pp. 709–738.

Fox Tree, Jean E. and Josef C. Schrock (1999). "Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes". In: *Journal of Memory and Language* 40.2, pp. 280–295. ISSN: 0749-596X. DOI: https://doi.org/10.1006/jmla.1998.2613. URL: https://www.sciencedirect.com/science/article/pii/S0749596X98926134.

Fraundorf, Scott H., Jennifer Arnold, and Valerie J. Langlois (2018). *Disfluency*. URL: https://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-9780199772810-0189.xml.

Fraundorf, Scott H and Duane G Watson (2011). "The disfluent discourse: Effects of filled pauses on recall". In: *Journal of memory and language* 65.2, pp. 161–175.

Fusaroli, Riccardo and Kristian Tylén (2016). "Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance". In: *Cognitive Science* 40.1, pp. 145–171. ISSN: 1551-6709. DOI: 10.1111/cogs.12251. (Visited on 06/06/2020).

Garcia, Alexandre et al. (2019). "A multimodal movie review corpus for fine-grained opinion mining". In: *CoRR* abs/1902.10102. arXiv: 1902.10102. URL: http://arxiv.org/abs/1902.10102.

Garrod, Simon and Anthony Anderson (1987). "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination". In: *Cognition* 27, pp. 181–218. DOI: 10.1016/0010-0277(87)90018-7.

Ginzburg, Jonathan et al. (2014). "Disfluencies as intra-utterance dialogue moves". In: *Semantics and Pragmatics*.

Godfrey, John and Edward Holliman (1993). "Switchboard-1 Release 2 LDC97S62". In: *Linguistic Data Consortium*.

Godfrey, John J, Edward C Holliman, and Jane McDaniel (1992). "SWITCHBOARD: Telephone Speech Corpus for Research and Development". In:

Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vol. 1. IEEE, pp. 517–520. URL: https://ieeexplore.ieee.org/document/225858.

Goryainova, Maria et al. (2014). "Morpho-Syntactic Study of Errors from Speech Recognition System". In: *International Conference on Language Resources and Evaluation*.

Grice, H Paul, Peter Cole, Jerry Morgan, et al. (1975). "Logic and conversation". In: *1975*, pp. 41–58.

Grosman, Iulia (2018). "Évaluation contextuelle de la (dis) fluence en production et perception: pratiques communicatives et formes prosodico-syntaxiques en français". PhD thesis. UCL-Université Catholique de Louvain.

Gupta, Aditya et al. (2021). "Disfl-QA: A Benchmark Dataset for Understanding Disfluencies in Question Answering". In: *CoRR* abs/2106.04016. arXiv: 2106.04016. URL: https://arxiv.org/abs/2106.04016.

Haddington, Pentti (2004). "Stance taking in news interviews". In: *SKY Journal of Linguistics* 17, pp. 101–142.

Heeman, Peter A and James Allen (1999). "Speech repains, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue". In: *Computational Linguistics* 25.4, pp. 527–572.

Hemphill, James F (2003). "Interpreting the magnitudes of correlation coefficients." In:

Hough, Julian, David Schlangen, et al. (2015). "Recurrent neural networks for incremental disfluency detection". In:

Hough, Julian et al. (2016). "Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1784–1788.

Jermann, Patrick and Marc-Antoine Nüssli (2012). "Effects of Sharing Text Selections on Gaze Cross-recurrence and Interaction Quality in a Pair Programming Task". In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. event-place: Seattle, Washington, USA. New York, NY, USA: ACM, pp. 1125–1134. ISBN: 978-1-4503-1086-4. DOI: 10.1145/2145204.2145371. (Visited on 09/19/2019).

Jermann, Patrick et al. (2011). "Collaborative Gaze Footprints: Correlates of Interaction Quality". In: *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL2011 Conference Proceedings.* Volume I - Long Papers, pp. 184–191.

Jiang, Xiaoming and Marc D Pell (2017). "The sound of confidence and doubt". In: *Speech Communication* 88, pp. 106–126.

Kapur, Manu (2008). "Productive Failure". In: *Cognition and Instruction* 26.3. Publisher: Routledge, pp. 379–424. ISSN: 0737-0008. DOI: 10.1080/07370000802212669. (Visited on 06/06/2020).

Kapur, Manu and Katerine Bielaczyc (2012). "Designing for Productive Failure". In: *Journal of the Learning Sciences* 21.1. Publisher: Routledge, pp. 45–83. DOI: 10.1080/10508406.2011.591717.

Kim, Seokhwan et al. (2021). "" How Robust ru?": Evaluating Task-Oriented Dialogue Systems on Spoken Conversations". In: *arXiv e-prints*, arXiv–2109.

Koiso, Hanae et al. (1998). "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs". In: *Language and speech* 41.3-4, pp. 295–321. DOI: 10.1177/002383099804100404.

Kruijff, Geert-Jan M. et al. (2010). "Situated Dialogue Processing for Human-Robot Interaction". In: *Cognitive Systems*. Ed. by Henrik Iskov Christensen, Geert-Jan M. Kruijff, and Jeremy L. Wyatt. Cognitive Systems Monographs. Berlin, Heidelberg: Springer, pp. 311–364. ISBN: 978-3-642-11694-0. DOI: 10.1007/978-3-642-11694-0_8. (Visited on 06/06/2020).

Kuhn, Deanna (2015). "Thinking Together and Alone". In: *Educational Researcher* 44, pp. 46–53. DOI: 10.3102/0013189X15569530.

Le Grezause, Esther (2017). "Um and Uh, and the expression of stance in conversational speech". PhD thesis.

Levelt, Willem JM (1983). "Monitoring and self-repair in speech". In: *Cognition* 14.1, pp. 41–104.

Leviathan, Yaniv and Yossi Matias (2018). "Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone." In: *Google AI Blog*.

Levow, Gina-Anne et al. (2014). "Recognition of stance strength and polarity in spontaneous speech". In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 236–241.

Lickley, Robin J (1994). "Detecting disfluency in spontaneous speech". PhD thesis. University of Edinburgh.

— (2015). "Fluency and Disfluency". In: *The handbook of speech production*, p. 445.

Lickley, Robin J and Ellen G Bard (1998). "When can listeners detect disfluency in spontaneous speech?" In: *Language and speech* 41.2, pp. 203–226.

Lickley, Robin J, Richard C Shillcock, and Ellen Gurman Bard (1991). "Processing Disfluent Speech: How and when are disfluencies found?" In: *Second European Conference on Speech Communication and Technology*.

Louvan, Samuel and Bernardo Magnini (2020). "Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey". In: *arXiv preprint arXiv:2011.00564*.

Maclay, Howard and Charles E Osgood (1959). "Hesitation phenomena in spontaneous English speech". In: *Word* 15.1, pp. 19–44.

Madaio, Michael, Justine Cassell, and Amy Ogan (2017a). "The impact of peer tutors' use of indirect feedback and instructions". In: Philadelphia, PA: International Society of the Learning Sciences.

— (2017b). ""I think you just got mixed up": confident peer tutors hedge to support partners' face needs". In: *International Journal of Computer-Supported Collaborative Learning* 12.4, pp. 401–421.

Madaio, Michael et al. (2017). "Using Temporal Association Rule Mining to Predict Dyadic Rapport in Peer Tutoring." In: *International Educational Data Mining Society*.

Mahl, George F (1956). "Disturbances and silences in the patient's speech in psychotherapy." In: *The Journal of Abnormal and Social Psychology* 53.1, p. 1.

Mairesse, François et al. (2007). "Using linguistic cues for the automatic recognition of personality in conversation and text". In: *Journal of artificial intelligence research* 30, pp. 457–500.

Mason, Winter and Siddharth Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk". In: *Behavior research methods* 44.1, pp. 1–23.

Mehta, Yash et al. (2020). "Recent trends in deep learning based personality detection". In: *Artificial Intelligence Review* 53.4, pp. 2313–2339.

Menon, Divya et al. (2019). "Going beyond digital literacy to develop computational thinking in K-12 education". In: *L. Daniela, Smart Pedagogy of Digital Learning, Taylor & Francis (Routledge), In press, 9780367333799*.

Meteer, Marie W et al. (1995). *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania Philadelphia, PA.

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781*. URL: `https://arxiv.org/abs/1301.3781`.

Mills, Gregory J (2014). "Establishing a communication system: Miscommunication drives abstraction". In: *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)*. World Scientific, pp. 193–194.

Mills, Gregory J and Patrick GT Healey (2006). *Clarifying spatial descriptions: Local and global effects on semantic co-ordination*. Universität Potsdam.

Moore, Johanna D, Leimin Tian, and Catherine Lai (2014). "Word-level emotion recognition using high-level features". In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 17–31.

Mugny, Gabriel and Willem Doise (1978). "Socio-cognitive conflict and structure of individual and collective performances". In: *European journal of social psychology* 8.2. Publisher: Wiley Online Library, pp. 181–192. DOI: 10.1002/ejsp.2420080204.

Nasir, Jauwairia et al. (2020a). "When Positive Perception of the Robot Has No Effect on Learning". In: *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 313–320. DOI: 10.1109/RO-MAN47096.2020.9223343.

— (2020b). "You Tell, I Do, and We Swap until we Connect All the Gold Mines!" In: *ERCIM News* 2020.120. URL: https://ercim-news.ercim.eu/en120/special/you-tell-i-do-and-we-swap-until-we-connect-all-the-gold-mines.

Nasir, Jauwairia et al. (Mar. 18, 2021). "What if Social Robots Look for Productive Engagement?" In: *International Journal of Social Robotics*. ISSN: 1875-4805. DOI: 10.1007/s12369-021-00766-w.

Nicholson, Hannele Buffy Marie (2007). "Disfluency in dialogue: attention, structure and function". In:

Norman, Utku et al. (2021). *JUSThink Alignment Analysis*. URL: https://zenodo.org/record/4675070.

Ochshorn, R. M. and M. Hawkins. "Gentle forced aligner [computer program]". In: ().

O'Shaughnessy, Douglas (1995). "Timing patterns in fluent and disfluent spontaneous speech". In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE, pp. 600–603.

Oviatt, Sharon (1995). "Predicting spoken disfluencies during human-computer interaction". In: *Computer Speech and Language* 9.1, pp. 19–36.

Park, Sunghyun et al. (2014). "Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach". In: *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI 2014. Istanbul, Turkey: Association for Computing Machinery. ISBN: 9781450328852. DOI: 10.1145/2663204.2663260. URL: https://doi.org/10.1145/2663204.2663260.

Pennebaker, James W, Martha E Francis, and Roger J Booth (2001). "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001, p. 2001.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

Pfeifer, Laura M and Timothy Bickmore (2009). "Should agents speak like, um, humans? The use of conversational fillers by virtual agents". In: *International Workshop on Intelligent Virtual Agents*. Springer, pp. 460–466.

Pickering, Martin J. and Simon Garrod (2004). "Toward a mechanistic psychology of dialogue". In: *Behavioral and Brain Sciences* 27.02, pp. 169–225. ISSN: 0140-525X. DOI: 10.1017/s0140525x04000056.

— (2006). "Alignment as the basis for successful communication". In: *Research on Language and Computation* 4.2-3. ISBN: 1116800690040, pp. 203–228. ISSN: 15707075. DOI: 10.1007/s11168-006-9004-0.

Pickett, Joseph P (2018). *The American heritage dictionary of the English language*. Houghton Mifflin Harcourt.

Plauché, Madelaine and Elizabeth Shriberg (1999). "Data-driven subclassification of disfluent repetitions based on prosodic features". In: *Proc. International Congress of Phonetic Sciences*. Vol. 2. Citeseer, pp. 1513–1516.

Pon-Barry, Heather (2008). "Prosodic manifestations of confidence and uncertainty in spoken language". In: *Ninth Annual Conference of the International Speech Communication Association*.

Purver, Matthew, Julian Hough, and Christine Howes (2018). "Computational Models of Miscommunication Phenomena". en. In: *Topics in Cognitive Science* 10.2, pp. 425–451. ISSN: 1756-8765. DOI: 10.1111/tops.12324. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12324 (visited on 01/24/2020).

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In: *OpenAI Blog* 1.8. URL: https://api.semanticscholar.org/CorpusID:160025533.

Rasipuram, Sowmya and Dinesh Babu Jayagopi (2016). "Automatic assessment of communication skill in interface-based employment interviews using audio-visual cues". In: *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, pp. 1–6.

Rohanian, Morteza and Julian Hough (2021). "Best of Both Worlds: Making High Accuracy Non-incremental Transformer-based Disfluency Detection Incremental". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3693–3703.

Rohanian, Morteza, Julian Hough, and Matthew Purver (2021). "Alzheimer's Dementia Recognition Using Acoustic, Lexical, Disfluency and Speech Pause Features Robust to Noisy Inputs". In: *arXiv preprint arXiv:2106.15684*.

Romano, Jeanine et al. (2006). "Appropriate Statistics for Ordinal Level Data: Should We Really Be Using t-test and Cohen's d for Evaluating Group Differences on the NSSE and other Surveys?" In: *annual meeting of the Florida Association of Institutional Research*. Vol. 177.

Roschelle, Jeremy and Stephanie D. Teasley (1995). "The Construction of Shared Knowledge in Collaborative Problem Solving". In: *Computer Supported Collaborative Learning* 128. Ed. by C. O'Malley. Publisher: Springer Place: Berlin, Heidelberg ISBN: 3540577408, pp. 69–97. ISSN: 01635735. DOI: 10.1007/978-3-642-85098-1_5.

Ruder, Sebastian (2020). *ICLR 2021 Outstanding Papers, Char Wars, Speech-first NLP, Virtual conference ideas*. URL: https://newsletter.ruder.io/issues/iclr-2021-outstanding-papers-char-wars-speech-first-nlp-virtual-conference-ideas-483703.

Saini, Divya (2017). "The Effect of Speech Disfluencies on Turn-Taking". In:

Sangin, Mirweis et al. (May 1, 2011). "Facilitating peer knowledge modeling: Effects of a knowledge awareness tool on collaborative learning outcomes and processes". In: *Computers in Human Behavior*. Group Awareness in CSCL Environments 27.3, pp. 1059–1067. ISSN: 0747-5632. DOI: 10.1016/j.chb.2010.05.032. (Visited on 04/05/2020).

Schiffrin, Deborah (1987). *Discourse markers*. 5. Cambridge University Press.

Schlangen, David and Gabriel Skantze (2011). "A general, abstract model of incremental dialogue processing". In: *Dialogue & Discourse* 2.1, pp. 83–111.

Schrank, Tobias and Barbara Schuppler (2015). "Automatic detection of uncertainty in spontaneous German dialogue". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Schuller, Björn et al. (2019). "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge". In: *Computer Speech & Language* 53, pp. 156–180.

Schwartz, Daniel L (1995). "The emergence of abstract representations in dyad problem solving". In: *Journal of the Learning Sciences (JLS)* 4.3. Publisher: Taylor & Francis, pp. 321–354. DOI: 10.1207/s15327809jls0403_3.

Serban, Iulian et al. (2016). "Building end-to-end dialogue systems using generative hierarchical neural network models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1.

Shalyminov, Igor, Arash Eshghi, and Oliver Lemon (2018). "Multi-Task Learning for Domain-General Spoken Disfluency Detection in Dialogue Systems". In: *CoRR* abs/1810.03352. arXiv: 1810.03352. URL: http://arxiv.org/abs/1810.03352.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423.

Shriberg, Elizabeth (1995). "Acoustic properties of disfluent repetitions". In: *Proceedings of the international congress of phonetic sciences*. Vol. 4, pp. 384–387.

— (2001). "To 'errrr'is human: ecology and acoustics of speech disfluencies". In: *Journal of the International Phonetic Association* 31.1, pp. 153–169.

— (2005). "Spontaneous speech: How people really talk and why engineers should care". In: *Ninth European Conference on Speech Communication and Technology*.

Shriberg, Elizabeth, Rebecca A Bates, and Andreas Stolcke (1997). "A prosody only decision-tree model for disfluency detection." In: *Eurospeech*. Vol. 97. Citeseer, p. 23832386.

Shriberg, Elizabeth et al. (1998). "Can prosody aid the automatic classification of dialog acts in conversational speech?" In: *Language and speech* 41.3-4, pp. 443–492.

Shriberg, Elizabeth E (1999). *Phonetic consequences of speech disfluency*. Tech. rep. SRI INTERNATIONAL MENLO PARK CA.

Shriberg, Elizabeth E and Robin J Lickley (1993). "Intonation of clause-internal filled pauses". In: *Phonetica* 50.3, pp. 172–179.

Shriberg, Elizabeth Ellen (1994). "Preliminaries to a theory of speech disfluencies". PhD thesis. Citeseer.

Skantze, Gabriel, Martin Johansson, and Jonas Beskow (2015). "Exploring turn-taking cues in multi-party human-robot discussions about objects". In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 67–74.

Smith, Vicki L and Herbert H Clark (1993). "On the course of answering questions". In: *Journal of memory and language* 32.1, pp. 25–38.

Stahl, Gerry (2007). "Meaning making in CSCL: Conditions and preconditions for cognitive processes by groups". In: *Proceedings of the 8th international conference on Computer supported collaborative learning*. ISBN: 978-0-6151-5436-7, pp. 652–661. ISSN: 15734552. DOI: 10.3115/1599600.1599723.

Stahl, Gerry, Timothy Koschmann, and Dan Suthers (2006). "Computer-supported collaborative learning: An historical perspective". In: *Cambridge handbook of the learning sciences*. Ed. by R. K. Sawyer. University Press, pp. 409–426.

Stolcke, Andreas and Elizabeth Shriberg (1996). "Statistical Language Modeling for Speech Disfluencies". In: *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Con-*

ference Proceedings (ICASSP). Vol. 1. IEEE, pp. 405–408. URL: https://ieeexplore.ieee.org/abstract/document/541118.

Swerts, Marc (1998). "Filled Pauses as Markers of Discourse Structure". In: *Journal of Pragmatics* 30.4, pp. 485 –496. ISSN: 0378-2166. DOI: https://doi.org/10.1016/S0378-2166(98)00014-9. URL: http://www.sciencedirect.com/science/article/pii/S0378216698000149.

Swerts, Marc and Ronald Geluykens (1994). "Prosody as a Marker of Information Flow in Spoken Discourse". In: *Language and Speech* 37.1, pp. 21–43. URL: https://journals.sagepub.com/doi/abs/10.1177/002383099403700102.

Swerts, Marc and Emiel Krahmer (2005). "Audiovisual prosody and feeling of knowing". In: *Journal of Memory and Language* 53.1, pp. 81–94.

Székely, Éva et al. (2019). "How to train your fillers: uh and um in spontaneous speech synthesis". In: *The 10th ISCA Speech Synthesis Workshop*.

Takmaz, Ece et al. (Nov. 2020). "Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4664–4677. DOI: 10.18653/v1/2020.emnlp-main.377. URL: https://aclanthology.org/2020.emnlp-main.377.

Thompson, Henry S et al. (1993). "The HCRC map task corpus: Natural dialogue for speech recognition". In: *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Tian, Leimin, Johanna D Moore, and Catherine Lai (2015). "Emotion recognition in spontaneous and acted dialogues". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 698–704.

Tottie, Gunnel (2014). "On the use of uh and um in American English". In: *Functions of Language* 21.1, pp. 6–29.

Tran, Trang (2020). "Neural Models for Integrating Prosody in Spoken Language Understanding". PhD thesis. University of Washington.

Tran, Trang et al. (2017a). "Joint modeling of text and acoustic-prosodic cues for neural parsing". In: *arXiv preprint arXiv:1704.07287*.

— (2017b). "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information". In: *arXiv preprint arXiv:1704.07287*.

Tran, Trang et al. (2019a). "On the Role of Style in Parsing Speech with Neural Models". In: *Proc. Interspeech 2019*, pp. 4190–4194.

— (2019b). "On the Role of Style in Parsing Speech with Neural Models". In: *Proceedings of Interspeech 2019*, pp. 4190–4194. DOI: 10.21437/

Interspeech . 2019 - 3122. URL: http : / / dx . doi . org / 10 . 21437 / Interspeech.2019-3122.

Tur, Gokhan and Renato De Mori (2011). *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.

Vasilescu, Ioana, Sophie Rosset, and Martine Adda-Decker (2010a). "On the functions of the vocalic hesitation euh in interactive man-machine question answering dialogs in French". In: *DiSS-LPSS Joint Workshop 2010*.

— (2010b). "On the Role of Discourse Markers in Interactive Spoken Question Answering Systems." In: *LREC*.

Vinciarelli, Alessandro and Gelareh Mohammadi (2014). "A survey of personality computing". In: *IEEE Transactions on Affective Computing* 5.3, pp. 273–291.

Visschers-Pleijers, Astrid J. S. F. et al. (2006). "Analysis of verbal interactions in tutorial groups: a process study". In: *Medical Education* 40.2, pp. 129–137. ISSN: 1365-2923. DOI: 10.1111/j.1365-2929.2005.02368.x.

Wollermann, Charlotte et al. (2013). "Disfluencies and uncertainty perception–evidence from a human–machine scenario". In: *Sixth Workshop on Disfluency in Spontaneous Speech*.

Yew, Elaine H. J. and Henk G. Schmidt (May 2009). "Evidence for constructive, self-regulatory, and collaborative processes in problem-based learning". In: *Advances in Health Sciences Education: Theory and Practice* 14.2, pp. 251–273. ISSN: 1573-1677. DOI: 10.1007/s10459-008-9105-7.

— (Mar. 2012). "What students learn in problem-based learning: a process analysis". In: *Instructional Science* 40.2, pp. 371–395. ISSN: 1573-1952. DOI: 10.1007/s11251-011-9181-6.

Yoshida, Etsuko and Robin J Lickley (2010a). "Disfluency patterns in dialogue processing". In: *DiSS-LPSS Joint Workshop 2010*.

— (2010b). "Disfluency patterns in dialogue processing". In: *DiSS-LPSS Joint Workshop 2010*.

Yuan, Jiahong and Mark Liberman (2008). "Speaker identification on the SCOTUS corpus". In: *Journal of the Acoustical Society of America* 123.5, p. 3878.

Zadeh, Amir. *CMU-Multimodal SDK*. https://github.com/A2Zadeh/CMU-MultimodalSDK/blob/master/mmsdk/mmdatasdk/dataset/standard_datasets/POM/pom_std_folds.py.

Zayats, Vicky et al. (2019a). "Disfluencies and Human Speech Transcription Errors". In: *arXiv preprint arXiv:1904.04398*.

— (2019b). "Disfluencies and Human Speech Transcription Errors". In: *Proc. Interspeech 2019*. DOI: 10 . 21437 / Interspeech . 2019. URL: https : // www.isca-speech.org/archive/Interspeech_2019/pdfs/3134.pdf.

Zhao, Tianyu and Tatsuya Kawahara (2019a). "Joint dialog act segmentation and recognition in human conversations using attention to dialog context". In: *Computer Speech 1& Language* 57, pp. 108–127.

— (2019b). "Joint dialog act segmentation and recognition in human conversations using attention to dialog context". In: *Computer Speech & Language* 57, pp. 108 –127. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2019.03.001. URL: http://www.sciencedirect.com/science/article/pii/S0885230818304030.

Zhu, Yukun et al. (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27. URL: https://ieeexplore.ieee.org/document/7410368.

DOCTORAL SCHOOL

Titre : Modèles Computationnels des Disfluences dans les Traitement de la Parole

Mots clés : Disfluences, Langue Parleé, Traitement du Langage Naturel.

Résumé :
Les locuteurs s'expriment rarement de la même manière qu'ils écrivent - en effet ils écrivent rarement de manière diffluente. Les disfluences sont des interruptions dans le flux régulier de la parole, telles que les pauses (silencieuses), les répétitions de mots ou les interruptions pour corriger une phrase précédemment dite. Bien qu'il s'agisse d'une caractéristique naturelle de la parole spontanée et malgré la riche littérature linguistique qui traite de leur caractère informatif, elles sont souvent considérées comme du bruit et éliminées lors du post-traitement des transcriptions de sortie des systèmes de reconnaissance de la parole. Jusqu'à présent, leur prise en compte dans un contexte de compréhension de la langue parlée (CLP) a rarement été explorée. L'objectif de cette thèse est de développer des modèles informatiques des disfluences dans la CLP. Pour ce faire, nous prenons inspiration dans les modèles psycholinguistiques des disfluences, qui se concentrent sur le rôle que les disfluences jouent dans l'expression (par le locuteur) et la compréhension (par l'auditeur) du discours. Plus précisément, lorsque nous utilisons le terme "modèles informatiques des disfluences", nous entendons développer des méthodologies qui traitent automatiquement les disfluences afin d'observer empiriquement 1) leurs impacts sur la production et la compréhension de la parole et 2) leurs interactions avec le signal primaire (lexical, ou la substance du discours). À cet effet, nous nous concentrons sur deux types de discours : les monologues et les dialogues orientés vers une tâche.
Nos résultats se concentrent sur des tâches de CLP, ainsi que sur les recherches pertinentes pour les systèmes de dialogues parlés. Lors de l'étude des monologues, nous utilisons une combinaison de modèles traditionnels et neuronaux pour étudier les représentations et l'impact des disfluences sur la performance du CLP. De plus, nous développons des méthodologies pour étudier les disfluences en tant qu'indices d'informations entrantes dans le flux du discours. Dans l'étude des dialogues orientés vers une tâche, nous nous concentrons sur le développement de modèles informatiques pour étudier les rôles des disfluences dans la dynamique auditeur-locuteur. Nous étudions spécifiquement les disfluences dans le contexte de l'alignement verbal, c'est-à-dire l'alignement des expressions lexicales des interlocuteurs et leurs rôles dans l'alignement comportemental, un nouveau contexte d'alignement que nous proposons de définir comme le moment où les instructions données par un interlocuteur sont suivis d'une action par un autre interlocuteur. Nous examinons également comment les disfluences dans les contextes d'alignement locaux peuvent être associées à des phénomènes au niveau du discours, tels que la réussite de la tâche. Nous considérons cette thèse comme l'un des premiers travaux, qui pourrait aboutir à l'intégration des disfluences dans les contextes d'alignement local.

Title : Computational Models of Disfluencies
Fillers and Discourse Markers in Spoken Language Understanding

Keywords : Disfluencies, Spoken Language Understanding (SLU), Fillers, Discourse Markers

Abstract : People rarely speak in the same manner that they write – they are generally disfluent. Disfluencies can be defined as interruptions in the regular flow of speech, such as pausing silently, repeating words, or interrupting oneself to correct something said previously. Despite being a natural characteristic of spontaneous speech, and the rich linguistic literature that discusses their informativeness, they are often removed as noise in post-processing from the output transcripts of speech recognisers. So far, their consideration in a Spoken Language Understanding (SLU) context has been rarely explored. The aim of this thesis is to develop computational models of disfluencies in SLU, focusing on tasks related to social interaction. To do so, we take inspiration from psycholinguistic models of disfluencies, which focus on the role that disfluencies play in the production (by the speaker) and perception (of the listener) of speech. Specifically, when we use the term ”computational models of disfluencies”, we mean to develop methodologies that automatically process disfluencies to empirically observe 1) their impact on the production and perception of speech, and 2) how they interact with the primary signal (the lexical, or what was said in essence). To do so, we focus on two discourse contexts ; monologues and task-oriented dialogues.
Our results mainly contribute to tasks in SLU related to social interaction, but also research that could be relevant to Spoken Dialogue Systems. When studying monologues, we use a combination of traditional and neural models to study the representations and impact of disfluencies on SLU performance. Additionally, we develop methodologies to study disfluencies as a cue for incoming information in the flow of the discourse. In studying task-oriented dialogues, we focus on developing computational models to study the roles of disfluencies in the listener-speaker dynamic. We specifically study disfluencies in the context of verbal alignment ; i.e. the alignment of the interlocutors' lexical expressions, and the role of disfluencies in behavioural alignment ; a new alignment context that we propose to mean when instructions given by one interlocutor are followed with an action by another interlocutor. We also consider how these disfluencies in local alignment contexts can be associated with discourse level phenomena ; such as success in the task. We consider this thesis one of the many first steps that could be undertaken to further research disfluencies in SLU contexts.

Institut Polytechnique de Paris
91120 Palaiseau, France