UTRECHT UNIVERSITY

MASTER THESIS

Dialogue Act Recognition for Conversational Agents

Author:
Lynn Hacquebord
ICA-3761045

Supervisor:
Dr. F.P.M. Dignum

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Game and Media Technology
Department of Information and Computing Sciences

March, 2017

UTRECHT UNIVERSITY

Abstract

Faculty of Science
Department of Information and Computing Sciences

Master of Science

Dialogue Act Recognition for Conversational Agents

by Lynn Hacquebord

The recognition of dialogue acts, actions that are performed through speech, is important for conversational agents to function well. Unfortunately, the studies that have been done on this topic vary a lot in methodology, so they are difficult to compare. Additionally, their results are often presented rather shallowly, leaving out a lot of valuable information. This thesis investigates the various methods that have been used for this task to determine what does and does not work, what the problem areas are and how they may be resolved. In addition, an alternative approach to dialogue act recognition is proposed that addresses some of the issues encountered. All of this is done within the context of conversational agents which is limiting in some ways, but also provides opportunities for improving the process.

Contents

Abstract iii

1 Introduction 1

2 Background Information 3
2.1 Linguistic Background . . . 3
2.1.1 Morphology, Syntax and Morphosyntax . . . 3
2.1.2 Semantics . . . 4
2.1.3 Pragmatics . . . 5
2.2 Approaches: Rules vs. Statistical Methods . . . 7
2.3 Dialogue System Architecture . . . 8
2.3.1 Dialogue Manager . . . 8
2.3.2 Natural Language Understanding . . . 10

3 Related Work 13
3.1 Dialogue Act Classes . . . 14
3.2 Corpora . . . 16
3.2.1 Switchboard . . . 16
3.2.2 ICSI Meeting Corpus . . . 17
3.2.3 HCRC Map Task Corpus . . . 17
3.2.4 NPS Internet Chatroom Conversations . . . 17
3.3 Features . . . 18
3.4 Classification Algorithms . . . 18

4 Analysis 21
4.1 Methodology . . . 21
4.1.1 Corpus . . . 21
4.1.2 Features . . . 23
4.1.3 Evaluation . . . 26
4.1.4 Implementation Details . . . 27
4.2 Results . . . 27
4.2.1 First n Words . . . 27
4.2.2 First n Part-of-speech Tags . . . 28
4.2.3 Previous n Speakers and Dialogue Acts . . . 28
4.2.4 Individual Features . . . 29
4.2.5 Combined Features . . . 29
4.2.6 Dialogue Acts . . . 30
4.2.7 Top 5 Results . . . 31
4.3 Discussion . . . 31
4.3.1 The Best Value for n . . . 34
4.3.2 Question Mark & Length . . . 34
4.3.3 N-grams . . . 34
4.3.4 Issues . . . 35

5 Alternative Approach **45**
5.1 Proposed Approach . . . 46
5.2 Methodology . . . 47
5.2.1 Clustering . . . 47
5.2.2 Termination . . . 48
5.2.3 Assigning dialogue acts . . . 48
5.2.4 Algorithm . . . 49
5.2.5 Implementation Details . . . 49
5.3 Results . . . 50
5.4 Limitations . . . 51
5.5 Discussion . . . 53

6 Future Work **55**

7 Conclusion **57**

A Corpus and Dialogue Acts **59**

B Analysis Results **63**
B.1 Average Recall . . . 64
B.2 Confusion Matrix . . . 69

C Classification Tree **73**

Bibliography **85**

List of Figures

4.1 The average recall when only the first n words are used as features. 28
4.2 The average recall when only the first n part-of-speech tags are used as features. 28
4.3 The average recall when only the previous n speakers and dialogue acts are used as features. 29
4.4 The average recall for each feature individually. POS = part-of-speech, q. mark = question mark, prev. sp./DA = previous speaker and dialogue act. 30
4.5 The average recall for combinations of features. The first nine words were used as the base and all other features were added on one by one. POS = part-of-speech, q. mark = question mark, prev. sp./DA = previous speaker and dialogue act. 30

5.1 A small brach of the tree that focuses on acknowledgement-like utterances. 50

List of Tables

3.1 The 42 clustered SWBD-DAMSL dialogue act tags. 15
3.2 An overview of four different corpora used to train dialogue act classifiers. 16
3.3 An overview of the different classification algorithms used for dialogue act recognition. 19

4.1 The types of markings occurring in the transcripts and how the texts are modified to filter them out. 22
4.2 Examples of utterances before and after cleaning. 23
4.3 The size of each of the ten subsets the corpus was split into. 24
4.4 A comparison of the micro and macro average precision, recall and F-score between the regular situation and one where a classification is considered correct as long as the dialogue act is in the top 5. 31
4.5 The average number of utterances available for training and testing as well as the average precision, recall and F-score per class. 32
4.6 The average number of utterances available for training and testing as well as the average precision, recall and F-score per class when a classification is counted as correct as long as the dialogue act is found in the top 5. 33

5.1 The average recall of the tree classifier when using the existing dialogue acts associated with the leaves as classifications. 51
5.2 The average recall of the tree classifier when the leaves themselves are used as classifications. 52
5.3 The average recall of a regular classifier that uses the leaves of the tree as dialogue act labels for the utterances. 53

A.1 Each dialogue act and its corresponding tag. 59
A.2 The total number of utterances per dialogue act. 60
A.3 The number of utterances per dialogue act in each fold. . . . 61
A.4 An overview of the part-of-speech tags. 62

B.1 Average recall per dialogue act with the first n words as features for each $n \in [1, 10]$. 64
B.2 Average recall per dialogue act with the first n part-of-speech-tags as features for each $n \in [1, 10]$. 65
B.3 Average recall per dialogue act with the previous n speakers and dialogue acts as features for each $n \in [1, 10]$. 66
B.4 Average recall per dialogue act for every type of feature individually. 67

B.5 Average recall per dialogue act for combinations of features. The first nine words were used as the baseline and all other features were added on one by one. 68
B.6 The confusion matrix of a classifier that uses all seven types of features. Part one of three. 70
B.7 The confusion matrix of a classifier that uses all seven types of features. Part two of three. 71
B.8 The confusion matrix of a classifier that uses all seven types of features. Part three of three. 72

C.1 The characteristics of the nodes in the classification tree. All grey rows as well as the child nodes marked in bold correspond to leaves. 73

List of Abbreviations

NLP	**N**atural **L**anguage **P**rocessing
NLU	**N**atural **L**anguage **U**nderstanding
NLG	**N**atural **L**anguage **G**eneration
DM	**D**ialogue **M**anager
SRL	**S**emantic **R**ole **L**abelling
pp	**P**ercentage **P**oint(s)

Chapter 1

Introduction

Conversational agents, programs with which users can engage in conversation using natural language, are an attractive concept because they can serve many purposes, from software interfaces to personal assistants to interactive game characters. Consequently a lot of research has been put into them in the last 50 years, but only with limited success so far.

As it turns out, understanding and producing natural language utterances is vastly more complex than early researchers expected. One reason for this complexity is that natural languages have evolved over time (and continue doing so) without conscious planning or forethought, leading to a lot of ambiguity, subtlety and exceptions among other things. Furthermore, humans break the rules of their language all the time for any number of reasons including laziness, ignorance and creative expression. Thus, unlike formal languages it is nigh impossible to fully capture a natural language and all its intricacies with a limited set of manually defined rules.

Another complicating factor is the sheer amount of knowledge necessary to actually grasp the meaning of an utterance. Each word is defined in relation to other words which all need definitions of their own as well. For example, a ball could be described as a round object, but that does not say much without knowing what "round" and "object" mean. As a result, understanding and reasoning about the meaning(s) of just a single word can require a vast network of concepts and the relations between them.

From these issues it becomes clear that the field of natural language processing (NLP) faces a lot of challenges ranging from understanding related tasks such as determining the structure of a sentence and resolving ambiguity, to processes that deal with generation such as choosing the right words to naturally express some idea. The NLP problem that this thesis focuses on is the recognition of so called *dialogue acts* by conversational agents. Dialogue acts are the actions that speakers perform through their speech such as giving or requesting certain pieces of information and expressing emotions like gratefulness and regret. They are vital to understanding what the speaker's communicative goals and intentions are. There has been a decent amount of research on the topic of dialogue act recognition, but in general the accuracy of the developed systems is not at a satisfactory level yet. Therefore, the main question is:

- **How can dialogue act recognition systems be made more accurate?**

There is one major restriction, though: the recognition system has to be useable for conversational agents. Since agents cannot look into the future, this means that only information contained in the preceding utterances may be used for the decision process. This complicates matters a little because subsequent utterances can contain valuable clues for pinpointing the correct

dialogue act. Additionally, there are two other problems. Firstly, the results of existing studies are not easily comparable because they vary a lot in their approach. Even if one system reportedly has a better performance than another, it could still fare worse when both are tested under equal circumstance. Secondly, most studies do not really offer any insight into the problem because they only report their system's overall performance (e.g., its accuracy, precision and/or recall). As a result, it is difficult to pinpoint which areas are problematic and why. Without that information, determining how to increase the performance is very hard. Therefore, the problem domain has to be analysed first before any possible improvements can be made. Specifically, the following questions have to be answered:

- **How well do existing systems recognise the individual dialogue acts?**
  Looking through the literature this question remains largely unanswered because most studies only give their classifier's overall accuracy. The problem is that this does not say much about the actual distribution: some dialogue acts could be recognised really well, whereas others might get very poor results. A breakdown of the performance per class would therefore be useful.

- **Why do existing systems fail at finding the right dialogue act in certain cases?**
  The system could fail for many reasons. The utterance might simply look too similar to instances of a different dialogue act for example, or it could be that there are not enough samples to train a system based on a statistical approach very well. Knowing where the problems occur is important, because it gives leads on how to improve the system's accuracy.

- **How do different utterance features affect the recognition rate?**
  The characteristics or features of an utterance play an important role in dialogue act recognition because they influence which dialogue act is seen as the most likely option. For example, an utterance that ends in a question mark is likely to be a question. Unfortunately, different studies tend to use different feature sets, so their results can be difficult to compare. It is therefore useful to compare the impact of the features under equal circumstances. Here, too, a breakdown of the results per dialogue act would be helpful, since there is the possibility that some features give a slight overall improvement, while also having a significant negative impact on some individual classes. In that case it is up to the situation if the trade-off is worth it or not.

This thesis is structured as follows. First, to get more acquainted with the topic chapter 2 gives some background information on conversational agents and important linguistic concepts, while chapter 3 contains an overview of the work that has been done on dialogue act recognition so far. Next, an analysis of existing approaches is given in chapter 4 and an alternative approach is presented in chapter 5. Finally, chapter 6 discusses some possibilities for future work and chapter 7 concludes the thesis.

Chapter 2

Background Information

This chapter provides some background information on natural language processing (NLP) and dialogue systems that is necessary to understand the remaining parts of the thesis. Section 2.1 gives an overview of several important linguistic concepts, section 2.2 discusses the two main approaches to NLP and section 2.3 discusses the general architecture of dialogue systems.

2.1 Linguistic Background

Every language, whether natural or artificial, is bound by certain rules – collectively referred to as a *grammar* – that dictate the ways in which it can be used. Properly understanding and producing utterances in a given language is only possible with adequate knowledge of these rules (Fromkin, Rodman, and Hyams, 2013, p. 9). Perhaps the most common usage of the term grammar is in reference to the rules that dictate morphology and syntax. However, more broadly it can also include phonology, semantics and pragmatics (Fromkin, Rodman, and Hyams, 2013, p. 9; Jurafsky and Martin, 2008, pp. 2-4). Phonology, which deals with how sounds combine to form words, will not be discussed further since the focus of this thesis is on written communication.

2.1.1 Morphology, Syntax and Morphosyntax

Morphology focuses on how words can be formed from smaller linguistic units (morphemes) such as roots, stems, prefixes and suffixes (Fromkin, Rodman, and Hyams, 2013, p. 37). For example, nouns such as "cat" get the suffix "-s" in their plural form, adjectives such as "beautiful" can be turned into adverbs by adding "-ly" and verbs such as "like" can be transformed in many different ways based on things like tense, person and number.

Syntax is concerned with how words can be grouped together to form phrases and sentences (Fromkin, Rodman, and Hyams, 2013, p. 77). Some languages, like English, have fairly strict rules on word order, whereas others are much more liberal. For example, in English sentences follow a subject-verb-object format ("the cat ate the food"), determiners ("the", "my", "this", etc.) must be placed before the noun they belong to and negation is done by putting the word "not" directly after certain verbs such as "do", "can" and "be".

The distinction between morphology and syntax is not always very obvious. The morphological properties of a verb, for example, often depend on its syntactic relation with other words as the following sentence shows:

The cat sits on the roof.

Here, the verb "sit" gets the suffix "-s" because of its relation to the singular noun "cat", the subject of the sentence. In situations such as these where both morphology and syntax are involved the term *morphosyntax* is used instead.

2.1.2 Semantics

The field of semantics studies meaning at different levels from morphemes to full sentences (Fromkin, Rodman, and Hyams, 2013, p. 140). Two subfields are *lexical semantics* and *compositional semantics*. Lexical semantics looks at the meaning of individual words and the relationships between them (Fromkin, Rodman, and Hyams, 2013, p. 153). Some examples of lexical relations are:

- **Synonymy**: words with the same meaning (e.g., pipe and tube).
- **Antonymy**: words with opposite meanings (e.g., wet and dry).
- **Hyponymy/hypernymy**: a hierarchical relation also commonly called *IsA*. The hyponym is the word whose meaning is included in that of its hypernym (e.g., "purple" is a hyponym of "colour", "colour" is a hypernym of "purple").

The meanings of words are accompanied by basic properties called semantic features that give additional information on how words relate to each other (Fromkin, Rodman, and Hyams, 2013, p. 158). For example, "red" and "blue" are both colours, "cat" and "dog" are animals and "yesterday" and "tomorrow" both indicate time. These features are important because they show which words can be combined together at the semantic level.

Semantic features are especially relevant for verbs since they restrict which noun phrases can be taken as arguments. The verb "play", for instance, requires one argument (a subject) and this argument, at the very least, has to have the property of being animate, because inanimate objects are not capable of playing. The arguments of a verb can be given specific names, also called thematic roles, to more clearly indicate how the verb and its arguments are related (Fromkin, Rodman, and Hyams, 2013, p. 163). Some examples are:

- **Agent**: the (typically sentient) argument that is performing the action.
- **Theme**: the argument that undergoes the action.
- **Recipient**: the argument that receives something as a result of the action.

Compositional semantics studies how a sentence's meaning is formed by combining the meanings of its individual words (Fromkin, Rodman, and Hyams, 2013, pp. 143-144). The meaning of a sentence can be viewed as its truth value, which can be calculated using set theory: all (meaningful) words or phrases represent a set and if the intersection of these sets is not empty, the sentence is true. Consider the following sentence:

A brown dog barks.

Here, "brown" refers to the set of all brown things, "dog" the set of all dogs and "bark" the set of all things that bark. The intersection then gives a set containing only dogs, all of which are brown and bark.

Compositional semantics cannot handle all situations equally well. Some examples of troublesome cases are paradoxes, idioms and anomalies. Paradoxes do not have a truth value because they contradict themselves and idioms cannot be decomposed because they have a fixed meaning. Anomalies are syntactically correct sentences that do not make any sense semantically such as "Colourless green ideas sleep furiously." (Fromkin, Rodman, and Hyams, 2013, p. 147). They are closely related to the metaphor, another difficult case that seems to be an anomaly at first, but still retains some meaning (e.g., "time is money").

2.1.3 Pragmatics

Semantics only looks at the literal meaning of utterances in isolation. Often, however, the exact same sentence can have different meanings depending on, among other things, who said it to whom at what point in the conversation and with what intention. In other words, *context* can have a significant influence on utterance meaning. For example, the sentence "It's cold in here." could be a simple factual statement, a request to warm the place or maybe even a sarcastic statement indicating that it is actually not cold at all. Pragmatics is all about figuring out what is meant when context is taken into account (Fromkin, Rodman, and Hyams, 2013, p. 140).

Context can influence meaning in different ways. A very direct one is through *deixis*, the use of words whose meaning changes based on the context (Fromkin, Rodman, and Hyams, 2013, p. 166). Pronouns ("he", "they", etc.), demonstratives ("these", "that", etc.), prepositions ("behind", "before" etc) and certain adverbs ("here", "tomorrow", etc.) are all examples of deictic words: while they have some meaning of their own they need to be supplemented by the context to get the whole picture. For example, the word "he" is generally only used to refer to a male person, but context is required to figure out which *specific* person it is referring to.

Deixis is just one manifestation of a more general attitude towards communication: to bring messages across as quickly, easily and, sometimes, as (socially) safely as possible. As a result, speakers tend to leave a lot of information unspoken and just assume the addressee will correctly fill in the gaps by reasoning about the context and what is actually said. This act of implying, but not outright stating something is called *implicature*. Some examples are given below (Fromkin, Rodman, and Hyams, 2013, p. 171):

Sue: Does Mary have a boyfriend?
Bill: She's been driving to Santa Barbara every weekend.

John: Do you know how to change a tire?
Jane: I know how to call a tow truck.

Dana: Do these slacks make my butt look big?
Jamie: You look great in chartreuse.

In each of these short dialogues the first person asks a yes/no question, but the second answers with a statement, rather than the expected yes or no. The first then has to figure out on their own which of the two is actually being implied.

Implicit communication works out because the pragmatic level has rules just like the morphological, syntactic and semantic level (Fromkin, Rodman, and Hyams, 2013, p. 171). One such rule is to only say things that are relevant in the current context. By following this rule (and believing that everyone else does too) it becomes easier to determine the implicit content of an utterance (if there is any). For instance, in the above dialogue between John and Jane, Jane's comment is expected to be directly relevant to John's, so since she is not answering him explicitly, it can be assumed that she is doing so implicitly.

Another use of the relevance rule is that it allows people to make *presuppositions*: assumptions about the state of the world based on implicitly represented information (Fromkin, Rodman, and Hyams, 2013, p. 174). For example, two presuppositions that can be made from Bill's answer in the dialogue above are that Mary is a woman and that there exists a location somewhere called "Santa Barbara". If either of these were not true, then Bill's utterance would not be relevant in the given context.

Besides the literal and implied meanings of utterances another essential component is the *intention* behind saying something: what is a speaker trying to achieve through communication? When Sue asks Bill about Mary, her intention is to get to know more about Mary's love life. Bill recognises this, so he makes a statement that not only answers Sue's question, but also gives some additional information (namely, that she seems to be seeing someone in Santa Barbara) instead of just a blunt yes or no.

An important insight into conversations is that it is possible to perform actions by speaking. Speakers use these actions to fulfil their communicative goals. For example, Sue's goal, as mentioned before, is to learn more about the state of Mary's love life and to achieve this she performs a "request for information" type of action. Such actions are fittingly called *speech acts*, although they are more commonly referred to as dialogue acts in the computer science literature.

The idea of speech acts was first popularised by Austin (Austin, 1962) who illustrated it with the following sentences:

I name this ship the Queen Elizabeth.
I give and bequeath my watch to my brother.
I bet you sixpence it will rain tomorrow.

Saying one of these sentences immediately results in carrying out the action indicated by its main verb (naming, giving, betting). Such performative verbs are the most obvious example, but their presence is not required to execute an action, as the next utterances show:

Sorry. (apology)
Call this ship the Queen Elizabeth. (order)
You can't do that. (protest)
Can you get me a drink? (request)

The last example is noteworthy because it is an indirect speech act: the speaker is probably not interested in knowing whether the addressee is physically capable of getting a drink, but is making a polite request to do so. Since such indirect and figurative language use is very common in conversations, it is not possible to fully understand all utterances by literal meaning alone. Speech acts can help fill this gap in representation between intended and actual meaning.

2.2 Approaches: Rules vs. Statistical Methods

There are two general approaches to NLP: one that uses rules and one that uses statistical methods. The former was the norm in the early days of NLP research. However, the limitations of this approach and the increasing popularity of machine learning led to a shift towards the latter in the late 1980s which has persisted to this day.

Rule-based systems rely on handwritten rules for language processing. A rule consists of a precondition and an action, and basically operates like an if-then statement: when the precondition is satisfied, the action is performed (Cambria and White, 2014). A simple example would be that if the input contains the word "hello", then a rule is triggered that forces the system to respond with "hi" or some other greeting. Rule-based systems are generally easy to manage, debug and understand (Chiticariu, Li, and Reiss, 2013), but as the number of rules increases, maintaining them becomes very difficult (Cambria and White, 2014). It can also be hard to define rules for complex domains, because it is easy to miss certain cases or exceptions and rules may end up interfering with each other (Jia, 2004). Manually describing all the rules is therefore a very time consuming tasks that often requires expert domain knowledge (Cambria and White, 2014).

Statistical approaches utilise machine learning techniques to automatically learn the general, probabilistic patterns underlying natural language. This is much faster and less labor intensive than manually defining rules, provided that there is a large representative data set available to train the system. Statistical methods are also more capable of dealing with ungrammatical or oddly phrased utterances, which are not uncommon in human dialogue (Nadkarni, Ohno-Machado, and Chapman, 2011). On the downside, a lot of training data is needed to gain good results which may not always be readily available. Furthermore, the trained models are not as easy to understand for humans because they mostly just consist of a bunch of numbers (Chiticariu, Li, and Reiss, 2013).

Rule-based and statistical methods do not have to be used exclusively: combining them is an option as well. The motivation behind hybrid approaches is that they can take advantage of the positive aspects of both without some or all of the negatives. For instance, a robust, broad-coverage statistical system could be used to preprocess the input after which the results are passed to a more narrow but precise rule-based system for further processing (Adolphs et al., 2008).

Pipeline architectures are an easy way to construct such a hybrid system. A pipeline is a sequence of subtasks that take the output of their direct predecessor as input. The implementation of each module is independent of the other modules, so some may use statistical approaches while others rely on rules depending on which of the two is the most effective for the given subtask. The main downside of pipelines is that not all problems can be cleanly cut into small pieces because some parts may, for example, need feedback from others to function well (Nadkarni, Ohno-Machado, and Chapman, 2011).

2.3 Dialogue System Architecture

Dialogue systems roughly consist of three components: a natural language understanding (NLU) module, a dialogue manager (DM) and a natural language generation (NLG) module. The DM forms the core of the system: it interprets user input, updates any state it may have as well as other relevant systems and then picks a proper response. The NLU and NLG modules essentially function as translators between human and computer language. The former extracts relevant information from natural language input and passes this along to the DM, while the latter takes the response from the DM and transforms it back into a natural language utterance. NLG will not be discussed further in this section because the focus is on processing natural language input, not generating output.

2.3.1 Dialogue Manager

DMs can be designed in various ways ranging from simple to very complex. Each kind has its own advantages and disadvantages, so it depends on the application which one will be the most suitable choice. Four different types of models that DMs can be based on are the finite-state, frame-based, information state and plan-based models (Jurafsky and Martin, 2008, p. 827).

Finite-state and Frame-based Models

Finite-state and frame-based systems are the most basic. The first only consists of a set of predefined states and transitions. The system produces the utterances associated with its current state and then the user's response triggers the transition to the next state. The second operates by filling in the fields in a form (also called frame). It keeps track of the parameters that are still undefined (e.g., travel date, destination) and asks the user questions to obtain the missing information. These questions do not have to be asked in a predefined order and the user can supply multiple pieces of information at once, so frame-based approaches are a bit more flexible than finite-state systems.

In both models the system is completely in control of the dialogue and the user may generally only respond with a restricted number of words and phrases. The result is a simple, easy to design system that does not need very sophisticated technology for things like speech recognition and NLU (McTear, 2002). Dialogue act recognition is likely not very relevant for these models either because the system dictates which action(s) the user can take out of a small set of possibilities. As a consequence, however, such a system is very inflexible and limited in its capability. Furthermore, it may become unwieldy if anything slightly more complex is needed such as allowing the user to go back and make changes (McTear, 2002).

Plan-based Model

Plan-based approaches view dialogues from a planning perspective where utterances are considered to be actions with which certain communicative goals can be achieved. This viewpoint of utterances is similar to that of dialogue acts. By chaining such actions together a plan can be formed to reach a given communicative goal. For example, if a travel planner's goal is to give users the information they need, it may come up with a plan that includes a couple "ask" actions to fill in any required parameters and then an "inform" action at the end to present the outcome.

A major difference with the other models is that these plans can be generated completely dynamically. As a result, it is possible to support more complex conversations where the user can take control too such as ones involving collaboration or negotiation. There are two downsides, however. Firstly, the difficulty of utterance interpretation increases a lot when more elaborate language is allowed and secondly, plan-based systems tend to be too slow and hard to manage when the problem domain is large. Therefore, systems based on this model are mostly limited to small, restricted domains that require complex dialogues. For applications that only need basic conversations it is better to choose one of the simpler models.

Information State Model

The information state model consists of an information state, dialogue moves, update rules and update strategy. The information state contains many kinds of information such as the dialogue history, speakers' beliefs and intentions, and what type of response to give next. Similar to dialogue acts, dialogue moves are actions such as "ask" and "answer" that are performed through natural language utterances and other forms of communication such as body language. The information state is changed by applying certain update rules to it once user input has been received. Not all rules are used in every situation, however: the dialogue move(s) extracted from the user input determine which rules are applicable and the update strategy decides which ones are actually triggered and in what order.

The information state approach was designed to be a combination of finite state and plan-based approaches with the intention to get the best of both worlds (Larsson and Traum, 2000). It is more flexible than finite-state systems because there are no fixed, predefined states and transitions. At the same time it is also easier to manage and interpret than plan-based systems

because the update rules and strategies limit what the system can do. In addition, the model can work with more complex types of information such as beliefs, desires and intentions that are common in plan-based approaches.

2.3.2 Natural Language Understanding

One of the most common approaches to NLU at the time of writing is to simply look for predefined keywords and phrases in the input. This cannot really be called "understanding", however: it is more akin to a dog that has learned to recognise a small number of commands. Nevertheless this method is generally fine for simple DMs such as finite-state or frame-based ones. These systems typically do not allow the user much freedom of expression, so there is not much of a need for true understanding.

Information state and plan-based DMs are a different matter, however. These at the very least need to know the dialogue act of the input utterance, because dialogue acts are one of their core components. Additionally, semantic information may also be desired depending on the expected complexity of the conversations. These systems therefore require a more advanced NLU module that is capable of analysing the input on the semantic and pragmatic level.

Semantic Analysis

The main goal of the semantic analysis is to be able to answer *who* in a sentence did *what* to *whom when where why* and *how*. These wh-words all point out semantic roles and the relations between them, which are typically indicated by verbs. For example, in the sentence "He sold her a book." the verb "sold" is the relation and "he", "her" and "book" can be given the roles "seller", "buyer" and "goods" respectively. Determining which parts of a sentence fulfil what roles is often referred to as semantic role labelling (SRL) or shallow semantic parsing.

SRL is still an open problem that is complicated by the sheer amount of knowledge the system needs to have about words, concepts and the relations between them. Several large, free resources that can be used for this purpose are PropBank (Kingsbury and Palmer, 2002), FrameNet (Baker, Fillmore, and Lowe, 1998) and VerbNet (Kipper-Schuler, 2005), but none of them are even close to complete. Aside from that ambiguity is also an issue as many words can have multiple different meanings and functions in a sentence. The word "rose", for instance, can be either a noun referring to a type of flower or the past tense of the verb "to rise".

A common approach to SRL is to divide it into four subtasks (Das et al., 2014; Màrquez et al., 2008). The first step is to find the *target* of the sentence or phrase, which is usually its main verb or predicate. Next is to determine the target's intended meaning in the given situation, which dictates which semantic roles or *arguments* have to be fulfilled. Then, all word sequences that correspond to an argument have to be identified and finally each of the resulting arguments is assigned a role label. All four tasks are typically handled in different ways, though some approaches combine two or more tasks into one. Both hand-written rules as well as statistical methods have been used for these tasks with varying degrees of success.

SRL has not been used yet to aid the dialogue act recognition process, but it might prove useful in the future as there are many utterances that are difficult to categorise based on things like syntax and keywords alone. A simple example is a yes-no question that is not answered with a yes or no, but a statement that implies one of the two such as the question-answer pairs found in section 2.1.3. In those cases semantic information as well as the ability to reason would be very helpful to determine the utterance's dialogue act.

Pragmatic Analysis

The aim of the pragmatic analysis is to determine how an utterance's meaning is influenced by the dialogue context. This can involve different subfields such as deixis and conversational implicature. However, one of the more important areas is dialogue act recognition as dialogue acts play a major role in both the information state and plan-based dialogue managers.

Dialogue acts offer two advantages to dialogue systems. Firstly, they help reduce the complexity of the dialogue manager by forming an abstraction between its tasks and the huge amount of possible natural language utterances (Larsson and Traum, 2000). For example, all user input that requests information will have to be handled in more or less the same way even if they have a completely different structure, word choice or meaning. Secondly, dialogue acts constrain the possible responses the system can give. A question generally has to be followed by an answer after all and it makes little sense to suddenly reply with a greeting in the middle of conversation.

For simple, restricted systems a rule-based approach to dialogue act recognition might work because some keywords ("yes", "no", "hello", etc.) and basic sentence and dialogue structures can be strong predictors for certain dialogue acts. However, as is common with rule-based systems this method is unlikely to scale well to more complex systems where the user has more freedom of expression. Therefore, in practice this approach is ignored by the literature in favour of a statistical one where the system is trained to identify dialogue acts with supervised machine learning techniques. This method is discussed more in-depth in chapter 3.

Chapter 3

Related Work

Dialogue act recognition has received a decent amount of attention in the past two decades. However, like many other NLP tasks it is not a trivial problem because of the complexity of natural languages. Thus, while there have been some advancements in the field there is still a lot of room for improvement.

As mentioned in chapter 2.3.2, the general approach to dialogue act recognition is to use a statistical classifier that is trained with supervised machine learning techniques. Classification is then done in two steps. Firstly, *features* are extracted from the utterance, which are specific characteristics such as the utterance's length or the kinds of words that occur in it. Secondly, these features are used by a classification algorithm to determine the most likely dialogue act label. Some classifiers may also return multiple results ordered from most to least likely instead of just a single one. Training and testing of the classifier is typically done using a large annotated data set called a *corpus*.

The dialogue act labels, corpus, classification algorithm and features all influence the performance of the classifier. The first two set the difficulty of the problem: simple, constrained language use and a small set of broad dialogue act categories result in an easier task than unrestricted language and a large, very precise set of dialogue acts. The latter two, on the other hand, determine how capable the classifier actually is at solving this problem: some algorithms are more suited to certain tasks than others and the features need to provide enough information to accurately discriminate between the different dialogue act classes.

Unfortunately, there is not really a standardised way to conduct research on dialogue act recognition. As a result, there is a lot of variety in the methodologies used by different studies, making it difficult to compare them to each other. In addition, these studies do not really give an in-depth report on their results, so it is not clear which areas they are struggling with and why. Therefore, while they are a useful starting point, the information they provide is not enough to determine how to improve on the existing systems. Nevertheless, this chapter will attempt to give an overview of what has been worked on so far. Section 3.1 discusses several different dialogue act taxonomies, section 3.2 reviews a few large corpora that can be used to train the classifier and sections 3.3 and 3.4 list commonly used features and algorithms respectively.

3.1 Dialogue Act Classes

Austin (Austin, 1962) described several basic speech act categories in his work on speech acts. These were later improved on by Searle (Searle, 1976) resulting in the following five categories:

- **Representatives**: acts that commit the speaker to something being true such as *suggesting*, *hypothesising*, *insisting*, *swearing*, *concluding* and *complaining*.
- **Directives**: acts through which the speaker attempts to get the addressee to do something such as *ordering*, *requesting*, *pleading*, *inviting*, *advising*, *defying* and *challenging*.
- **Commissives**: acts that commit the speaker to some future course of action such as *promising*, *planning*, *threatening*, *swearing*, *vowing* and *betting*.
- **Expressives**: acts that express the speaker's psychological state about some state of affairs such as *thanking*, *congratulating*, *apologising*, *condoling* and *welcoming*.
- **Declarations**: acts that change the state of the world in some way such as *resigning*, *nominating*, *firing*, *marrying*, *declaring* and *convicting*.

Representative, commissive and expressive acts come in varying degrees of strength. Insisting that something is true is a much stronger commitment than suggesting it, after all. Likewise, *ordering* is a more forceful attempt than *pleading* and *vowing* is more drastic than *planning*. The verbs associated with these categories are also not restricted to just a single one. Swearing, for example, can indicate either a representative ("I swear it's true!") or a commissive ("I swear I will do it!").

A more extensive annotation scheme is DAMSL (Core and Allen, 1997). Unlike Searle's categorisation (Searle, 1976), DAMSL allows an utterance to have more than one label because in practice it is possible to perform multiple actions with just one utterance. Another difference is that it supports subclassing, so each system can add extra classes when needed while still remaining comparable to others at a high level.

SWBD-DAMSL (Stolcke et al., 2000) is a modified version of DAMSL that contains hundreds of possible dialogue act tags if not more. 220 of them were actually used to annotate the Switchboard corpus (Godfrey, Holliman, and McDaniel, 1992). These were then clustered to remove a lot of very small classes, resulting in a more manageable set of just 42 dialogue act tags. Table 3.1 gives an overview of the remaining classes. SWBD-DAMSL is often used as a basis for other tag sets and then modified or added to as needed (e.g., Shriberg et al., 2004; Wu et al., 2005).

There is no general agreement on how to categorise dialogue acts and one criticism of expert defined taxonomies is that they are based on intuition rather than hard data (Andernach, Poel, and Salomons, 1997). As a result, it can be difficult for humans to tag utterances with the right dialogue act consistently. Therefore, there has also been some experimentation with clustering algorithms to automatically find categories, which were then given a semantic label by a human judge afterwards (Rus et al., 2012).

TABLE 3.1: The 42 clustered SWBD-DAMSL dialogue act tags.

Dialogue Act	Description
Uninterpretable	Utterances that cannot be interpreted for any kind of reason. This is commonly the case when speakers do not finish their sentences.
Collaborative completion	The speaker completes the other's words.
Tag-question	Questions such as "Okay?", "Right?", "Don't they?" and "Isn't it?".
Hold before answer or agreement	Utterances such as "Uhm..." and "Let's see..." that precede an answer.
Quotation	A repetition of someone else's words.
Agree/accept	Utterances such as "Right.", "Okay." and "Yeah." that signal agreement with or acceptance of the previous speaker's words.
Action-directive	Utterances that push the addressee to do something such as "Go ahead. or "You've got to..."
Maybe/accept-part	Utterances that express doubt over whether to accept what was previously said such as "Maybe.", "I guess." and "I don't know.".
Reject	A rejection of what was previously said. Typically in the form of a "No.".
Acknowledge (backchannel)	An acknowledgement that the other person said something. E.g., "Okay.", "Uh-huh." or "Yeah.".
Repeat-phrase	An utterance that repeats (part of) the previous one.
Appreciation	Expressions of appreciation for something, such as "That's good." or "Oh wow.".
Downplayer	Downplaying response to sympathy or compliments such as "That's all right.", "It happens.".
Summarize/reformulate	A summarisation/reformulation of the other's utterance(s).
Backchannel in question form	A rhetorical question that pushes the other person to tell more about the current topic. E.g., "Oh, really?" or "Is that right?".
Response acknowledgement	Acknowledgement of an answer given by the previous speaker. E.g., "Oh, okay." or "Oh, I see.".
Signal-non-understanding	Utterances that indicate that the speaker did not understand something such as "Huh?", "Pardon me?" or "What do you mean?".
Offers, options and commits	Utterances that are an offer/commit the speaker to do something or present a choice between options. E.g., "Let me...", "We could...", "I'll do it.". This is a cluster of three very small classes.
Apology	Apologies such as "I'm sorry." or "Excuse me.".
Conventional-closing	Utterances that close the conversation such as "Bye." and "Nice talking to you.".
Conventional-opening	Utterances that appear during the opening part of the conversation such as "Hello.", "My name is..." and "How are you doing?".
Thanking	Expressions of gratitude such as "Thanks".
Hedge	Utterances that "soften" what was said previously or prevent any kind of commitment. E.g., "I don't know.", "Maybe..." and "I'm not sure.".
Affirmative non-yes answer	A descriptive/narrative statement which acts as an affirmative answer to a question.
Dispreferred answer	Utterances that express (partial) disagreement with the other person.
Negative non-no answer	A descriptive/narrative statement which acts as a negative answer to a question.
No answer	A negative answer usually in the form of a "No.". "Huh-uh." is also fairly common.
Other answer	An answer that does not fall in any of the other answer categories.
Yes answer	A positive answer in the form of a "Yes." or "Yeah.".
Other	Utterances that do not fall in any of the other categories. Most commonly in the form of an "Okay.".
Rhetorical-question	Questions asked to make a point, not to get an answer. E.g., "Can't you do anything right?".
Open-question	Open ended questions such as "What do you think?" or "How about you?".
Or-clause	An or-question (questions that offer an option between multiple choices) tacked onto a yes-no question. E.g., "Or is it more of a company?"
Wh-question	Questions that start with a wh-word, such as "who", "what", or "how".
Declarative wh-question	Wh-questions written in the declarative form.

Continued on next page

Table 3.1 – continued from previous page

Dialogue Act	Description
Yes-no-question	Questions that can be answered with "yes" or "no".
Declarative yes-no-question	Yes-No-Questions in the declarative form.
Statement-non-opinion	Descriptive and/or narrative statements that cannot be disputed.
Statement-opinion	Any kind of viewpoint, from personal opinions to proposed general facts, that can be disputed.
Self-talk	Utterances where the speaker is talking to themselves such as "Oh, what was it?".
3rd-party-talk	Anything said by someone other than the two primary speakers.
Non-verbal	Non-verbal utterances such as laughter and coughing.

3.2 Corpora

There are many different corpora that could be used for dialogue act recognition. Only a few of them are already annotated with dialogue act tags, however. Most studies end up using one of these corpora to save time, but some opt to annotate a corpus themselves that better suits their needs. Four publicly available corpora that have been annotated with dialogue acts are Switchboard (Godfrey, Holliman, and McDaniel, 1992), the ICSI Meeting corpus (Janin et al., 2003), the HCRC Map Task Corpus (Anderson et al., 1991) and NPS Internet Chatroom Conversations (Forsythand and Martell, 2007). Table 3.2 gives an overview of these corpora.

TABLE 3.2: An overview of four different corpora used to train dialogue act classifiers.

Corpus	Size	Speakers	Type	dialogue act labels
Switchboard	1 155 conversations 205K utterances 1.4M words	2	Spoken Casual	42 clustered 220 total
ICSI Meeting	75 conversations 795K words	3-10 (avg. 6)	Spoken Slightly structured Mostly casual	11 general 39 specific
HCRC Map Task	128 conversations 26 621 utterances	2	Spoken Task-oriented	12
NPS Chat	15 conversations 10 567 utterances 45K words	N/A	Written Casual	15

3.2.1 Switchboard

Switchboard (Godfrey, Holliman, and McDaniel, 1992) is a large collection of casual telephone conversations between two people about all kinds of subjects. Since the participants just had to chat instead of achieving a particular goal, the type of speech is mostly unrestricted and can be any of a large variety of dialogue acts. The conversations are available in two forms: the original audio recordings and their transcripts. 1155 dialogues (205K utterances, 1.4M words) from the corpus have been annotated with 220 tags from the SWBD-DAMSL tag set (Stolcke et al., 2000). There is also a clustering of these tags available which reduces that number to 42 by removing a lot of very small classes (Stolcke et al., 2000).

The highest reported accuracy on the Switchboard corpus appears to be 89.2% (Margolis, Livescu, and Ostendorf, 2010). However, that study only used four dialogue act classes (Incomplete, Statement, Question and

Backchannel), which likely simplified the problem. Similarly, another study achieved an accuracy of 80.72%, but that was also after clustering of the tag set (Webb and Ferguson, 2010). The best result that has been achieved so far on the full clustered tag set seems to be 77.85% (Gambäck, Olsson, and Täckström, 2011).

3.2.2 ICSI Meeting Corpus

The ICSI Meeting corpus (Janin et al., 2003) contains 75 meetings (795K words) of ICSI working teams with an average of six people per group. The conversations are a bit more structured than those found in Switchboard because the speakers had agenda points to talk about, but they are still fairly casual (Shriberg et al., 2004). Like Switchboard, both audio recordings and transcripts are available. The corpus has been annotated with the MRDA tag set (Shriberg et al., 2004), which is a modification of SWBD-DAMSL. This set contains 11 general tags and 39 specific tags. Each utterance is labeled with one of the general tags and a variable number of specific tags.

The best accuracy attained on this corpus seems to be 89.27%, but as with Switchboard, this was with only five dialogue acts instead of the full tag set (Verbree, Rienks, and Heylen, 2006). Another study that used just the 11 general tags achieved 80.5% (Tavafi et al., 2013) and one that used 62 tags obtained by clustering the full set recorded 66% (Ji and Bilmes, 2005).

3.2.3 HCRC Map Task Corpus

The HCRC Map Task Corpus (Anderson et al., 1991) consists of 128 audio recordings of pairs working together to complete a certain task. As such, the dialogues are *task-oriented*, which results in more structured and restrained language use than casual conversations. Transcriptions are available with multiple kinds of annotations including dialogue acts.

The corpus has been annotated with a tag set containing just 12 dialogue acts (Carletta et al., 1997). These dialogue acts focus primarily on the task at hand: some examples are *Instruct*, *Explain*, *Query-YN* (yes-no-question) and *Reply-Y* (yes answer). Thus, there are no classes that have more of a social function such as greetings and apologies. The highest accuracy achieved on this corpus appears to be 73.91% (Serafin and Di Eugenio, 2004). The corpus seemss to have fallen out of favour in the past decade, as it is not used very often anymore.

3.2.4 NPS Internet Chatroom Conversations

NPS Internet Chatroom Conversations (Forsythand and Martell, 2007) contains 10567 utterances (called "posts") from 15 online chatrooms. What sets the corpus apart is that it focuses on computer-mediated communication instead of traditional spoken or written conversations. Thus, the domain has a couple unique quirks such as the use of emoticons and abbreviations (Wu et al., 2005).

The corpus has been annotated with a tag set consisting of 15 dialogue acts (Wu et al., 2005). Most of the dialogue acts in the set were chosen from other tag sets, including SWBD-DAMSL. A couple others such as *Emotion*

were added to fulfill the specific needs of the domain. The highest performance on the corpus so far seems to be 71.9% (Moldovan, Rus, and Graesser, 2011).

3.3 Features

The main focus of most research on dialogue act recognition is finding the features that lead to the best classification results. A feature can be any kind of characteristic from the utterance such as its length or the presence of a certain word. The following ones are all commonly used to varying degrees of success, though this is by no means an exhaustive list:

- **n-grams**: n-grams are sequences of n words with n usually between 1 and 4. They are by far the most common choice because the presence of certain words or phrases can be a powerful cue. They do not necessarily have to be sequences of *words*: other tokens such as part-of-speech tags are also an option. Used by Stolcke et al., 2000; Webb, Hepple, and Wilks, 2005; and Verbree, Rienks, and Heylen, 2006.

- **First n words**: there are indications that humans can generally tell the dialogue act of an utterance within the first few words they hear (Jurafsky and Martin, 2008, p. 814; Gisladottir et al., 2012). Thus, the initial words of a sentence are assumed to be quite informative and a good choice to use as features. Used by Ang, Liu, and Shriberg, 2005; Moldovan, Rus, and Graesser, 2011; and Rus et al., 2012.

- **Length**: the number of words in the utterance. Some dialogue acts such as simple yes/no responses are consistently very short whereas others tend to be longer, so length may help distinguish these groups from each other. Used by Webb, Hepple, and Wilks, 2005; Verbree, Rienks, and Heylen, 2006; Tavafi et al., 2013.

- **Previous n dialogue acts**: some dialogue acts such as questions and answers often co-occur with each others, so the preceding dialogue acts can be an indication of the current one. Such contextual elements could be especially helpful in cases where the exact same utterance can have different dialogue acts depending on the situation. Used by Verbree, Rienks, and Heylen, 2006; Kim, Cavedon, and Baldwin, 2010; Petukhova and Bunt, 2011.

- **Presence of punctuation marks**: some punctuation marks are very characteristic of certain dialogue acts. Question marks, for example, indicate that an utterance has a high chance of being some kind of interrogative dialogue act. Used by Gambäck, Olsson, and Täckström, 2011; Rus et al., 2012; Omuya, Prabhakaran, and Rambow, 2013.

3.4 Classification Algorithms

Many different classification algorithms have been used for dialogue act recognition, some of which are listed in table 3.3. Support vector machines (SVM) appear to be used the most, sometimes combined with hidden Markov

models (HMM) into a hybrid system. Also popular are conditional random fields (CRF) and naive Bayes. Bayesian networks, maximum entropy, logistic regression and decision trees are used less commonly. Curiously enough, it does not look like neural networks have been tried so far despite having become a very popular choice for machine learning tasks in recent years. Given that dialogue act recognition does not receive as much attention as some other fields such as computer vision, it is likely that this is simply because no one has gotten around to it yet.

TABLE 3.3: An overview of the different classification algorithms used for dialogue act recognition.

Classification Algorithm	**Examples of Use**
Support vector machines (SVM)	Fernandez and Picard, 2002; Margolis, Livescu, and Ostendorf, 2010; Tavafi et al., 2013
A combination of an SVM and hidden Markov model. (SVM-HMM)	Surendran and Levow, 2006; Kim, Cavedon, and Baldwin, 2010; Tavafi et al., 2013
Conditional random fields (CRF)	Kim, Cavedon, and Baldwin, 2010; Tavafi et al., 2013
Naive Bayes	Kim, Cavedon, and Baldwin, 2010; Moldovan, Rus, and Graesser, 2011; Samei et al., 2014
Bayesian networks	Klüwer, Uszkoreit, and Xu, 2010
Maximum entropy	Ang, Liu, and Shriberg, 2005; Sridhar, Bangalore, and Narayanan, 2009
Logistic regression	Boyer et al., 2010
Decision trees	Moldovan, Rus, and Graesser, 2011; Samei et al., 2014

The results of the classification algorithms cannot be accurately compared unless they were trained and tested in the same environment. This is unfortunate, because most studies use only one of them. A few did test multiple classifiers, however. Two studies found that naive Bayes is outperformed by decision trees (Moldovan, Rus, and Graesser, 2011; Samei et al., 2014) and one that it does worse than both SVM-HMMs and CRFs (Kim, Cavedon, and Baldwin, 2010). In that same study CRF achieved better results than SVM-HMM. In another, however, it did worse than SVM-HMM, but better than a regular SVM (Tavafi et al., 2013). In the latter case the circumstances were not entirely equal, though, because the CRF took the dialogue structure (e.g., the sequence of dialogue acts) into account, while the SVM did not.

Chapter 4

Analysis

Several experiments were carried out to gain more insight into the problem domain and the efficacy of the different types of features. Firstly, the influence of different values of n was examined for the features that rely on a variable number n. Secondly, all seven feature types were tested individually to see how effective they are on their own. Thirdly, different combinations of features were investigated to find the one that leads to the best performance. The methodology is given in section 4.1, the results are presented in section 4.2 and a discussion is found in section 4.3.

4.1 Methodology

Multiple classifiers were trained using a large corpus annotated with dialogue act tags. Each classifier used a different set of features, but the same classification algorithm and the same data for training and testing so that the effects of the different features on the performance could be properly compared. The corpus and features that were used are detailed in sections 4.1.1 and 4.1.2 respectively. Section 4.1.3 discusses how the tests were evaluated and section 4.1.4 gives some details about the implementation.

4.1.1 Corpus

The analysis made use of the Switchboard corpus with the SWBD-DAMSL dialogue act tag set (see chapter 3) to train the classifiers. It was chosen because it is very large, freely available and commonly used by related studies. Of the 42 dialogue act tags 3 were excluded: *3rd-party-talk*, *Non-verbal* and *Uninterpretable or Abandoned*. 3rd-party-talk utterances are spoken by someone other than the two dialogue participants and non-verbal ones are things like laughter and coughs. These dialogue acts were left out because the first is not really a dialogue act, the second does not normally occur in written conversations and the third is uninterpretable. Table 3.1 gives an overview of all 42 tags.

Utterance Modifications

Many of the Switchboard utterances contain parts that are marked with extra pieces of information. Square brackets, for example, indicate that the speaker is repeating or correcting themselves. Since these markings are only used in this particular corpus and are not really needed anyway, the utterances were first cleaned into a more readable format. Table 4.1 shows the different markings and how they were filtered, and table 4.2 contains

some examples. The remaining number of utterances per dialogue act after the modifications can be found in appendix A.

TABLE 4.1: The types of markings occurring in the transcripts and how the texts are modified to filter them out.

Marking	Description	Modification
< >	Non-verbal sounds such as laughter.	The brackets and the content they enclose are removed.
{ }	Mark certain types of words such as conjunctions (and, or, but, etc.) and fillers (uh, oh, um, etc.). The opening bracket is followed by a capital letter to indicate the type.	The brackets and the capital letter are removed.
(( ))	Mark parts where the transcriber could not hear or was unsure about what was said.	All double round brackets are removed because such situations do not occur in written conversations.
[ ]	Indicate that the speaker is correcting or repeating themselves. The original and corrected word or phrase are separated by a +.	Only the correction is kept to make the text more fluid.
–	Indicate that the utterance is a continuation of a previous one (e.g., because of interruptions or speaking at the same time). They complicate the classification process, because one part is usually a short one that does not have much of a function on its own. Therefore, multi-part utterances are combined into a single one.	The dashes are removed and the original utterance is combined with its continuation. The combination is positioned in the dialogue at the spot of the longest of its two parts.
/	Mark the end of an utterance.	All slashes are removed.
#	Marks parts where multiple people are speaking at the same time.	All hashtags are removed because overlaps do not occur in written conversations.
*	Indicates commentary by the transcriber or someone else.	All asterisks and the corresponding commentary are removed.

Subsets

The corpus was split into ten subsets of approximately the same size for 10-fold cross validation. This was done by randomly assigning each dialogue to one of the ten groups. The splits were made at the dialogue level instead of the utterance level because contextual information (e.g., from the

TABLE 4.2: Examples of utterances before and after cleaning.

Marking	Original Version	Cleaned Version
< >	Wow <laughter>.	Wow.
{ }	{C and } I think that's probably a little over kill for this day and age.	and I think that's probably a little over kill for this day and age.
(( ))	I used to have a (( )) Chevy.	I used to have a Chevy.
[ ]	[I, + uh, I] drive [[a, + a truck,] + a Ford truck.]	uh, I drive a Ford truck.
–	A: and you just get so depressed for the US auto – B: Yeah. A: – makers when you do that.	A: and you just get so depressed for the US auto makers when you do that. B: Yeah.
/	Uh-huh. /	Uh-huh.
#	A: # Uh-huh. # B: # With a # three fifty in it.	A: Uh-huh B: With a three fifty in it.
*	I guess the job that I'm in it's stuff to stay on any kind of a regular schedule. / *[[listen; possible typo - stuff = tough?]]	I guess the job that I'm in it's stuff to stay on any kind of a regular schedule.

preceding utterances) can be used for the classification process. Most dialogues are roughly the same length, though, so the number of utterances in each set approaches the same ratio as the dialogues anyway.

Since a random selection process does not necessarily lead to a balanced split, many different combinations were generated. The final one was chosen according to two criteria: firstly, all sets had to approach the 1/10 ratio as closely as possible on both the dialogue and utterance level, and secondly, every dialogue act had to appear at least once in each set. The second requirement is important because several classes contain only a handful of utterances. As a result, they could easily end up unrepresented in some sets. Table 4.3 shows the size of each of the ten chosen subsets. The number of utterances in each fold per dialogue act can be found in appendix A.

4.1.2 Features

The following features were chosen to be tested because they are also commonly used by other studies (see chapter 3.3):

- **First n words**: a total of n features that each record one of the utterance's first n words. Every word was lemmatised so that all its variations would result in the same feature. For example, "the cats played" would result in the same three features as "the cat played". Utterances that contained less than n words were padded with blank fillers to make sure there would always be exactly n features.

TABLE 4.3: The size of each of the ten subsets the corpus was split into.

Subset	Nr. of dialogues	%	Nr. of utterances	%
1	114	9.9%	17900	10.3%
2	109	9.4%	16902	9.7%
3	120	10.4%	18127	10.4%
4	115	10.0%	17270	9.9%
5	118	10.2%	17269	9.9%
6	114	9.9%	16915	9.7%
7	119	10.3%	18074	10.4%
8	114	9.9%	16488	9.5%
9	115	10.0%	17309	9.9%
10	117	10.1%	17774	10.2%

- **Part-of-speech tags of the first n words**: same as the previous category, but the part-of-speech-tag is recorded instead of the word. This captures structural information instead of the exact words. A list of all the possible part-of-speech tags can be found in appendix A.
- **Utterance length**: the number of words in the utterance.
- **Presence of a question mark**: a binary feature that records whether the utterance contains a question mark or not.
- **Previous n speakers**: a total of n binary features, one for each of the n previous speakers. They record whether their allocated speaker is the same as the current one or not.
- **Previous n dialogue acts**: a total of n features that record the dialogue acts of the previous n utterances.
- **Presence of specific n-grams**: one binary feature for each of a select number of n-grams that records whether said n-gram is present in the utterance or not. The n-grams are lemmatised as well.
- **Presence of specific part-of-speech tag n-grams**: similar to the previous category, but part-of-speech tags are used instead of words.

The features were first tested individually and then added one by one to see how well they would do when combined together. All features that rely on a variable number n with the exception of the n-grams were tested for each $n \in [1, 10]$. The values of n that led to the best results were used during the remaining experiments.

A total of 400 n-grams were automatically selected from the training data: 100 for each $n \in [1, 4]$. Since the ten subsets of the corpus were all associated with a slightly different set of training data, the n-grams differed a little between them as well. The n-grams were chosen according to the following algorithm:

1. Extract *all* n-grams from the training data for a specific value of n.
2. Discard any n-gram that only occurs once in the entire training set, because it is unlikely to be very characteristic of any dialogue act.

3. Calculate a value for each n-gram that represents how informative it is and use it to rank the n-grams.

4. Keep only the 100 n-grams with the highest value. The threshold of 100 is rather arbitrary, but it was chosen because it gave a decent balance between performance and training time.

The value of an n-gram depends on how predictive it is for a specific dialogue act. This is measured according to two criteria: firstly, the n-gram must frequently co-occur with said dialogue act and secondly, it must not co-occur with too many other dialogue acts. For example, the unigram "the" might be present in most utterances tagged as statements, but it is still a bad predictor for this dialogue act because it is also very common in many other utterances. By contrast, the unigram "sorry" is a good candidate as it frequently occurs in apologies, but rarely in other dialogue acts.

A basic way to model the first criterium is as the percentage of all utterances tagged with the dialogue act that contain the given n-gram. The higher this percentage, the more commonly the n-gram co-occurs with the dialogue act. This is essentially the conditional probability $P(G|D)$ that the n-gram g is present in any utterance u from the set U_d of all utterances tagged with dialogue act d:

$$P(G = g|D = d) = \frac{|\{u|g \subseteq u, u \in U_d\}|}{|U_d|}$$

One issue with this representation is that it does not take into account how common the dialogue act is across the data set. If an n-gram has the same conditional probability for two dialogue acts, but one dialogue act occurs three times more frequently in the data set than the other, then 75% of all utterances containing the n-gram will be tagged with the common dialogue act, whereas only 25% will be tagged with the rarer one. Thus, the former is more likely to be correct when the n-gram is present than the latter. This can be modelled as what is essentially the opposite situation: the conditional probability $P(T|G)$ that an utterance u from the set U_g of all utterances containing n-gram g is tagged with dialogue act d.

$$P(D = d|G = g) = \frac{|\{u|g \subseteq u, u \in U_d\}|}{|U_g|}$$

The second criterium can simply be modelled as the inverse $I(g)$ of the total number of dialogue acts in the set D_g of all dialogue acts that the n-gram g co-occurs with at least once:

$$I(g) = \frac{1}{|D_g|}$$

Multiplying all three with each other then gives the value of n-gram g for dialogue act d:

$$V(g, d) = \frac{P(G = g|D = d) \cdot P(D = d|G = g)}{|D_g|}$$

This function is calculated for each dialogue act tag and then the highest value is used to rank the n-gram:

$$W(g) = \max_{d \in D} V(g, d)$$

While this method of ranking the n-grams is quite simple, initial tests showed that it worked fairly well, so it was kept as is. That said, better results may be possible with more sophisticated selection algorithms.

4.1.3 Evaluation

The test results are given in the form of a confusion matrix. From this matrix the following values are calculated to determine how well the classifier did:

- **Precision**: the fraction of all utterances classified as a certain dialogue act that actually received the correct tag. In other words, it measures how "precise" the classifier is when classifying. For example, a simple classifier that always assigns the majority class will be 100% accurate for that class, but 0% for all others, so it is not very precise. The precision is formally defined as:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives + false positives}}$$

- **Recall**: the fraction of all utterances belonging to a certain dialogue act that was correctly classified. Basically, it shows how well a dialogue act is being recognised. The recall is formally defined as:

$$\text{recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}$$

- **F-score**: the harmonic mean of the precision and recall. The F-score is simply a way to combine the precision and recall in one value. It is formally defined as:

$$\text{F-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision + recall}}$$

- A **confusion matrix** that contains percentages instead of absolute numbers. This matrix shows which classes are confused with each other and how often this happens, which is useful for pinpointing problem areas. The diagonal of this matrix contains the recall of each dialogue act.

The precision, recall and F-score are calculated for each dialogue act individually and then averaged to get the overall results, as these values are not defined for a multi-class situation. Two averages are calculated: the *micro average* and *macro average*. The macro average simply averages the results of each class, while the micro average is weighted by class size. Note that in a multi-class scenario the average recall is equal to the *average accuracy*, which is also frequently used to judge classification performance.

The micro average is a better reflection of the performance at the utterance level, but it does not show how well the different dialogue acts are actually recognised because it is influenced more by large classes than small

ones. As a result, it is possible to have a high micro average by just recognising one or two big classes really well and neglecting all the others. By contrast, the macro average gives a more balanced view of the performance on the dialogue act level, but does not say much about how many utterances were actually correctly classified.

The classifiers were tested with 10-fold cross-validation to reduce the influence of outliers and get a more generalised view. The final precision, recall, F-score and confusion matrix were therefore calculated by averaging the results of the individual folds.

4.1.4 Implementation Details

The classifiers were implemented in Python (2.7.12) using the NLTK[1] (3.2.1) library for natural language processing tasks and the scikit-learn[2] (0.17.1) library for any required machine learning algorithms. All classifiers used logistic regression as their classification algorithm with a one-vs.-rest approach to multi-class classification.

The initial plan was to use linear SVMs because they appear to be the most common choice of algorithm in the literature, but these have a downside: they cannot give the *probability* that an utterance belongs to a certain class, only the distance to the support vector. Knowing the probability for each dialogue is useful because the agent could base certain decisions on this information. For example, if all dialogue acts turn out to have a low probability, the agent could choose to ask for clarification instead of acting on apparently uncertain data. Since logistic regression *is* capable of giving probabilities and functions and performs similarly to linear SVMs, it was chosen instead.

4.2 Results

All results are compared to a baseline accuracy that is given by a classifier that simply always assigns the majority class. The most common dialogue act in Switchboard is *statement-non-opinion*, which has a frequency of 36%. However, after the corpus has been adjusted as described in chapter 4.1.1, this number rises to 40.5%, which is the final baseline. This section gives the overall results of the tested features. Additional information can be found in the appendix B.

4.2.1 First n Words

Figure 4.1 shows the average recall when only the first n words are used as features with n in the range [1, 10]. Initially the recall increases quite a bit as n is incremented, but beyond five the curve quickly flattens out. The best results were achieved for $n = 9$ and $n = 10$ with a micro average of 72.4% and a macro average of 31.6%. The micro average is an improvement of 31.9 percentage points (pp) over the baseline and 8.9pp over just using the first word.

[1] http://www.nltk.org/
[2] http://www.scikit-learn.org/

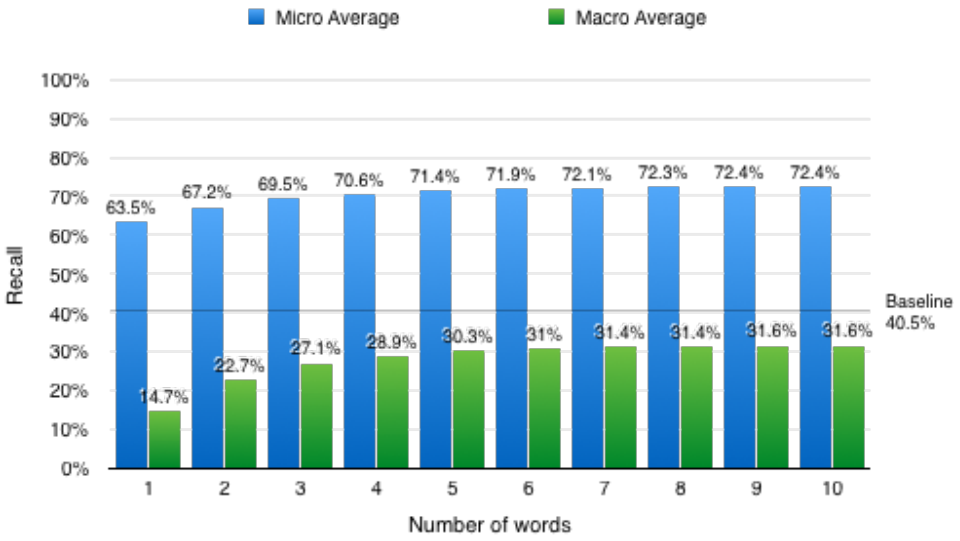


FIGURE 4.1: The average recall when only the first n words are used as features.

4.2.2 First n Part-of-speech Tags

The average recall when only the first n part-of-speech tags are used as features is shown in figure 4.2. There is a steep increase from $n = 1$ to $n = 2$, but after that the curve evens out. As with the first n words features, the best results are obtained for $n = 9$ and $n = 10$: a micro average of 65.1% and a macro average of 15.6%. The micro average improves 24.6pp over the baseline and 12.1pp over using just the first part-of-speech tag.

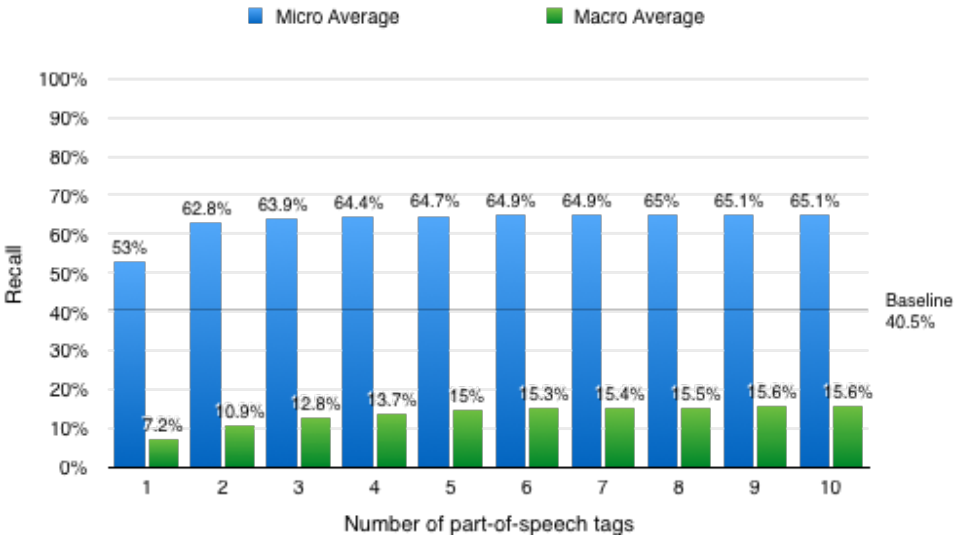


FIGURE 4.2: The average recall when only the first n part-of-speech tags are used as features.

4.2.3 Previous n Speakers and Dialogue Acts

Figure 4.3 shows the recall when the previous n speakers and dialogue acts are used as features with $n \in [1, 10]$. The curve is mostly just flat with very little variance. Only $n = 1$ gives slightly better results than the rest with a

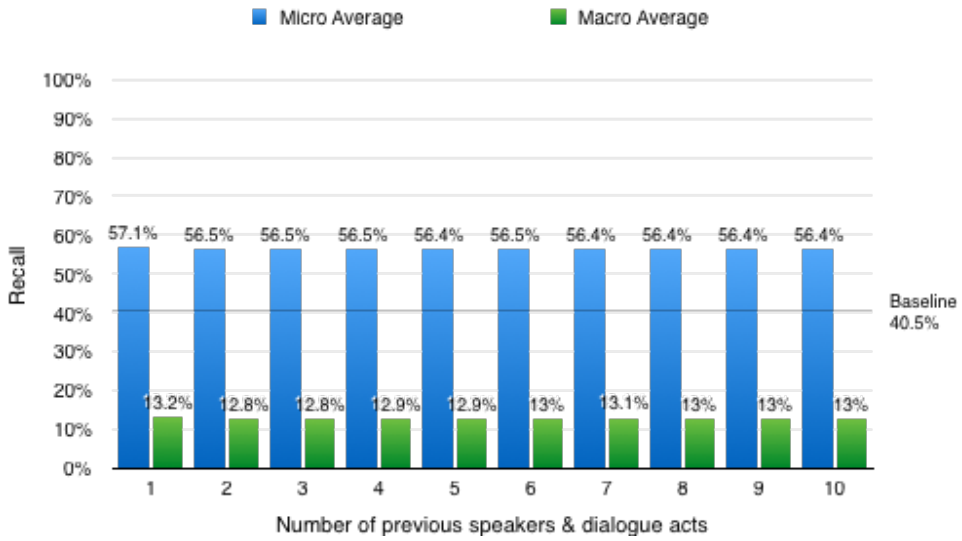


FIGURE 4.3: The average recall when only the previous n speakers and dialogue acts are used as features.

micro average of 57.1% and a macro average of 13.2%. This micro average improves 16.6pp over the baseline, but only 0.7pp over the lowest results obtained when n lies between 7 and 10.

4.2.4 Individual Features

The recall for each feature individually can be seen in figure 4.4. The best micro and macro average are achieved by using the first nine words as features. The question mark feature results in a micro average that is just barely above the baseline. It only results in the correct classification of the *statement-non-opinion* and *yes-no-question*. There is a similar problem with the classifier that uses the length feature: it only gets *statement-non-opinion* and *acknowledge* right, so its macro average is lower than the previous dialogue act feature, even though its micro average is higher. The first nine part-of-speech tags, n-grams and part-of-speech n-grams all result in comparable performance on the micro side, while the n-grams do a bit better on the macro side.

4.2.5 Combined Features

Figure 4.5 shows the recall for different combinations of features. The base classifier used only the first nine words as features, because those led to the highest accuracy on their own. The other features were added one by one and all improved the recall at least a little. The previous speaker and dialogue act features led to the largest increase: 4.5pp on the micro side and 10.7pp on the macro side. The question mark raised the recall by 0.6pp (micro) and 1.7pp (macro), and the previous speaker & dialogue act added 0.2pp (micro) and 1.4pp (macro). The first nine part-of-speech tags, length and part-of-speech n-grams all caused only a small increment of 0.1pp for micro and 0.9pp, 0pp and 0.3pp for macro respectively. The best recall is obtained by combining all seven types of features: a micro average of 78% (37.5pp above the baseline) and a macro average of 46.6%.

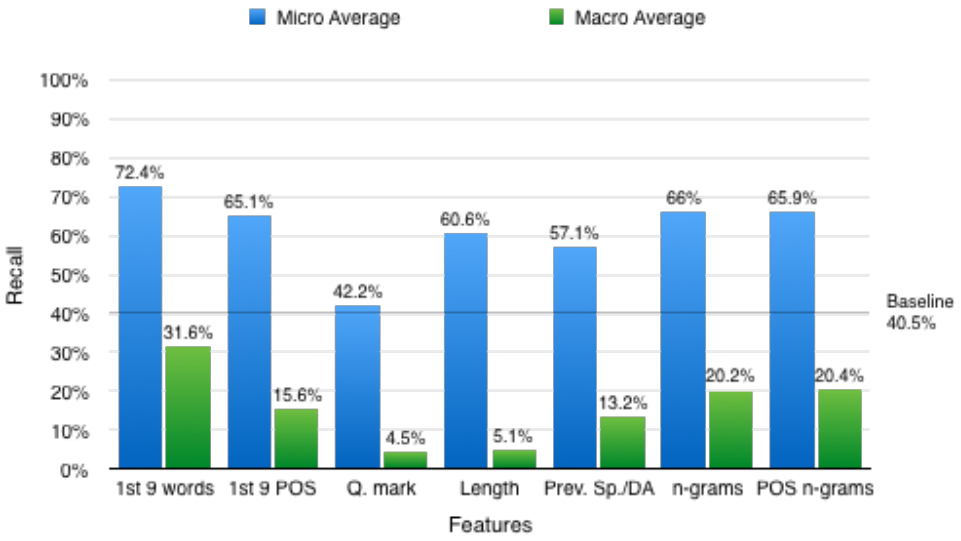


FIGURE 4.4: The average recall for each feature individually. POS = part-of-speech, q. mark = question mark, prev. sp./DA = previous speaker and dialogue act.

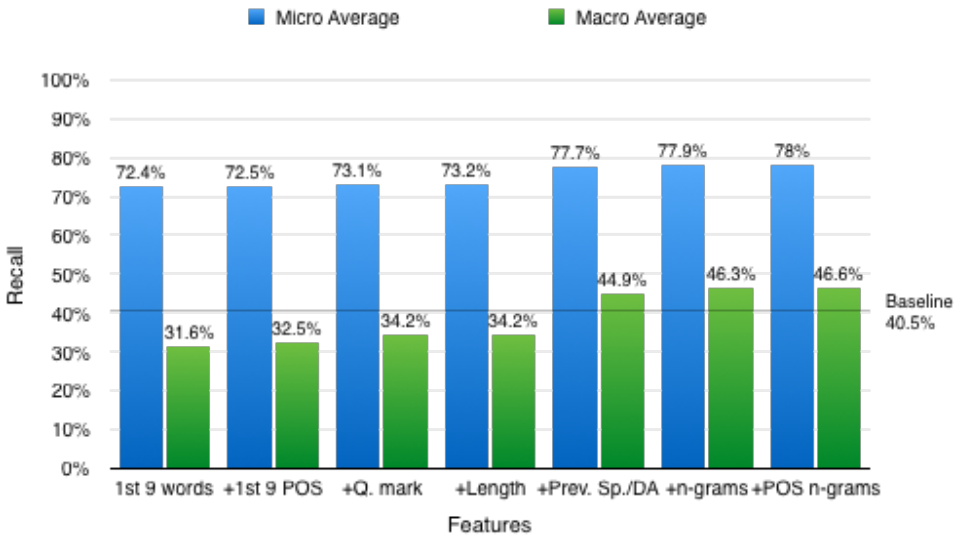


FIGURE 4.5: The average recall for combinations of features. The first nine words were used as the base and all other features were added on one by one. POS = part-of-speech, q. mark = question mark, prev. sp./DA = previous speaker and dialogue act.

4.2.6 Dialogue Acts

So far only the overall classification results have been shown, but it is also important to look at the performance of the individual dialogue acts. Table 4.5 contains the average precision, recall and F-score per class as well as the average number of utterances used for training and testing. The full confusion matrix can be found in appendix B.2. These results are from the classifier that used the full feature set, as it performed the best overall. Some dialogue acts were recognised better with a different setup, but in most cases the difference was less than 1pp, with the largest being 3.3pp.

There is a large amount of variety in the results of the individual classes. For some, such as *statement-non-opinion*, *acknowledge* and *conventional-closing* the recall is 90% or more, while others, such as *maybe*, *reject* and *summarize* have very few correct classifications. Generally, the classes that have many training samples available tend to get better results than those that have only a small amount, but this is not a hard rule.

The most common source of misclassifications is the majority class *statement-non-opinion*: utterances of 22 out of 39 classes incorrectly get that label over 10% of the time, often more. Others to a lesser extent are *statement-opinion* and *acknowledge* (both 7 out of 39), *yes-no-question* (6 out of 39) and *wh-question* (4 out of 39).

4.2.7 Top 5 Results

All results until now have only taken the best classification of each utterance into account. However, it can also be useful to look at some of the other high-ranking results, as these give an indication of how close the classifier managed to get to the correct answer. Table 4.4 compares the precision, recall and F-score of the regular situation to one that considers the top 5 (out of 39) highest ranking classifications for each utterance. If the right answer is in this top, the classification is counted as correct. Table 4.6 shows a breakdown by dialogue act.

Average	Precision	Recall	F-score
Macro	62.1%	46.6%	50.7%
Macro (top 5)	95.5%	82.3%	87.1%
Micro	76.2%	78.0%	76.3%
Micro (top 5)	97.9%	97.9%	97.8%

TABLE 4.4: A comparison of the micro and macro average precision, recall and F-score between the regular situation and one where a classification is considered correct as long as the dialogue act is in the top 5.

Overall, both micro and macro average precision and recall increase a lot when the whole top 5 is taken into account. The correct answer can be found among the top 5 97.9% of the time. Most individual dialogue acts see a big increase as well with over half having a precision over 80% and more than a quarter over 95%. Only four dialogue acts have a precision lower than 60%: *maybe/accept-part*, *downplayer*, *dispreferred answer* and *declarative wh-question*.

4.3 Discussion

Overall, combining all features together led to the highest accuracy: 78%, which is 37.5pp above the baseline. The macro average recall of only 46.6% suggests that many classes are still not recognised well, though. This is confirmed by the results for the individual dialogue acts: some are great, others very poor, with a lot in between as well. That said, the correct classification of an utterance can be found among the top 5 best results 97.9% of the time, which is promising.

TABLE 4.5: The average number of utterances available for training and testing as well as the average precision, recall and F-score per class.

Dialogue Act	Training	Testing	Precision	Recall	F-score
Collaborative completion	644	72	27.5%	8.4%	12.4%
Tag-question	83	9	65.7%	52.4%	57%
Hold before answer/agreement	498	55	75.1%	40.1%	52%
Quotation	878	98	62.3%	22.6%	33%
Agree/accept	9126	1014	67.3%	41.2%	51.1%
Action-directive	601	67	63.6%	29.6%	40%
Maybe/accept-part	93	10	5%	1%	1.7%
Reject	273	30	48.3%	6.6%	11.1%
Acknowledge (backchannel)	32568	3619	82.9%	96.1%	89.1%
Repeat-phrase	626	70	22.1%	7.1%	10.6%
Appreciation	4071	452	78.5%	69.5%	73.7%
Downplayer	86	10	62.9%	40.3%	47.9%
Summarize/reformulate	842	94	21.9%	2.4%	4.3%
Backchannel in question form	940	104	81%	68.1%	73.6%
Response acknowledgement	1129	125	67.2%	51.3%	58%
Signal-non-understanding	213	24	72.4%	65.3%	67.7%
Offers, options and commits	84	9	40.3%	14.4%	18.4%
Apology	68	8	78.6%	68%	69.2%
Conventional-closing	2164	240	96.1%	92.4%	94.2%
Conventional-opening	186	21	92.1%	83.2%	86.7%
Thanking	68	8	74.1%	58.6%	62.3%
Hedge	1096	122	80.7%	68.9%	74.2%
Affirmative non-yes answer	689	77	49.1%	41.3%	44.7%
Dispreferred answer	184	20	36.7%	5.5%	9.2%
Negative non-no answer	266	30	60.2%	35.2%	43.9%
No answer	1109	123	82.2%	86.8%	84.3%
Other answer	251	28	58%	21.5%	31.2%
Yes answer	2544	283	86.7%	76.4%	81.2%
Other	763	85	83%	60.6%	70%
Rhetorical-question	508	56	43.1%	14.9%	21.8%
Open-question	580	64	76%	64.5%	69.5%
Or-clause	178	20	79.6%	74.6%	76.4%
Wh-question	1703	189	76.4%	82.3%	79.2%
Declarative wh-question	76	8	10%	1.3%	2.2%
Yes-no-question	3622	402	80.3%	78.9%	79.6%
Declarative yes-no-question	1121	125	39.4%	12.6%	19%
Statement-non-opinion	63459	7051	80.3%	90%	84.8%
Statement-opinion	23147	2572	64.8%	57.4%	60.9%
Self-talk	92	10	50.4%	25.2%	30.9%

TABLE 4.6: The average number of utterances available for training and testing as well as the average precision, recall and F-score per class when a classification is counted as correct as long as the dialogue act is found in the top 5.

Dialogue Act	Training	Testing	Precision	Recall	F-score
Collaborative completion	644	72	93.2%	70.4%	80.1%
Tag-question	83	9	95%	85.1%	89.6%
Hold before answer/agreement	498	55	96.4%	73.6%	83.3%
Quotation	878	98	97.6%	76.5%	85.7%
Agree/accept	9126	1014	98.6%	98.4%	98.5%
Action-directive	601	67	95.7%	78.2%	85.9%
Maybe/accept-part	93	10	90.7%	34.3%	48.9%
Reject	273	30	97.3%	73.2%	83.4%
Acknowledge (backchannel)	32568	3619	99.1%	99.7%	99.4%
Repeat-phrase	626	70	92.1%	76.6%	83.7%
Appreciation	4071	452	97.3%	96.7%	97%
Downplayer	86	10	84.9%	56.5%	67%
Summarize/reformulate	842	94	97%	68.4%	80.1%
Backchannel in question form	940	104	98.3%	91.9%	94.9%
Response acknowledgement	1129	125	98.2%	95.6%	96.9%
Signal-non-understanding	213	24	95.2%	88.9%	91.8%
Offers, options and commits	84	9	95.8%	70%	80.1%
Apology	68	8	91.9%	88%	89.3%
Conventional-closing	2164	240	99.7%	97.2%	98.4%
Conventional-opening	186	21	97.1%	94.8%	95.8%
Thanking	68	8	87.9%	99.1%	92.8%
Hedge	1096	122	98.3%	88.3%	93%
Affirmative non-yes answer	689	77	90.2%	84.9%	87.3%
Dispreferred answer	184	20	99.2%	50.4%	66.4%
Negative non-no answer	266	30	94.1%	82.3%	87.7%
No answer	1109	123	96.7%	98.9%	97.7%
Other answer	251	28	96.3%	76.6%	85.1%
Yes answer	2544	283	99.3%	99.7%	99.5%
Other	763	85	98.3%	80.3%	88.3%
Rhetorical-question	508	56	95.2%	78.7%	86%
Open-question	580	64	97.6%	92.3%	94.8%
Or-clause	178	20	93.8%	97%	95.3%
Wh-question	1703	189	95%	98.2%	96.6%
Declarative wh-question	76	8	95%	27.3%	41.4%
Yes-no-question	3622	402	96%	96.4%	96.2%
Declarative yes-no-question	1121	125	97.2%	78.8%	87%
Statement-non-opinion	63459	7051	97.7%	99.9%	98.8%
Statement-opinion	23147	2572	97.6%	99.6%	98.6%
Self-talk	92	10	88.9%	68.4%	76.4%

4.3.1 The Best Value for n

The individual tests showed that when only the first n words or part-of-speech tags are used as features, a higher n leads to better overall results. The curve appears to flatten out around $n = 9$, though, and given its shape it seems likely that higher values of n will not increase the accuracy that much further. However, since n was only tested for values up to 10 this cannot be said with absolute certainty. One issue, though, is that nine words or part-of-speech tags may be quite a lot since the feature set extracted from an utterance becomes more unique with each additional feature. For example, when you only take the first five words of the following two sentences they appear to be equal, but if you take the first nine they are suddenly a bit more different:

What do you think about cycling?
What do you think about going to the movie tonight?

When the training samples are too specific it is harder for the classifier to generalise properly and overfitting may occur as a result. Overall this does not seem to be a problem yet at $n = 9$, since the performance of the classifier is still slowly increasing at that point. Certain individual dialogue acts could be affected by it, however, because many peak at a smaller n. In most of these cases the drop in recall remains limited to less than 1pp, but a few are hit harder, such as the *or-clause* which loses 7.5pp. Unfortunately, there is no single n which is best for all dialogue acts, so all further tests were simply done with $n = 9$ because it gave the highest results overall.

The situation is a lot more straightforward with the previous n speakers and dialogue acts: the best result is obtained for $n = 1$ and after that the accuracy slowly declines as n is incremented. Thus, only the previous speaker and dialogue act were used for further tests. While these features did poorly on their own, they gave a major boost to the classifier's accuracy when combined with the other features, showing the importance of contextual information for dialogue act recognition.

4.3.2 Question Mark & Length

The question mark and length features performed very poorly on their own, but that is to be expected since they are very limited: the question mark can only separate the dialogue acts into two groups (present/not present) and the length theoretically in many, but in practice also only in two (short/long). When combined with other features the question mark increases the accuracy a little, while the length does not have much effect. A possible reason for this is that utterance length is already implicitly present in the first nine word features, because blank values are used to fill up the remaining features when an utterance contains less than nine words.

4.3.3 N-grams

The automatically extracted n-grams led to a decent classifier with a micro average accuracy of 66% for the word n-grams and 65.9% for the part-of-speech n-grams, far above the baseline 40.5% and the second best results for the individual tests. The macro averages were only 20.2% and 20.4%

respectively, though, so many classes remained poorly recognised. Still, the results show that the method of automatically picking n-grams worked decently well.

One downside, though, is that there were *a lot* of n-grams: 400 in total. Since there is one feature per n-gram, the resulting feature vector is very large and as a result it takes a long time to train the classifier. Reducing the number of n-grams is a possibility, but that would probably lead to a decrease in accuracy. This issue makes n-grams a less attractive choice, because a much simpler classifier that uses just nine features managed to achieve far better results.

When combined with the other features the n-grams gave only a slight increase in accuracy. This could be because the first nine word and part-of-speech tag features already get better performance in almost all classes compared to the n-grams, so the n-grams do not have much to add. It might be worthwhile to focus the extraction process more on n-grams that do not occur as much at the start of the utterance, so they complement the other features instead of covering the same bases.

A possible improvement to the extraction algorithm that can reduce the number of n-grams is to filter out overlapping n-grams that fulfil the same purpose. For example, almost all possible n-grams of the sentence "It's been nice talking to you." as well as those of many variants such as "I enjoyed talking to you" ended up in the final set used as features. This leads to a lot of redundancy that may be preventable by only keeping the n-grams that represent the core of a phrase.

4.3.4 Issues

The results show that many dialogue acts were not recognised very well yet. Several reasons for this can be identified from the experiments. These include a lack of data, ambiguity, inconsistent annotations, long-range dependencies and inconsistent language use. There are also more general issues with the chosen approach to dialogue act recognition. At its core these are all caused by the fact that the method is too simplistic and hard to integrate with other components of conversational agents.

Lack of Data

Overall, having a larger set of data available for training and testing seems to improve the classification results. The dialogue acts whose test set contains more than a hundred utterances almost all have a recall over 50% with many being much higher than that. Likewise, most dialogue acts that get very poor results only occur a few times in the whole corpus, with some not even having ten utterances in their entire test set. Not only does this have the downside of making it harder to achieve good performance, it also makes the collected classification results much less reliable because they are a lot more vulnerable to outliers.

Gathering more data for such minority classes could be a first step towards improving their recognition rate. However, what if this is not feasible because it would be too costly and time-consuming? This problem does not just hit small, rare classes either, as many applications have domain-specific needs for which there also may not be a lot of data available. Unfortunately,

methods reliant on machine learning are very data-hungry, so there is not really an easy way around this issue besides using a different, less data-reliant approach.

Ambiguous Forms

Some dialogue acts tend to be expressed in the same way, making it difficult to distinguish them from each other. For example, the simple utterance "yeah" could be an *acknowledge*, *accept/agreement*, *yes answer* and more. In some cases it helps to add the previous speaker and dialogue act as features because that gives a bit more context. The *yes answer*, for instance, went up from a meagre 2% all the way to 78% because it is the most likely choice when the previous speaker asked a yes-no-question. Unfortunately, this is often simply not enough, so recognising certain dialogue acts may require more in-depth understanding of the conversation which the current system does not offer. Some dialogue acts that are especially affected by this problem are:

- **Agree/accept**: agreement and acceptance are often expressed in the same way as acknowledges. What differs is that the latter is just a signal that the speaker is listening to its conversational partner, while the former is generally an evaluation of some idea that was conveyed. According to the annotation manual the two can be told apart by examining the next utterance: agreements must always be followed by another utterance that indicates agreement such as "Me too" or "You're right", while acknowledges can occur on their own. This is unfortunate, because conversational agents cannot look into the future, so they have no way of knowing what the next utterance will be.

- **Response acknowledgement**: the *response acknowledgement* is an acknowledgement that follows a question-answer pair. It is often confused for regular *acknowledgements* because it frequently shares the same form. The characterising difference is that the *response acknowledgement* always has to be preceded by a question-answer sequence. Thus, the utterances that follow directly after an answer type dialogue act such as *yes answer* are generally classified correctly. However, answers can also come in the form of one or more statements, particularly when in response to a *wh-question*. In such situations it is not clear when or even if a question-answer sequence has ended, so the utterances are misclassified as regular *acknowledges*.

- **Signal-non-understanding**: non-understanding is often indicated with common phrases such as "Huh?", "What?" and "Excuse me?". Those cases are not the problem, however: the utterances may also take on the form of a question that more specifically shows what it is that the speaker does not understand (e.g., "You're now in what?", "Horses?", "What was the second one?"). These utterances tend to be misclassified as regular questions because the "non-understanding" is not explicitly present in their form.

- **Rhetorical-question**: the rhetorical question is a figure of speech and thus not meant to be interpreted literally. It mostly takes the form of

questions or statements and is usually confused for those types of dialogue acts. Recognition of figurative language use does not seem to have been researched much so far, but it is a very complex problem that involves deeper understanding of not just language, but also culture. Therefore, it seems unlikely that a simple statistical classifier can ever recognise them well.

- **Open-question**: utterances tagged as *open-questions* are often easy to recognise because they tend to contain something along the lines of "what/how about...", "what do you think", "how do you feel about...". The ones that do not, however, just look like any kind of *wh-question* and are frequently misclassified as such. The difference between the two is that the *wh-question* restricts the type of answer (e.g., "how old are you?") whereas the *open-question* places few if any constraints on it (e.g., "how do you feel about guns?"). Outside of the aforementioned key phrases this distinction is not always easy to make based on form alone.

- **Repetitions** The *repeat-phrase* dialogue act is literally a repetition of parts of the previous utterance, so it can essentially take on the form of any other dialogue act. Since the tested classifier did not use any features that represent this characteristic the accuracy for this dialogue act is very low. It should be possible to detect some degree of similarity between two utterances, though, so it might be possible to solve the issue by adding that as a feature.

- **Any dialogue act that shares the form of statements**: there are a couple dialogue acts that are virtually indistinguishable from statements based on form or surrounding dialogue acts alone. These are *collaborative completion*, *quotation*, *summarize/reformulate* and *dispreferred answer*. All four are expressed as statements and do not have distinct surface characteristics to set them apart. Instead they differ based on their function and semantic content. Several other dialogue acts suffer from the same problem but to a lesser extent: some of their utterances contain characteristics that make them distinct enough, others do not. Some examples are *hold before answer*, *declarative wh-* and *yes-no-questions*, *action-directive*, *maybe/accept-part*, *offers/options/commits* and *non-yes/non-no answers*.

Annotation Inconsistencies

This problem ties in with the previous one: the human-made annotations of the corpus are inconsistent from time to time, possibly because a good chunk of the utterances are somewhat ambiguous in form or function. In fact, the human annotators of the corpus only agreed with each other in 84% of the cases (Stolcke et al., 2000) despite the availability of an extensive annotation manual. For example, consider the following dialogue fragment from the corpus:

A:	you expect it to do its job,	(statement-opinion)
B:	Yeah.	(acknowledge)
A:	and I think a lot of car manufacturers don't take that into consideration you know.	(statement-opinion)
B:	Yeah,	(agree/accept)

Here, the first "yeah" is tagged as an *acknowledge*, while the second is tagged as an *agree/accept*. Both follow directly after a *statement-opinion*. Unsurprisingly, the classifier incorrectly assigned the *acknowledge* label to the second "yeah", as the difference between the two – if any – seems to be very subtle at best.

Likewise, the line between *wh-questions* and *open-questions* can be blurry too as the following examples taken from the corpus show:

So, what kind of music you into?	(wh-question)
What kind of vacations do you like?	(open-question)
What do you do with your budget?	(wh-question)
what do you do with your credit cards?	(open-question)
How do you like Chinese food?	(wh-question)
How did you like Africa?	(open-question)
And how about the person in, uh, Houston?	(wh-question)
How about your family?	(open-question)

The utterances in each pair are very similar to each other, yet one is tagged as a *wh-question* and the other as an *open-question*. The classifier understandably was not able to make the distinction and misclassified the first three *open-questions* as *wh-questions*. Likewise, the last *wh-question* was classified as an *open-question*.

Another inconsistently handled class is *thanking*. Almost all expressions of gratitude contain "thanks" or "thank you", so intuitively they should be easy to recognise. However, it is common to thank someone when closing the conversation and in that case the utterances are classified as *conventional-closings*. Or at least they should be, according to the manual, but there are many examples where this is not the case such as the following:

A:	Hey thanks a lot,	(thanking)
A:	I'll talk to you later.	(conventional-closing)
B:	All right,	(conventional-closing)
B:	thank you,	(thanking)

As a result, almost all misclassified *thanking* utterances occur during the closing sequence and are (incorrectly?) assigned the *conventional-closing* label.

Long-Range Dependencies

The tested classifier does not take long-range dependencies into account. Neither between words in an utterance, nor between the utterances themselves. As a consequence, many utterances that rely on the previous speaker and dialogue act to be properly recognised end up being misclassified. There are different ways in which long-range dependencies can happen. One is where the speaker starts a "subsequence" as in the following dialogue fragment:

A:	Are you finished with it now?	(yes-no-question)
B:	Uh, the roof?	(decl. yes-no-question)
A:	Uh-huh.	(yes answer)
B:	Oh yeah,	(yes answer)

Here, speaker A asks a question after which speaker B starts a subsequence by responding with another question to get some clarification. After speaker A responds the subsequence ends and speaker B finally gives his answer to the initial question. Keeping track of sequences and subsequences may help in scenarios like this one. Determining when to start or end a sequence and which utterance belongs to which sequence is not a trivial task, however, as many conversations do not follow a rigid structure. In the following fragment, for example, speaker B changes its mind and gives a second answer to the same question:

A:	Do you go up through the Raton Pass, when you go up there?	(yes-no-question)
B:	Uh, no,	(no answer)
B:	well, wait a minute,	(hold before answer)
B:	yeah,	(yes answer)

As this dialogue shows, a system that simply ends a question-answer sequence as soon as an answer has been received does not always work. Another problem is that two or more sequences can be intertwined as the next fragment demonstrates:

B:	How many stories?	(wh-question)
B:	Just one?	(yes-no-question)
A:	It's just one story.	(statement-non-opinion)
A:	Yeah.	(yes answer)

Here, the questions and answers are out of order, resulting in intertwined sequences. This likely happened because A and B spoke (almost) simultaneously. In both dialogues, the final *yes answer* was misclassified as an *acknowledge* by the classifier because their unconventional structures resulted in long-range dependencies.

A basic statistical approach to dealing with long-range dependencies could be to use multiple preceding dialogue acts as features instead of just a single one, but the results in section 4.2.3 show that this does not seem to improve the classification accuracy. An alternative option is to simply avoid complicated situations by having the conversational agent enforce a more

rigid dialogue structure. Such a system would be easy to break without some mechanism to detect violations of the expected structure, however. A more complex approach would involve the agent actively keeping track of the flow of the dialogue and any open obligations such as unanswered questions. This might require semantic and contextual information as well, though, because it is difficult to fully understand the dialogue structure based on dialogue acts alone.

Negations used Affirmatively

A rather counterintuitive problem is that words such as "no" do not always indicate a negative response like *reject* or *no answer*: they can be used positively as well depending on the context. The following dialogue fragment shows a common situation where this happens:

A:	It's not, like you sit and knit every night.	(reformulate)
B:	No.	(agree/accept)
B:	Not at all.	(agree/accept)

Here, person A expresses something in a negative form, after which person B signals agreement by responding with a negative as well, because two negatives make a positive. Such situations occur often enough in the corpus compared to actual *rejects* that the classifier associates the word "no" more with acceptance and agreement than rejection. As a result, the *reject* dialogue act is recognised very poorly. The *no answer* class is less affected by this problem as it usually follows after a *yes-no-question*, where this problem does not occur as much.

The easiest way to circumvent this issue is by simply limiting words such as "yes" and "no" to their conventional meaning and ignoring the edge cases. This should be fine for applications that are task-oriented or that simulate more formal conversations because these typically prefer language that is clear and not very ambiguous. Additionally, the use of "no" to indicate agreement is usually not required anyway. For example, in the above dialogue fragment the intentions of speaker B would have been just as clear if the response had been "yeah" instead of "no". Another way to avoid the problem is to design the dialogue system to never give a response that could trigger an ambiguous reply from the user in the first place. This would probably be the most user-friendly approach, but like the first method it is more suited to structured conversations than unrestricted, casual ones.

Pure Pattern Recognition is Insufficient

Several of the issues discussed so far point towards a major weakness of the tested approach: in the end, all a classifier can do is recognise patterns. Methods based on machine learning have their advantages as discussed in chapter 2.2, but they only work well for utterances with distinctive surface features that appear frequently enough in the data set. However, the analysis has shown that there are cases where higher-level knowledge or even reasoning are needed to distinguish the dialogue acts.

The non-yes/no answers to yes-no-questions are a prime example. The classifier has a lot of difficulty telling these apart from statements because most of the answers do not have any defining features as a group. The only thing they have in common besides looking like statements is their function: answers provide information that was requested by a closely preceding question. Shallow features are usually not enough to detect this characteristic however, because it tends to be hidden within the utterance's semantic content. For example, consider the following dialogue fragments:

B:	do you wear waders when you fish?	(yes-no-question)
A:	Uh, I probably ought to.	(negative non-no answer)
B:	It is in the morning?	(declarative yes-no-question)
A:	Uh, around nine or ten I think.	(affirmative non-yes answer)
A:	Well, do you watch much TV?	(yes-no-question)
B:	Well, I watch in the evening with my kids.	(affirmative non-yes answer)

The first fragment can only be interpreted properly if the meaning of "ought to" is known, as the answer does not contain any negations or other clues. The second at the very least requires knowledge about the meanings of "morning", "nine" and "ten" as well as the insight that the latter two refer to a time period in the morning. The last fragment complicates matters even further, because what is "much" is subjective. Thus, although the answer was marked "affirmative", this is really up to personal interpretation.

There are many other situations where the addition of high-level knowledge is advantageous or even necessary. One example are long-range dependencies such as questions not being answered until several utterances later. As section 4.2.3 has shown, simply adding more features with preceding dialogue acts only hurt the results. However, if the overall dialogue structure is actually known and understood to an extent it should be easier to deal with these cases.

Another example is that the way the dialogue participants interact with each other depends on the setting in which the conversation takes place as well as the social context that surrounds it. For instance, someone that is telling a story will primarily use statements, a narcissistic personality is unlikely to ever seriously apologise and a soldier will receive a lot of action directives from his superior while not using any himself. Such contextual information has the potential to be very useful for narrowing down the possible dialogue acts.

Theoretically, it is possible to correctly classify all of these cases with just pattern recognition as long as enough representative training data is available. However, in practice this is not going to work well. The problem is that even if such a data set exists, the classifier cannot generalise well from it. For example, the patterns it learns about language use surrounding the time of the day will not be of much use when the topic is about watching TV and vice versa. Therefore, outside of very narrow problem domains an impossibly large data set would be required to capture all the possible situations that might occur.

Despite these downsides, pattern recognition does have its place in dialogue act recognition systems. It should not be its sole component, however. Instead, a more scalable approach should use it as a specialised tool to extract useful information from utterances such as whether they have an interrogative or declarative form. If necessary, other valuable information can be supplied by different components such as a semantic role labeller or general knowledge base. The dialogue act can then be chosen based on the combination of the resulting data.

No Multi-label Classification

So far, the assumption has been that each utterance contains only one dialogue act, but in practice this view is too limiting. Consider the following sentence:

That's great!

In the Switchboard corpus such utterances are usually marked as *appreciation*, but there are also quite a few cases that were labeled as *statement-opinion*. Neither is really incorrect, since the sentence is both an expression of appreciation as well as an opinion. Merging these two dialogue acts does not make sense though, because both classes contain many utterances that only belong to one of the two such as the following:

Wow!	(appreciation)
I think black cars are better than blue cars.	(statement-opinion)

The first is not a statement but an expression of emotion, so it should not be marked as a *statement-opinion*. Likewise, the second is an opinion, but does not express appreciation for something said previously, so it cannot be labeled as an *appreciation*.

Ideally, it should thus be possible to assign multiple dialogue acts to a single utterance. One way to realise this with the current classifier is to simply pick all dialogue acts whose probability is above some threshold. There are two problems with this approach, however. The first is that it does not take the possible relations between dialogue acts into account. For example, both the *accept/agree* and *reject* might have a high probability for a "No" utterance, but because they are opposites only one of them can be correct.

The second is that it can be difficult to find a good threshold. One reason for this is that the probabilities are often heavily biased towards dialogue acts that appear more frequently in the corpus. As a result, the correct dialogue acts can still have a relatively low probability. For example, one "No" utterance was classified as follows:

Accept/agree:	38.2%
No answer:	21.2%
Acknowledge:	18.1%
Reject:	12.3%
...	...

The actual dialogue act in this particular case was *reject*, but it is ranked below three others that have a much higher frequency in the corpus. The *accept/agree* and *acknowledge* cannot be paired with the *reject* or *no answer* however, because their meanings contradict each other. Therefore, there is no threshold that would work well in situations such as this one.

It is not necessarily impossible to work around these problems, but it will add a lot of extra complexity. The question then is whether that is worth it, or if it would be better to design an alternative system that supports the assignment of multiple dialogue acts on its own. The first option is probably the quickest, but the second might be more maintainable in the long run.

Poor Support for dealing with Incomplete Information

In the literature, classifiers were often designed under the assumption that the conversation has already taken place and the full dialogue can be examined. However, this is not the case for conversational agents because they have to determine the dialogue act in the moment with only the dialogue history up to that point available. This presents a problem: what if there are multiple good candidates and there is simply not enough information to decide which one it should be? On the syntactic and semantic level this situation is quite common as even a simple sentence such as "He saw the girl with the binoculars." has multiple possible interpretations. It does not seem to occur as much on the pragmatic level, but there are certainly cases where it happens as well. One example is the *accept/agree* which, according to the SWDA annotation manual, often cannot be distinguished from the *acknowledge* without looking forward.

How should such an occurrence be solved, then? A simple solution is to just ignore it altogether and always go with the highest scoring dialogue act. Unfortunately, as mentioned previously the classifier is biased towards dialogue acts that appear more often in the data set. As a result the highest scoring one can often be wrong in ambiguous situations. If too many utterances are interpreted incorrectly because of that, the agent will not be able to function well, so this solution is not always acceptable.

Alternatively, an approach based on how humans handle communication might work. Psycholinguistic evidence suggests that when humans encounter ambiguous utterances, they do not actually bother resolving them until they absolutely have to to achieve their communicative goals (Ferreira and Patson, 2007). When applied to dialogue act recognition this would mean doing only a partial classification if there is not enough information to resolve all ambiguities. The agent will then have to work with a more general answer such as 'negative response' that corresponds to multiple dialogue acts. If, at some point, the agent really needs the full classification while it does not have all the required information, then a simple solution is to just ask the user for clarification.

Unfortunately, partial classification is not well supported by the classifier either because all it can do is rank the dialogue acts in order of likeliness. Just as it is difficult to determine from those results whether there are multiple correct labels, it is also difficult to determine whether there are multiple ambiguous labels. Therefore, if partial classification is a desired mechanism dialogue act recognition probably has to be handled in a different way.

Chapter 5

Alternative Approach

The analysis presented in chapter 4 has exposed multiple issues that cause the system to perform poorly at recognising many dialogue acts. To recap, most of them can be traced back to one of the following problems:

- **Ambiguity**: many dialogue acts overlap each other to some extent, which often makes them difficult to distinguish.
- **Inconsistent annotations**: there are many cases where two very similar utterances are labeled with different dialogue acts. As a result, it is difficult for the system to tell them apart.
- **Simplistic approach**: the system consists of a single classifier whose decision process is purely based on shallow features extracted from the input. In many situations this is not enough to make the correct choice.

The inconsistent annotations are likely caused by a combination of human error and the ambiguity between dialogue acts. Therefore, to fix this issue the problem of ambiguity has to be resolved first and then the annotations have to be adjusted accordingly. To reduce the workload and to prevent new human-caused errors from slipping in the second point should preferably be done with a more data-driven approach.

To resolve the ambiguity, the core of the issue needs to be examined first: why is there ambiguity between the dialogue acts at all? One possibility is that the boundaries between the categories are simply poorly defined with overlapping sections as a result. Another option is that some utterances contain more than one dialogue act, in which case the dialogue acts in question would not be strictly separable. It is likely that both reasons play a role to some extent. The first because the dialogue act taxonomy was designed based on the knowledge and intuition of human experts, not hard data. Given the complexity of language use it would not be very surprising if some errors got in. The second because there are a lot of utterances for whom multiple suitable dialogue acts can be found. A "yeah", for example, can double as both an agreement and an acknowledgement in certain situations and an appreciative utterance such as "That's great!" can be considered a type of opinion too.

Fixing the ambiguity involves two steps. First, the current boundaries between the dialogue acts have to be reexamined and adjusted where necessary. Second, utterances should be assigned multiple dialogue acts instead of just one if they are all a good fit. Both require modifications to the existing annotations, but as mentioned previously those have to be done anyway to reduce the number of inconsistencies. However, there is another

problem with the second step: as discussed in chapter 4.3.4, the classifier does not support multi-label classification very well. This ties in with the issue that the chosen approach is very simplistic. Several of the other difficulties the system is having suffer from this as well. They can likely be fixed by integrating the system better with other components of the conversational agent, but this is hard to do because it was not designed with those capabilities in mind.

To summarise, the following four tasks have to be dealt with to improve the system's performance:

- The dialogue act taxonomy has to be redefined.
- The data set has to be re-annotated.
- The system has to be able to assign multiple labels.
- The system has to integrate better with the agent's other components, so that it can make use of them to aid its decision process.

In addition, it would also be nice if the system can perform partial classifications as discussed in chapter 4.3.4 to deal with situations where not enough information is available to make an absolute decision. A different approach to dialogue act recognition is proposed that takes all of these things into account.

5.1 Proposed Approach

The proposed solution is to divide the general problem of dialogue act recognition into multiple smaller, more specific subproblems similar to how semantic role labelling is handled. However, instead of connecting these subproblems sequentially, they are organised in a hierarchical fashion with more general problems at the top and specific ones towards the bottom. The idea is to have a tree-like structure that gradually narrows down the possible dialogue acts for a given utterance. For example, a task at one of the higher level nodes could be to determine the utterance's sentence structure (declarative, interrogative, etc.). Based on the result it is then send to a subtree that deals with statement-like, question-like or other types of utterances. On the other hand, a task at a low level node could be something specific such as deciding if the utterance expresses agreement or acknowledgement. Utterances are then classified by moving from the root to one of the leaves, each of which corresponds to one or more dialogue acts.

This setup is better equipped to deal with several of the previously listed tasks. Firstly, it is easier for the system to make use of external resources such as the agent's knowledge base because it is no longer necessary to do everything in a single pass with a statistical approach. Instead, each subproblem can be handled in a different way that is specifically tailored to its needs. For example, determining the sentence structure might be done by a simple statistical method, whereas finding the difference between a regular statement and statement-like answer could be done by making use of the agent's knowledge and reasoning ability. An additional advantage is that the dialogue acts do not have to "compete" with each other as much for optimal results: an approach that is beneficial for

one but hinders another can still be used if the disadvantaged dialogue act is not relevant to the subproblem.

Secondly, assigning multiple labels to an utterance is better supported because each node, including the leaves, corresponds to a subset of related dialogue acts. All dialogue acts in such a set should be appropriate for the utterance at the given level of detail, so there is no need to try and filter some of them out with thresholds or other methods. Furthermore, since each node represents certain types of dialogue acts, the classification process can be halted at any point. When the resulting node is not a leaf, the output will simply be less specific. For example, if there is not enough information to determine whether a "Yeah" should be classified as an acknowledgement or an agreement, the system could stop at the current node and tag it with a more general label such as "positive response". Therefore, this setup also makes it easier to do partial classifications.

What remains, then, is modifying the dialogue act taxonomy and reannotating the data set accordingly. Both can actually be done simultaneously with the construction of the tree. The basic idea is to repeatedly partition the data set into smaller and smaller groups based on shared features. This results in a tree where each node corresponds to a single utterance subset such that the full set is found in the root and the smallest ones in the leaves. Whatever method was used in a particular node to further divide its subset can be kept to decide which child node any new utterances should be send to. Furthermore, dialogue acts can be assigned to the node based on the ones found most frequently in its utterance subset. The final result is a tree that can be used to classify utterances. Its leaves represent the new dialogue act taxonomy and the the utterances in the data set can be relabelled with the leaf they ended up in. The next section describes the approach in more detail.

5.2 Methodology

The process to construct the tree involves three components: clustering methods to divide the data set, termination criteria to decide when the tree should stop expanding further and a method to assign dialogue acts to each node. These are discussed in sections 5.2.1, 5.2.2 and 5.2.3 respectively. Next, section 5.2.4 gives a step by step overview of the construction algorithm and finally section 5.2.5 discusses some implementation details.

5.2.1 Clustering

The most important part of the system is the clustering process through which the tree is constructed. The simplest approach would be to use the same method for all nodes, but it is also possible to give each node its own clustering algorithm. The latter option has two advantages. First, it gives more control over the form of the tree as different methods and parameters may result in different partitions. Second, better results can be achieved by fine-tuning the algorithm to the characteristics of the subset it has to divide. The downside, though, is that manually selecting which algorithm to use where is a bit more time consuming.

In the chosen setup, each node was given its own clustering algorithm. Most utilised a classifier derived from the ones tested in chapter 4. A few others used simple, manually defined rules instead because statistical methods did not result in the desired partitions. Although classifiers are not normally used to cluster data, they work out in this situation because the already existing annotations can guide the system. In this case the clustering process consists of two stages. In the first, the classifier is trained using the labeled utterances. In the second all utterances in the subset are assigned to a cluster by the classifier.

In some situations utterances belonging to a specific dialogue act may be excluded or merged with another group during the training stage to improve the results. Merging related or very similar dialogue acts such as the declarative and regular yes-no-question ensures that they end up in the same partition. If desired they can then be divided further by targeting the specific differences between them. Excluding a dialogue act is useful if it should not be in a subset of its own in the intended subdivision. This is particularly helpful when a specific dialogue act needs to be split up over several nodes because it contains multiple different types of utterances.

An advantage of using classifiers over more conventional clustering algorithms such as k-means is that there is no need to define the number of clusters in advance. The maximum is equal to the number of class labels in the data set, but if a classifier cannot distinguish between two classes during the training stage, it will merge them into one.

5.2.2 Termination

The utterances cannot be partitioned into ever smaller subsets, so the process has to terminate at some point. Some simple methods would be to define a maximum depth or minimum subset size, but on their own those are unlikely to lead to the best results. Two more useful conditions to check for are the following:

- The clustering algorithm fails to divide the subset any further.
- The clustering algorithm does not find any new useful or meaningful clusters.

The former can be handled automatically, but the latter probably has to be judged by a human supervisor because what is "useful" or "meaningful" is subjective and depends on the problem domain. One application may be fine with a coarse grouping of utterances, for example, while another needs much more specific and precise results.

5.2.3 Assigning dialogue acts

After the tree has been constructed, one or more dialogue acts need to be assigned to each node. A simple way to do this automatically is to check how often utterances of each dialogue act are present in the node's subset relative to its size and then just choosing all dialogue acts that occur more often than some threshold. The downside of this method is that very small classes may be drowned out by much larger ones, but other than that it seemed to work out fairly well. Another approach would be to have a

human label the leaves, but this is obviously more time consuming. A combination of the two is possible as well where the automatically generated results are used as a base and then modified manually where necessary.

5.2.4 Algorithm

To summarise, the tree is constructed by the following algorithm:

1. Choose which clustering algorithm to use for the current node with what parameters.
2. If a classifier is used, train it using the current utterance subset. Depending on the desired partitioning, some utterances may be excluded from the training process or merged with another dialogue act group.
3. Assign each utterance in the subset to a cluster with the clustering algorithm.
4. For each cluster:
   (a) Create a new child node.
   (b) Pass all utterances in the cluster to the new node.
   (c) Determine which dialogue acts should be associated with this node. Optional for internal nodes.
   (d) Repeat from step 1 unless the termination condition has been fulfilled

5.2.5 Implementation Details

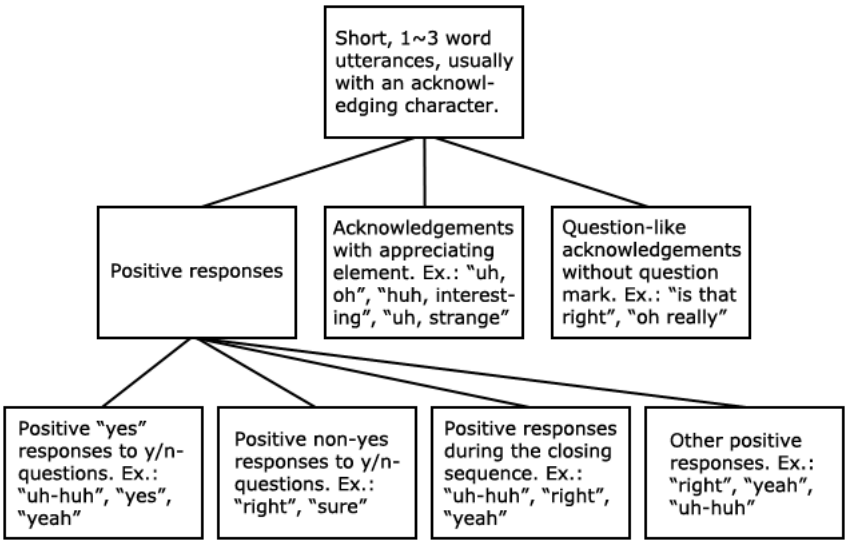
Each node was manually assigned a clustering method. In most cases a classifier sufficed, but occasionally a simple rule had to be used instead because the statistical approach failed to give the desired partition. Not all utterances in the node's subset were always used to train the classifier: often certain dialogue acts were merged with others or completely excluded. Appendix C contains the full list of nodes and corresponding parameters for the resulting tree.

Termination of the process was decided manually as well according to the following three criteria.

- No new clusters were found by the clustering algorithm. This can sometimes be remedied by changing some parameters or switching to a rule-based method, but other times it is simply difficult to partition the current subset any further.
- The resulting partition was undesirable. The classifier can often find at least some subdivision, but not all of them are great. If the new groups do not look sufficiently distinct, for example, they could be the result of overfitting.
- Only one dialogue act was left in the cluster, so there was no reason to continue dividing the group any further.

After the tree has been constructed some additional modifications may be made such as assigning dialogue acts and pruning some nodes. dialogue acts were chosen for each leaf according to the method described in section 5.2.3 with a threshold of 5%. This threshold was somewhat arbitrarily chosen: 10% felt a tad too strict while 1% was a bit too forgiving. The only nodes that were pruned were the ones that had less than ten utterances in their subset. These small groups were typically the result of overfitting and therefore not very useful.

FIGURE 5.1: A small brach of the tree that focuses on acknowledgement-like utterances.



5.3 Results

The final tree had a height of 4, 27 internal nodes and 104 leaves. Many, but not all of the leaves were associated with more than one dialogue act. In most cases just two, but occasionally up to five. As an example, figure 5.1 shows a small part of the tree that deals with acknowledgement-like utterances. Details on the full tree can be found in appendix C.

Table 5.1 shows the recall of the tree when a classification is counted as correct as long as the actual Switchboard dialogue act of the utterance is associated with the resulting leaf. The micro average is 92.7%, but the macro is brought down to 57.1% by some dialogue acts with poor to mediocre results. This method does not check if an utterance actually ends up in the intended leaf, however, which is what the next results are focused on.

The recall of the tree when utterances need to be assigned to the right leaf for the classification to be correct can be found in table 5.2. In this case, the macro average recall is 80.3% while the micro average is 90.5%. The *declarative wh-question* has the poorest recall with only 41.9% followed by the *dispreferred answer* at 43.2%. All other classes have a recall above 50% with the majority above 80%.

Instead of using the tree to classify the utterances it is also possible to train a regular classifier on the utterances relabelled with the leaf they were assigned to. The recall for that setup is shown in table 5.3. Both micro and

TABLE 5.1: The average recall of the tree classifier when using the existing dialogue acts associated with the leaves as classifications.

Dialogue Act	Recall	Dialogue Act	Recall
Collaborative completion	14.5%	Thanking	96.2%
Tag-question	51.5%	Hedge	73.7%
Hold before answer/agreement	49.1%	Affirmative non-yes answer	61.6%
Quotation	0.3%	Dispreferred answer	21.7%
Agree/accept	91.0%	Negative non-no answer	56.7%
Action-directive	33.9%	No answer	96.7%
Maybe/accept-part	0.6%	Other answer	46.5%
Reject	60.3%	Yes answer	87.0%
Acknowledge (backchannel)	98.9%	Other	60.3%
Repeat-phrase	26.4%	Rhetorical-question	25.2%
Appreciation	84.9%	Open-question	62.2%
Downplayer	31.6%	Or-clause	61.7%
Summarize/reformulate	11.6%	Wh-question	84.8%
Backchannel in question form	82.6%	Declarative wh-question	7.4%
Response acknowledgement	88.9%	Yes-no-question	89.3%
Signal-non-understanding	60.8%	Declarative yes-no-question	39.0%
Offers, options and commits	9.5%	Statement-non-opinion	98.6%
Apology	71.0%	Statement-opinion	97.2%
Conventional-closing	91.2%	Self-talk	20.0%
Conventional-opening	80.8%		
Macro average	57.1%	Micro average	92.7%

macro average, 86.3% and 73.4% respectively, are lower than the results obtained by the tree. The same is true for the recall of each class individually. The tree took 5.7ms on average to classify a single utterance, however, which is approximately 2.4x slower than the regular classifier which only needed 2.4ms.

5.4 Limitations

There are several limitations and issues with the current setup. Firstly, the resulting tree is more of a rough draft used as a proof of concept than a finished product. Longer sentences are currently the most problematic. Short utterances can often be divided into groups fairly well because they are recognisable thanks to specific keywords. It is much harder to find the right clusters for longer sentences, though, as those vary a lot more in form. Other clustering methods may be required to find better subdivisions for this group.

Secondly, clustering utterances with classifiers can only be done if there are preexisting annotations already because without those the classifiers cannot be trained. Therefore, a new data set without any annotations must be partitioned using a different method. Alternatively, it might be possible to annotate only a small amount of new data and use that to iteratively train the classifiers by having each iteration take the automatic annotations

TABLE 5.2: The average recall of the tree classifier when the leaves themselves are used as classifications.

Dialogue Act	Recall	Dialogue Act	Recall
Collaborative completion	82.4%	Thanking	99.1%
Tag-question	89.5%	Hedge	85.1%
Hold before answer/agreement	69.0%	Affirmative non-yes answer	78.4%
Quotation	90.0%	Dispreferred answer	43.2%
Agree/accept	95.8%	Negative non-no answer	67.2%
Action-directive	52.4%	No answer	98.6%
Maybe/accept-part	54.5%	Other answer	60.1%
Reject	82.3%	Yes answer	99.3%
Acknowledge (backchannel)	98.8%	Other	91.3%
Repeat-phrase	87.0%	Rhetorical-question	86.7%
Appreciation	90.9%	Open-question	78.6%
Downplayer	91.2%	Or-clause	66.1%
Summarize/reformulate	85.4%	Wh-question	85.9%
Backchannel in question form	93.7%	Declarative wh-question	41.9%
Response acknowledgement	98.7%	Yes-no-question	89.2%
Signal-non-understanding	78.4%	Declarative yes-no-question	59.4%
Offers, options and commits	65.9%	Statement-non-opinion	91.4%
Apology	85.2%	Statement-opinion	79.1%
Conventional-closing	93.7%	Self-talk	57.6%
Conventional-opening	90.4%		
Macro average	**80.3%**	**Micro average**	**90.5%**

from the previous one into account. This way not *all* of the new utterances would have to be manually annotated, but only a subset. The viability of this approach has not been tested, however.

Thirdly, there is a risk that dialogue acts with only a small number of utterances end up divided and spread too thin across multiple nodes. If a situation like that occurs, certain dialogue acts might not have enough utterances left in a single node to effectively train the classifier with. This problem was encountered once with the *conventional-opening*: a couple "How are you doing?" utterances were grouped together with *wh-questions*, but were not numerous enough to be distinguished from them.

Fourthly, the method that is currently used to automatically assign dialogue acts to a node is very simple and not without problems. The main issue is caused by the fact that dialogue acts are chosen based on the percentage of utterances they have in the node's subset. The result of this approach is that a dialogue act with a small amount of utterances can be completely drowned out by a larger class through sheer numbers despite being a viable choice for the node. This can be resolved by using a more complex algorithm or simply having a human judge the chosen dialogue acts and make modifications where necessary.

Finally, despite being a more data driven approach there is still quite a bit of human involvement in the process. This can be reduced by having the system automatically select which clustering methods and parameters to use, but that does not necessarily lead to better results. Either way, a

TABLE 5.3: The average recall of a regular classifier that uses the leaves of the tree as dialogue act labels for the utterances.

Dialogue Act	Recall	Dialogue Act	Recall
Collaborative completion	64.8%	Thanking	94.2%
Tag-question	78.5%	Hedge	80.2%
Hold before answer/agreement	64.9%	Affirmative non-yes answer	63.7%
Quotation	78.3%	Dispreferred answer	37.1%
Agree/accept	92.4%	Negative non-no answer	55.3%
Action-directive	42.9%	No answer	98.8%
Maybe/accept-part	45.1%	Other answer	55.4%
Reject	76.9%	Yes answer	89.3%
Acknowledge (backchannel)	96.5%	Other	86.3%
Repeat-phrase	74.6%	Rhetorical-question	77.5%
Appreciation	87.9%	Open-question	74.2%
Downplayer	81.1%	Or-clause	83.5%
Summarize/reformulate	63.8%	Wh-question	86.9%
Backchannel in question form	92.0%	Declarative wh-question	33.5%
Response acknowledgement	98.0%	Yes-no-question	87.5%
Signal-non-understanding	77.8%	Declarative yes-no-question	48.6%
Offers, options and commits	43.4%	Statement-non-opinion	88.1%
Apology	79.8%	Statement-opinion	71.5%
Conventional-closing	83.8%	Self-talk	39.5%
Conventional-opening	88.7%		
Macro average	73.4%	Micro average	86.3%

human is still needed to judge the quality of the final nodes and to give them an appropriate label.

5.5 Discussion

On average, over 92% of the utterances are classified correctly by the tree when the dialogue acts corresponding to the resulting leaf are used as classification. The macro average is still pretty mediocre, however, because many of the smaller classes do not score well. This could be because they are truly not classified correctly, but it is also possible that the method used to automatically assign dialogue acts did not work well in some instances. The threshold might have been too strict, for example, or small classes may have been drowned out by large ones.

The performance was therefore also measured by using the leaves themselves as classifications instead of the assigned dialogue acts. The micro average recall slightly decreased to 90.5% under these circumstances, but the macro average increased by a lot to 80.3%. The lower micro average can largely be explained by the poorer recall of the two statement dialogue acts. Since the statements tend to have a less well defined form than many other dialogue acts, they are more likely to end up in the wrong leaf. Despite that, they would often still be classified correctly by the previous method because many leaves are associated with at least one of the two statements.

The higher macro average shows that most utterances do get assigned to the correct leaf. The difference with the previous method is thus an indication that the dialogue act assignment algorithm may indeed have some issues with certain classes. Most likely the threshold is too strict for some which causes them to be excluded disproportionately often. Lowering the threshold would improve this situation a bit, but is also likely to result in an increase in undesirable dialogue act assignments. To resolve this problem, a different method may be needed altogether.

Compared to a regular classifier that also uses the leafs of the tree as classifications, the tree is a bit more accurate. This shows that there is merit in using the tree itself for the classification task and not just as a method to relabel the utterances. The better performance likely has to do with the fact that each classifier used by the tree can be tailored to the specific characteristics of the subset it needs to divide. There is still room for improvement in that area, though, as the current tree is more of a rough draft and probably nowhere near optimal.

On the whole, the current tree is shallow and broad: it has 104 leaves, but only a height of 4. Since there are over 2.5 times as many leaves as the number of dialogue acts in the SWDA taxonomy, the grouping of utterances created by the tree is probably more specific in many cases than the original one. This is especially noticeable with some of the larger classes such as *acknowledge*. This particular dialogue act contains a large variety of short utterances that function as acknowledges ranging from "yeah" to "oh, I see" to "really". The tree splits these up into distinct groups, which makes sense as they are all used in slightly different ways: "yeah" carries a neutral to positive load and may signal agreement, "oh I see" is a neutral response primarily used after answers and "really" functions as a rhetorical question.

Such specificity might not always be necessary, however. For example, for a conversational agent to function properly the distinction between a "yeah" used to signal agreement and a more wordy "I agree with that" may not be all that important. On the other hand, there could be applications that do benefit from it. A "yeah", for instance, might be seen as more informal and therefore inappropriate in certain scenarios. Using the tree as a classifier offers a clear advantage here over a regular one as it can easily be adapted to the requirements of specific applications by simply adding or removing nodes.

The tree does have some downsides, however. Compared to a regular classifier it is more time consuming to set up and design, and is slower at classifying utterances. Specifically, the higher the tree is, the longer the classification process takes as every node on the path to a leaf adds one additional classifier. The trade-off for this extra computation time is a higher classification accuracy, though, so it will really depend on the application whether the tree is a suitable choice.

Chapter 6

Future Work

There is still a lot of room left for improvement when it comes to dialogue act recognition. Issues with the data set such as ambiguity between dialogue acts were handled pretty well by re-partitioning and labelling the utterances with the proposed tree system. However, the resulting tree is not perfect and needs to be improved and optimised in certain areas. In particular, statement-like utterances were often difficult to divide into groups because their differences were not captured well by the chosen feature set. This shortcoming is partly caused by the fact that only shallow methods were used to create the subdivisions.

Better results might be achievable by integrating the system with an actual conversational agent so that more advanced techniques can be used. Social practices (Dignum and Dignum, 2015) are one example. The social context has a lot of influence on how a conversation progresses, so it can help narrow down the possible dialogue acts. Early results on this topic have shown promise so far. Integrating elements such as social practices with the tree should be fairly straightforward, as it only requires the modification or addition of nodes. Further integration with an agent also allows the usefulness of partial classifications to be tested. In theory this should be a beneficial functionality, but it has not been tested in practice yet.

Another area of research could be to determine how useful the tree is when dealing with data from a different domain. This includes how accurate it is when used as a classifier, but also how it can be adapted to a new domain. One way, for example, could be to just modify the existing nodes of the tree with the new data. However, from a modularity point of view it might be more desirable to keep the general and domain-specific data separated from each other as much as possible. Whether that is feasible will have to be seen, though.

Additionally, a different problem of dealing with a new domain is that new data may have to be annotated, which is a time-consuming task. The tree can speed up this process, but the specific form tested so far relies on existing annotations, so it would not be useable in this case. Besides constructing a tree manually, it might also be possible without those annotations by switching to a different clustering method. Alternatively, maybe it could be build iteratively with just a small amount of annotated data as a starting point. The initial tree would likely be far from perfect, but still useable for annotating some of the remaining data. Any subsequent iterations can then use the annotations from the previous one to attempt to refine the tree further. Given how laborious manually annotating a lot of data is, it could be worth the time to investigate such alternative methods to automate (parts of) the process.

Chapter 7

Conclusion

All tested features improved the classification accuracy, but some had a bigger positive effect than others. Individually, using the first nine words as features achieved the best results. Adding the previous speaker and dialogue act on to that led to a major improvement, while the question mark feature gave a more modest increase. The addition of the remaining features did result in a higher classification accuracy, but only by a small amount. The combination of all features resulted in an accuracy of 78%, which is fairly close to the performance of humans (84%).

There is a lot of variety in how well the different dialogue acts were recognised, however. Some had a near perfect score, others received mediocre results and several were rarely even classified correctly at all. The classification accuracy was often higher for dialogue acts that occurred more frequently in the corpus, but this was not an absolute rule. Other causes of recognition issues include ambiguity between dialogue acts, inconsistent language use, differences between certain dialogue acts not being accurately captured by the feature set, inconsistent annotations and long-range dependencies.

To tackle some of these issues a different approach to dialogue act recognition was proposed where a hierarchically organised group of classifiers was used instead of just a single one. This tree-like structure offers multiple advantages. It is easier to integrate with different components of the conversational agent, supports multi-label classification and is more capable of performing partial classifications. In addition, it can be used to semi-automatically redefine the dialogue act taxonomy and relabel the data set, which was necessary to resolve the problems with ambiguity. This resulted in over a hundred utterances subsets that each corresponded to one or more dialogue acts. The classification accuracy on this new data set ranged from 86.3% to 92.7% depending on the classification method that was used. This is much better than the results for the old data set, but because of the change in methodology the two are not directly comparable.

Appendix A

Corpus and Dialogue Acts

TABLE A.1: Each dialogue act and its corresponding tag.

Dialogue Act	**Tag**
Collaborative completion	^2
Tag-question	^g
Hold before answer/agreement	^h
Quotation	^q
Agree/accept	aa
Action-directive	ad
Maybe/accept-part	am
Reject	ar
Acknowledge (backchannel)	b
Repeat-phrase	b^m
Appreciation	ba
Downplayer	bd
Summarize/reformulate	bf
Backchannel in question form	bh
Response acknowledgement	bk
Signal-non-understanding	br
Offers, options and commits	co
Apology	fa
Conventional-closing	fc
Conventional-opening	fp
Thanking	ft
Hedge	h
Affirmative non-yes answers	na
Dispreferred answers	nd
Negative non-no answers	ng
No answers	nn
Other answers	no
Yes answers	ny
Other	o
Rhetorical-questions	qh
Open-questions	qo
Or-clause	qrr
Wh-questions	qw
Declarative wh-questions	qw^d
Yes-no-questions	qy
Declarative yes-no-questions	qy^d
Statement-non-opinion	sd
Statement-opinion	sv
Self-talk	t1

TABLE A.2: The total number of utterances per dialogue act.

Dialogue Act	**Utterances**	**%**
Collaborative completion	716	0.41%
Tag-question	92	0.05%
Hold before answer/agreement	553	0.32%
Quotation	975	0.56%
Agree/accept	10140	5.83%
Action-directive	668	0.38%
Maybe/accept-part	103	0.06%
Reject	303	0.17%
Acknowledge (backchannel)	36187	20.79%
Repeat-phrase	695	0.4%
Appreciation	4523	2.6%
Downplayer	96	0.06%
Summarize/reformulate	936	0.54%
Backchannel in question form	1044	0.6%
Response acknowledgement	1254	0.72%
Signal-non-understanding	237	0.14%
Offers, options and commits	93	0.05%
Apology	75	0.04%
Conventional-closing	2404	1.38%
Conventional-opening	207	0.12%
Thanking	76	0.04%
Hedge	1218	0.7%
Affirmative non-yes answer	766	0.44%
Dispreferred answer	204	0.12%
Negative non-no answer	295	0.17%
No answer	1232	0.71%
Other answer	279	0.16%
Yes answer	2827	1.62%
Other	848	0.49%
Rhetorical-question	564	0.32%
Open-question	644	0.37%
Or-clause	198	0.11%
Wh-question	1892	1.09%
Declarative wh-question	84	0.05%
Yes-no-question	4024	2.31%
Declarative yes-no-question	1245	0.72%
Statement-non-opinion	70510	40.52%
Statement-opinion	25719	14.78%
Self-talk	102	0.06%
Total	174028	100%

TABLE A.3: The number of utterances per dialogue act in each fold.

DA \ Fold	1	2	3	4	5	6	7	8	9	10
Collaborative completion	67	86	83	77	70	81	71	60	57	64
Tag-question	9	10	7	10	3	6	15	7	11	14
Hold before answer/agree-ment	71	51	57	50	48	44	61	55	61	55
Quotation	127	106	99	81	91	112	96	106	75	82
Agree/accept	1010	941	1113	1035	932	919	1116	1007	1047	1020
Action-directive	61	54	57	73	55	63	102	60	87	56
Maybe/ accept-part	10	17	10	14	6	10	8	5	10	13
Reject	35	29	30	34	34	20	40	25	29	27
Acknowledge (backchannel)	3761	3547	3925	3583	3493	3622	3733	3596	3390	3537
Repeat-phrase	70	73	87	76	74	55	66	56	88	50
Appreciation	476	442	483	488	403	446	470	423	417	475
Downplayer	11	9	7	11	12	1	15	6	14	10
Summarize/ reformulate	98	94	90	108	94	105	77	89	95	86
Backchannel in question form	107	104	134	95	114	101	96	94	104	95
Response acknowledge-ment	132	119	118	101	149	118	154	101	131	131
Signal-non-understanding	23	17	23	28	22	29	18	20	31	26
Offers, options and commits	2	9	7	17	6	9	18	12	8	5
Apology	7	10	9	7	8	4	10	5	10	5
Conventional-closing	281	137	276	271	231	229	233	220	232	294
Conventional-opening	25	10	22	22	19	21	16	47	13	12
Thanking	9	5	6	9	7	8	7	8	11	6
Hedge	120	116	127	132	126	112	120	103	134	128
Affirmative non-yes answer	76	67	74	75	81	86	88	67	70	82
Dispreferred answer	21	27	18	22	25	11	28	15	19	18
Negative non-no answer	15	28	33	32	41	20	30	24	35	37
No answer	133	111	98	116	162	125	125	119	112	131
Other answer	26	27	28	31	37	22	26	20	29	33
Yes answer	265	282	317	271	318	254	299	255	292	274
Other	89	67	77	95	91	96	84	88	79	82
Rhetorical-question	65	50	47	58	50	44	53	65	64	68
Open-question	68	62	83	67	65	45	74	62	56	62
Or-clause	16	17	20	20	23	21	31	9	20	21
Wh-question	200	206	178	189	177	184	177	160	211	210
Declarative wh-question	7	11	10	7	6	8	7	7	8	13
Yes-no-question	382	377	391	391	445	378	443	352	415	450
Declarative yes-no-question	140	129	110	114	115	105	144	126	127	135
Statement-non-opinion	6961	7091	7164	6901	7322	6932	7264	6450	7075	7350
Statement-opinion	2916	2342	2700	2544	2309	2466	2650	2554	2631	2607
Self-talk	8	22	9	15	5	3	9	10	11	10

TABLE A.4: An overview of the part-of-speech tags.

Tag	**Description**
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Appendix B

Analysis Results

B.1 Average Recall

The following tables show the average recall per dialogue act for different classifiers. The blank cells represent 0% and the numbers in bold are the best values.

TABLE B.1: Average recall per dialogue act with the first n words as features for each $n \in [1, 10]$.

Dialogue Act	1	2	3	4	5	6	7	8	9	10
Collaborative completion		0.3%	3.1%	3.8%	4.4%	4.9%	5.3%	5.3%	5.2%	**5.4%**
Tag-question				**16.2%**	14.2%	14.9%	14.2%	14.2%	14.2%	14.2%
Hold before answer/agreement	15.9%	31.5%	31.8%	33.5%	35%	36%	36.2%	36.9%	**37.4%**	**37.4%**
Quotation			0.7%	0.7%	1.2%	**1.3%**	1.1%	1.2%	**1.3%**	**1.3%**
Agree/accept	8.9%	21.1%	24.8%	27.7%	28.5%	**29.3%**	28.8%	28.8%	28.9%	28.9%
Action-directive	12.4%	15.9%	20%	21.7%	22%	23%	23.1%	22.9%	**24.1%**	23.5%
Maybe/accept-part										
Reject				**0.3%**	**0.3%**	**0.3%**	**0.3%**	**0.3%**	**0.3%**	**0.3%**
Acknowledge (backchannel)	94.7%	95.8%	**96.5%**	**96.5%**	**96.5%**	96.3%	**96.5%**	**96.5%**	**96.5%**	**96.5%**
Repeat-phrase		1%	**3.4%**	3.1%	3.1%	2.8%	2.9%	2.9%	3%	3%
Appreciation	12.8%	47.8%	63.3%	66.3%	67.6%	68%	68.4%	**68.6%**	**68.6%**	**68.6%**
Downplayer			26.3%	**27%**	**27%**	24%	**27%**	**27%**	**27%**	**27%**
Summarize/reformulate		0.1%	1%	0.8%	1.1%	**1.2%**	1.1%	1%	**1.2%**	**1.2%**
Backchannel in question form	16.6%	37.2%	43.6%	52.7%	54.8%	56.9%	55.7%	55.7%	**57.4%**	**57.4%**
Response acknowledgement		26.7%	31.7%	31.4%	31.6%	**32.1%**	31.6%	31.6%	31.6%	31.6%
Signal-non-understanding	15.3%	17.3%	26.3%	27.7%	29.1%	**31.2%**	30.5%	30.5%	30.5%	30.5%
Offers, options and commits			11.8%	14.3%	13.8%	**14.6%**	12.6%	12.1%	12.6%	12.6%
Apology	19.6%	19.6%	58.7%	56.1%	62.9%	**64.9%**	61.9%	61.9%	61.9%	61.9%
Conventional-closing	31.4%	38%	42.1%	48%	51.5%	51.2%	54%	54.2%	54.3%	**54.4%**
Conventional-opening	58.6%	52.9%	**58.8%**	58.4%	58.6%	57.4%	58.6%	58.6%	58.6%	58.6%
Thanking		3.1%	2%	**6.6%**	5.1%	5.1%	5.1%	5.1%	5.1%	5.1%
Hedge					36%	56.5%	67.4%	68.9%	**69.1%**	**69.1%**
Affirmative non-yes answer		0.2%	0.3%	**1.4%**	1.2%	1.3%	1.2%	1.2%	1.2%	1.2%
Dispreferred answer										
Negative non-no answer		10.5%	13.5%	15.1%	15.1%	14.6%	15.6%	15.9%	15.6%	**16.2%**
No answer	76.2%	88.4%	88.6%	89.1%	89%	**89.4%**	88.9%	88.8%	88.9%	88.8%
Other answer										
Yes answer		**4.6%**	2.8%	2.8%	2.5%	2.7%	2.5%	2.5%	2.5%	2.5%
Other		5.5%	5.4%	5.5%	5.6%	**6.4%**	5.4%	5.4%	5.5%	5.3%
Rhetorical-question		3.5%	6%	7.7%	8.5%	**9.8%**	8.7%	8.4%	8%	6.7%
Open-question		33.5%	46.2%	58.3%	61.3%	61%	62.6%	63.1%	63%	**63.5%**
Or-clause		67.7%	**71%**	69.8%	66.3%	64.7%	62.7%	60.8%	63.5%	63.3%
Wh-question	58%	77.1%	78.3%	79%	**79.6%**	79.3%	79.4%	79.2%	79.4%	79.4%
Declarative wh-question										
Yes-no-question	54.7%	64.8%	65.7%	**66.4%**	65.8%	65.4%	65.2%	65.2%	65%	65.1%
Declarative yes-no-question		0.2%	0.7%	1.2%	1.3%	1.8%	4.1%	4%	4.2%	**4.4%**
Statement-non-opinion	**97.4%**	93.7%	91.7%	90.3%	89.8%	89.3%	89.3%	89.2%	89.2%	89.1%
Statement-opinion	2.5%	15.9%	30%	37.7%	42.1%	44.6%	46.2%	47%	47.9%	**48.5%**
Self-talk		10.8%	9.3%	**12%**	11.5%	7.9%	10.6%	11.3%	11.3%	11.3%
Macro average	14.7%	22.7%	27.1%	28.9%	30.3%	31%	31.4%	31.4%	**31.6%**	**31.6%**
Micro average	63.6%	67.2%	69.5%	70.6%	71.4%	71.9%	72.2%	72.3%	**72.4%**	**72.4%**

TABLE B.2: Average recall per dialogue act with the first n part-of-speech-tags as features for each $n \in [1, 10]$.

Dialogue Act	1	2	3	4	5	6	7	8	9	10
Collaborative completion			1.5%	2.3%	3%	**3.3%**	3.2%	3.2%	3.2%	3.2%
Tag-question										
Hold before answer/agreement			18.6%	24%	25%	26.4%	26.4%	26.4%	**26.6%**	**26.6%**
Quotation										**0.1%**
Agree/accept			15.1%	15.6%	**15.7%**	**15.7%**	**15.7%**	**15.7%**	**15.7%**	**15.7%**
Action-directive			8.3%	8.3%	8.2%	8.3%	8.3%	8.3%	**8.4%**	**8.4%**
Maybe/accept-part										
Reject										
Acknowledge (backchannel)	**98.4%**	97.1%	98%	98.2%	98.1%	98.1%	98.1%	98.1%	98.1%	98.1%
Repeat-phrase			**2.5%**	**2.5%**	**2.5%**	**2.5%**	**2.5%**	**2.5%**	**2.5%**	**2.5%**
Appreciation		7.7%	9.1%	21.4%	25.7%	28.4%	29.4%	29.5%	**29.6%**	**29.6%**
Downplayer										
Summarize/reformulate										
Backchannel in question form			1.4%	**8.5%**	7.9%	7.9%	7.9%	7.9%	7.9%	7.9%
Response acknowledgement										
Signal-non-understanding			2.4%	**3%**	**3%**	**3%**	**3%**	**3%**	**3%**	**3%**
Offers, options and commits										
Apology										
Conventional-closing			4.9%	5.1%	5.1%	5.1%	7.5%	8.1%	8.1%	**8.2%**
Conventional-opening										
Thanking										
Hedge					34.8%	36.4%	40.2%	40.3%	**40.5%**	**40.5%**
Affirmative non-yes answer										
Dispreferred answer										
Negative non-no answer										
No answer		74.7%	**86.3%**	**86.3%**	**86.3%**	**86.3%**	**86.3%**	**86.3%**	**86.3%**	**86.3%**
Other answer										
Yes answer										
Other										
Rhetorical-question		0.2%			0.2%	0.4%	0.2%	**0.8%**	0.6%	**0.8%**
Open-question		26.8%	28%	26.4%	34.9%	**40.7%**	37%	36.3%	36.1%	36%
Or-clause			4.4%	8.7%	9.5%	9.5%	9%	10.2%	10.6%	**11.2%**
Wh-question	56.1%	66.9%	68.7%	71.4%	70.7%	69.8%	71.3%	71.3%	**72.3%**	**72.3%**
Declarative wh-question										
Yes-no-question	52.1%	51.2%	50.4%	**52.9%**	52.7%	52.5%	52.2%	52.1%	51.9%	51.8%
Declarative yes-no-question										
Statement-non-opinion	75.8%	**98.1%**	97.3%	96.5%	95.6%	94.9%	94.4%	94%	93.8%	93.5%
Statement-opinion	0.3%	0.4%	1%	3.2%	5.3%	7.7%	8.8%	10.2%	11.4%	**12.1%**
Self-talk										
Macro average	7.2%	10.9%	12.8%	13.7%	15%	15.3%	15.4%	15.5%	**15.6%**	**15.6%**
Micro average	53%	62.7%	63.9%	64.5%	64.8%	64.9%	64.9%	65%	**65.1%**	**65.1%**

TABLE B.3: Average recall per dialogue act with the previous n speakers and dialogue acts as features for each $n \in [1, 10]$.

Dialogue Act	1	2	3	4	5	6	7	8	9	10
Collaborative completion										
Tag-question										
Hold before answer/agreement										
Quotation	**30.8%**	21.5%	21.2%	20.4%	19.9%	21.3%	20.8%	21.2%	21.7%	21.1%
Agree/accept	**8.2%**	6%	6.5%	6.3%	6.5%	6.7%	6.8%	6.8%	7%	7.2%
Action-directive			0.7%	1.7%	2%	2.1%	2.1%	2.2%	**2.3%**	**2.3%**
Maybe/accept-part										
Reject										
Acknowledge (backchannel)	**76.3%**	69%	70.1%	69.3%	68.8%	68.7%	68.4%	68.3%	68.1%	68%
Repeat-phrase										
Appreciation										
Downplayer	**18.4%**	7.2%	9.4%	10.8%	10.6%	9.7%	9%	7.8%	6.2%	7.7%
Summarize/reformulate										
Backchannel in question form					**0.1%**					
Response acknowledgement		0.3%	1.7%	1.9%	2.7%	3%	2.9%	3.1%	**3.2%**	3.1%
Signal-non-understanding										
Offers, options and commits										
Apology										
Conventional-closing	**85.1%**	84.8%	84.8%	84.7%	84.9%	84.9%	84.8%	84.9%	84.9%	84.9%
Conventional-opening	**52.7%**	50.2%	50.2%	49.2%	49.7%	49.7%	49.7%	49.8%	49.8%	49.8%
Thanking										
Hedge										
Affirmative non-yes answer		0.1%	0.3%	0.9%	0.8%	1.2%	1%	1.3%	1.3%	**1.5%**
Dispreferred answer							**0.7%**	**0.7%**	**0.7%**	
Negative non-no answer			0.3%	0.3%	0.3%	0.7%	0.7%	0.7%	0.7%	**1.2%**
No answer		2.7%	4%	5%	5.5%	6.2%	6.7%	7%	7.4%	**7.8%**
Other answer				0.3%	0.3%	0.4%	0.4%	0.4%	**0.7%**	0.4%
Yes answer	**70.5%**	68.9%	67.5%	65.3%	64.8%	63.6%	62.8%	61.8%	61.4%	60.6%
Other	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**	**62.9%**
Rhetorical-question		**1.8%**	0.9%	0.7%	0.7%	0.7%	0.9%	0.9%	0.7%	0.7%
Open-question										
Or-clause		5.5%	3.6%	6.5%	7.1%	8.6%	**12%**	9.5%	9.2%	9.2%
Wh-question										**0.1%**
Declarative wh-question										
Yes-no-question						**0.1%**	**0.1%**	**0.1%**	**0.1%**	**0.1%**
Declarative yes-no-question										
Statement-non-opinion	84.4%	**84.5%**	83.9%	84%	84.2%	84.1%	84.3%	84.3%	84.3%	84.3%
Statement-opinion	24.7%	32.1%	32%	32.6%	32.7%	33%	32.9%	32.9%	**33.2%**	**33.2%**
Self-talk										
Macro average	**13.2%**	12.8%	12.8%	12.9%	12.9%	13%	13.1%	13%	13%	13%
Micro average	**57.1%**	56.5%	56.5%	56.5%	56.5%	56.5%	56.4%	56.4%	56.4%	56.4%

TABLE B.4: Average recall per dialogue act for every type of feature individually.

Dialogue Act	F. 9 words	F. 9 POS	Q. mark	Length	Prev. Sp./DA	n-grams	POS n-grams
Collaborative completion	**5.2%**	3.2%					
Tag-question	**14.2%**						7.8%
Hold before answer/agreement	**37.4%**	26.6%				31.8%	31.9%
Quotation	1.3%				**30.8%**	0.1%	0.1%
Agree/accept	**28.9%**	15.7%			8.2%	12%	21%
Action-directive	**24.1%**	8.4%				10.6%	1.4%
Maybe/accept-part							
Reject	**0.3%**						
Acknowledge (backchannel)	96.5%	98.1%		**99.2%**	76.3%	93.7%	97.2%
Repeat-phrase	**3%**	**2.5%**					
Appreciation	**68.6%**	29.6%				41.1%	29%
Downplayer	27%				18.4%	28.6%	**31.6%**
Summarize/reformulate	**1.2%**					**0.1%**	
Backchannel in question form	**57.4%**	7.9%				12.3%	38.5%
Response acknowledgement	**31.6%**					30.4%	3.8%
Signal-non-understanding	**30.5%**	3%				16.8%	18%
Offers, options and commits	**12.6%**						
Apology	61.9%						
Conventional-closing	54.3%	8.1%			**85.1%**	23.1%	13.4%
Conventional-opening	**58.6%**				52.7%	40.7%	1.9%
Thanking	5.1%					2%	**31.5%**
Hedge	**69.1%**	40.5%				65.2%	66%
Affirmative non-yes answer	**1.2%**						
Dispreferred answer							
Negative non-no answer	**15.6%**						
No answer	88.9%	86.3%				**89.5%**	86.5%
Other answer							
Yes answer	2.5%				**70.5%**	0.2%	1%
Other	5.5%				**62.9%**		0.8%
Rhetorical-question	**8%**	0.6%				0.4%	5.3%
Open-question	**63%**	36.1%				59%	53%
Or-clause	**63.5%**	10.6%				35.4%	25.6%
Wh-question	**79.4%**	72.3%				43.1%	66%
Declarative wh-question							
Yes-no-question	65%	51.9%	**74.9%**			30.4%	38.7%
Declarative yes-no-question	**4.2%**					1.4%	0.2%
Statement-non-opinion	89.2%	93.8%	**99.9%**	98.4%	84.4%	92.7%	90.9%
Statement-opinion	**47.9%**	11.4%		0.5%	24.7%	25.6%	22.4%
Self-talk	**11.3%**					1%	10.3%
Macro average	**31.6%**	15.6%	4.5%	5.1%	13.2%	20.2%	20.4%
Micro average	**72.4%**	65.1%	42.2%	60.6%	57.1%	66%	65.9%

TABLE B.5: Average recall per dialogue act for combinations of features. The first nine words were used as the baseline and all other features were added on one by one.

Dialogue Act	F. 9 words	+ F. 9 POS	+ Q. mark	+ Length	+ Prev. Sp./DA	+ n-grams	+ POS n-grams
Collaborative completion	5.2%	6.3%	6.7%	6.6%	7.4%	8.3%	**8.4%**
Tag-question	14.2%	20.6%	42.6%	42.6%	49.2%	51.5%	**52.4%**
Hold before answer/ agreement	37.4%	37.4%	36.9%	37.1%	38.4%	**40.3%**	40.1%
Quotation	1.3%	1.7%	1.6%	1.7%	22.3%	**23%**	22.6%
Agree/accept	28.9%	29%	29%	29%	40.8%	40.9%	**41.2%**
Action-directive	24.1%	26.2%	25.2%	25.5%	27.4%	28.1%	**29.6%**
Maybe/accept-part						1%	**1%**
Reject	0.3%	0.3%	0.3%	0.3%	6.4%	**6.9%**	6.6%
Acknowledge (backchannel)	96.5%	96.5%	**97.2%**	**97.2%**	96.1%	96.1%	96.1%
Repeat-phrase	3%	4.5%	4.4%	4.3%	6.7%	**7.5%**	7.1%
Appreciation	68.6%	69.4%	**69.8%**	69.7%	69.4%	69.3%	69.5%
Downplayer	27%	29.8%	29.8%	29.8%	34.1%	38%	**40.3%**
Summarize/reformulate	1.2%	1%	1.1%	1.2%	1.5%	2.2%	**2.4%**
Backchannel in question form	57.4%	57.2%	65.5%	65.5%	66.5%	67%	**68.1%**
Response acknowledgement	31.6%	31.6%	31.6%	31.6%	51.3%	**51.8%**	51.3%
Signal-non-understanding	30.5%	40.7%	53.3%	53.3%	62.6%	64.2%	**65.3%**
Offers, options and commits	12.6%	13.5%	13.5%	13.5%	11.9%	**14.4%**	**14.4%**
Apology	61.9%	61.9%	66.7%	66.7%	63.8%	64.9%	**68%**
Conventional-closing	54.3%	54.5%	54.5%	54.6%	91.8%	**92.5%**	92.4%
Conventional-opening	58.6%	60.5%	59.5%	59.5%	79.4%	**83.5%**	83.2%
Thanking	5.1%	4%	4%	4%	47.5%	**58.6%**	**58.6%**
Hedge	69.1%	68.6%	68.9%	**69.2%**	66.6%	68.7%	68.9%
Affirmative non-yes answer	1.2%	0.8%	0.8%	0.8%	41%	**42.8%**	41.3%
Dispreferred answer		0.9%		0.9%	5.1%	4.7%	**5.5%**
Negative non-no answer	15.6%	16.4%	16.4%	16.4%	34%	34.5%	**35.2%**
No answer	88.9%	**89%**	88.9%	88.9%	86.8%	87%	86.8%
Other answer					23.7%	**25.1%**	21.5%
Yes answer	2.5%	2%	2%	2%	76%	**76.4%**	**76.4%**
Other	5.5%	5.5%	5.4%	5.4%	60.1%	60.3%	**60.6%**
Rhetorical-question	8%	10.1%	10.1%	10.1%	12.4%	13.9%	**14.9%**
Open-question	63%	62.3%	61.6%	61.1%	60.8%	63.6%	**64.5%**
Or-clause	63.5%	64.5%	65.3%	65.1%	**75.4%**	75.3%	74.6%
Wh-question	79.4%	78.9%	81.3%	80.9%	80.9%	81.3%	**82.3%**
Declarative wh-question							**1.3%**
Yes-no-question	65%	66.1%	78.3%	78.3%	77.7%	78.3%	**78.9%**
Declarative yes-no-question	4.2%	4.1%	7.4%	7.6%	8%	10.9%	**12.6%**
Statement-non-opinion	89.2%	89%	89.2%	89.2%	**90.1%**	90%	90%
Statement-opinion	47.9%	48.1%	48.3%	48.5%	56.3%	57%	**57.4%**
Self-talk	11.3%	14.6%	16%	16.5%	21.6%	**26.5%**	25.2%
Macro average	31.6%	32.5%	34.2%	34.2%	44.9%	46.3%	**46.6%**
Micro average	72.4%	72.4%	73.1%	73.2%	77.7%	77.9%	**78%**

B.2 Confusion Matrix

The following three tables are the three parts of the confusion matrix of a classifier with the following features: first nine words and part-of-speech tags, question mark, length, previous speaker and dialogue act, n-grams and part-of-speech n-grams. The rows show the actual utterance labels, while the columns show the labels assigned by the classifier. Each cell contains a percentage that indicates how often its corresponding dialogue act combination occurred. This value is averaged over all ten folds. The diagonals are shown in bold and contain the average recall for the corresponding classes. Blank cells represent 0% and the others are color coded from red (worst) to green (best). See table A.1 for an overview of which tag represents which dialogue act.

TABLE B.6: The confusion matrix of a classifier that uses all seven types of features. Part one of three.

Tag	^2	^g	^h	^q	aa	ad	am	ar	b	b^m	ba	bd	bf
^2	**8.4%**	0.2%	0.1%	0.6%	1.7%	0.7%			8.9%	2.7%	1.4%		0.5%
^g		**52.4%**		0.7%	4.1%				11.2%		0.7%		1.1%
^h		0.2%	**40.1%**	0.2%	0.9%	1%			4.7%	0.2%	1.4%		
^q	0.4%	0.1%		**22.6%**	0.8%	0.8%		0.2%	1.5%		0.8%		
aa	0.1%				**41.2%**			0.1%	47.2%	0.1%	1.2%		
ad	0.2%		0.9%	0.6%	0.9%	**29.6%**		0.1%	1%	0.4%	0.3%		
am	2%				17.2%		**1%**		8.5%	2%			
ar	0.3%				35.5%			**6.6%**	1.5%	0.3%	3.7%		0.7%
b					1.6%				**96.1%**		0.3%		
b^m	2.8%		0.1%	0.3%	2.8%	0.5%			41.4%	**7.1%**	2.2%		0.7%
ba	0.1%				9.4%	0.1%			4.9%	0.3%	**69.5%**	0.1%	0.1%
bd					4.7%			0.9%	6.3%		17.3%	**40.3%**	0.7%
bf	2.1%			0.4%	1.3%	0.4%			1.5%	1.3%	0.9%		**2.4%**
bh		0.2%			0.9%	0.1%			11.4%	0.1%	2.2%		0.2%
bk					0.8%				44.9%	0.2%	1.1%	0.1%	
br		1.1%	0.3%						1.6%		0.6%		
co			2.2%		4.7%	6.4%							
fa						4%			1.1%		1%		
fc	0.1%				0.1%				1.5%		0.2%		
fp									1.4%	1.3%	1.5%		
ft											1.1%		
h	0.2%		0.4%	0.1%	0.5%	0.2%	0.1%		0.5%				
na	0.1%		0.2%		3.9%				7%		1.2%		
nd			0.4%		3%		0.5%	1.5%	6.2%				
ng			0.4%		0.7%			0.3%		0.3%			
nn				0.1%	8.8%			1.1%	1.2%	0.1%	0.6%		
no			0.7%		1.1%		0.3%		1%	0.3%			
ny					2.1%				20.7%				
o	0.4%		0.6%	0.2%	3.4%	0.4%		0.1%	10.5%	0.8%	2.9%		
qh	0.2%	0.2%	0.2%	1.5%	0.4%	0.5%			1.2%		0.9%		0.2%
qo			0.2%	0.2%		0.2%				0.2%	0.2%		0.2%
qrr			0.5%										
qw	0.1%		0.1%	0.1%		0.1%			0.2%	0.1%	0.2%		
qw^d			2.6%	0.8%							1%		
qy	0.3%	0.4%		0.1%	0.2%	0.2%			0.6%	0.1%	0.2%		0.1%
qy^d	0.7%			0.2%	0.3%	0.2%			0.7%	0.8%	1.1%		0.6%
sd				0.1%	0.4%				0.2%	0.1%	0.2%		
sv	0.1%			0.1%	1.1%	0.1%			0.2%		1.2%		0.1%
t1			6.3%		1.1%				1.6%	1.3%	0.5%		

TABLE B.7: The confusion matrix of a classifier that uses all seven types of features. Part two of three.

Tag	bh	bk	br	co	fa	fc	fp	ft	h	na	nd	ng	nn
^2	0.4%		0.5%							0.7%			
^g	8.3%		1.7%							1.6%			
^h		1.4%		0.5%					0.4%	4.1%		0.9%	1%
^q				0.4%	0.1%			0.1%	0.3%				0.5%
aa		0.2%								0.1%			0.5%
ad				0.1%	0.5%	0.7%	0.2%		0.4%	0.2%			
am									3.8%	3.3%			
ar					0.7%				1.9%				20%
b		0.7%								0.1%			0.1%
b^m		0.3%	0.1%				0.2%			1.4%		0.3%	3.7%
ba	0.1%	0.2%								0.1%			
bd		3.3%				2.6%			1.7%				2.8%
bf	0.1%		0.1%			0.1%							
bh	**68.1%**	0.1%	0.3%			0.1%						0.1%	
bk	0.1%	**51.3%**								0.2%			0.2%
br	1%		**65.3%**		0.9%		0.4%			0.3%			0.8%
co				**14.4%**		1.4%				0.6%			
fa			1.4%		**68%**			1%					
fc	0.1%	0.2%		0.1%		**92.4%**		0.5%					
fp						1.1%	**83.2%**						
ft						34.8%		**58.6%**					
h					0.1%				**68.9%**	0.3%		0.2%	
na		0.4%								**41.3%**	0.4%	0.9%	0.4%
nd				0.5%						4.1%	**5.5%**	1.9%	1.5%
ng			0.3%							8%	1.1%	**35.2%**	2.4%
nn						0.1%				0.1%		0.4%	**86.8%**
no									15.9%	10.1%	0.4%	5%	
ny										0.2%		0.1%	0.1%
o		1%	0.1%		0.3%	0.1%	0.3%	0.1%	0.3%	0.1%			2.5%
qh	1.2%								0.4%	0.2%		0.2%	
qo			0.2%										
qrr												0.5%	
qw			0.9%			0.1%				0.1%			
qw^d			1.3%				1.4%						
qy	2.7%		0.5%			0.1%				0.1%			
qy^d	1.7%		0.7%						0.1%	0.6%		0.3%	
sd									0.2%	0.2%			
sv										0.1%			
t1			0.7%										0.7%

TABLE B.8: The confusion matrix of a classifier that uses all seven types of features. Part three of three.

Tag	no	ny	o	qh	qo	qrr	qw	qw^d	qy	qy^d	sd	sv	t1
^2			0.2%	0.1%		0.1%	0.5%		1.8%	0.4%	50%	20.1%	
^g				1.4%			1%		13.8%	1%	1.1%		
^h	0.5%		1.2%	0.2%	0.2%		0.7%		0.5%		33.2%	5.6%	1%
^q				0.9%	0.2%	0.1%	2.2%		2.3%	0.3%	52.2%	13.1%	0.2%
aa		0.5%	0.2%								4.6%	4%	
ad	0.2%		0.9%				0.3%		2.9%	0.5%	40.9%	18.1%	
am											18.6%	43.7%	
ar											19%	9.9%	
b		0.7%	0.1%								0.3%	0.1%	
b^m		0.6%	0.3%				0.4%		1.9%	0.8%	28%	4.1%	
ba		0.1%		0.1%							7.8%	7%	
bd											11.7%	7.9%	
bf				0.3%			0.3%		1.9%	1.2%	58.5%	27.3%	
bh		0.1%		0.6%	0.1%		0.5%		13.3%	0.3%	1.1%	0.3%	
bk		0.2%	0.2%								0.8%		
br							11.3%		12.5%	2.5%	0.7%	0.8%	
co											64.5%	5.8%	
fa			2.4%						1%		20.1%		
fc							0.1%		0.3%		2.9%	1.5%	
fp			2.4%				1.5%				6.3%	1.3%	
ft											5.4%		
h	1.1%		0.1%	0.3%			0.2%			0.2%	22.4%	4.2%	0.1%
na	0.5%	0.5%	0.3%						0.2%	0.3%	36.1%	6.5%	
nd	1.4%	0.4%								0.8%	62.6%	9.9%	
ng	3.5%										43.4%	4.4%	
nn		0.2%									0.6%	0.1%	
no	**21.5%**	0.3%	0.4%	0.4%			0.3%			1.2%	37%	4.2%	
ny		**76.4%**	0.3%								0.1%		
o		0.5%	**60.6%**				0.1%		1%		10.8%	2.9%	
qh			0.2%	**14.9%**	1.6%		17.8%	0.2%	13.9%	0.8%	26.3%	17.3%	0.2%
qo				0.6%	**64.5%**		26.9%		3.6%	0.2%	2.3%	0.8%	
qrr				0.9%	1.5%	**74.6%**	0.9%		12.4%	0.3%	6.2%	2.2%	
qw			0.1%	1.6%	5.3%	0.1%	**82.3%**	0.1%	2.8%	0.3%	4.3%	1.1%	0.2%
qw^d					0.8%		8.5%	**1.3%**	3.9%	5.1%	65.2%	8.2%	
qy		0.2%	0.2%	0.5%	0.3%	0.8%	1.2%		**78.9%**	3.4%	7%	2.4%	
qy^d		0.1%	0.1%	0.2%			0.7%	0.1%	17.1%	**12.6%**	47.6%	13.9%	
sd									0.1%		**90%**	8.3%	
sv				0.1%					0.1%	0.1%	39.1%	**57.4%**	
t1				0.5%			24.3%		4.6%	0.7%	29.7%	3.1%	**25.2%**

Appendix C

Classification Tree

TABLE C.1: The characteristics of the nodes in the classification tree. All grey rows as well as the child nodes marked in bold correspond to leaves.

Node	Characteristics	
Root	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark
	Excluded DAs	^2, ^q, am, ar, b^m, bf, co, ft, na, nd, ng, no, ny, o, qh, qw^d, qy^d, t1
	Merged DAs	None
	Children	**^g1**, **^h1**, aa1, ad1, b1, ba1, **bd1**, **bh1**, **bk1**, br1, **fa1**, fc1, **fp1**, **h1**, **nn1**, **qo1**, **qrr1**, qw1, qy1, sd1, sv1
^g1	Type	Leaf
	Associated DAs	Tag-question (90%), yes-no-question (10%)
	Description	Tag questions
	Examples	"Right?", "Isn't it?", "Didn't he?"
^h1	Type	Leaf
	Associated DAs	Hold before answer (83.1%)
	Description	Utterances that indicate the speaker is thinking about what to say next.
	Examples	"Let's see", "Let me think", "I'm trying to think"
bd1	Type	Leaf
	Associated DAs	Downplayer (72.1%), appreciation (7%), statement-opinion (7%), statement-non-opinion (7%)
	Description	Utterances that downplay what was said before.
	Examples	"That's okay", "That's all right"
bh1	Type	Leaf
	Associated DAs	Backchannel in question form (83.5%), yes-no-question (10.3%)
	Description	Short questions that request confirmation about something.
	Examples	"Really?", "Is that right?", "They do?"
bk1	Type	Leaf
	Associated DAs	Response acknowledgement (64.5%), acknowledge (29.6%)
	Description	Acknowledgements of a response (typically an answer)
	Examples	"Oh okay", "Oh I see"
fa1	Type	Leaf
	Associated DAs	Apology (78.3%)
	Description	Apologies
	Examples	"(I'm) sorry", "Excuse me"
fp1	Type	Leaf
	Associated DAs	Conventional-opening (93.9%)
	Description	Utterances that are common during the opening sequence of a conversation.
	Examples	"Hello", "Hi", "My name is ...", "How you doing?"
h1	Type	Leaf
	Associated DAs	Hedge (79.3%), other answer (9%), statement-non-opinion (8.1%)
	Description	Utterances that express uncertainty or a lack of knowledge.
	Examples	"I don't know", "I'm not sure"
nn1	Type	Leaf

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Associated DAs	No answer (55.2%), accept (24.5%), reject (8.8%)
	Description	Negative responses that involve the words "no" or "nope".
	Examples	"No", "Nope"
qo1	Type	Leaf
	Associated DAs	Open-question (91.7%)
	Description	Open-ended wh-questions
	Examples	"How/what about ...?", "(What) do you think (about) ...?", "How do you feel about ...?"
qrr1	Type	Leaf
	Associated DAs	Or-clause (92%), yes-no-question (5.2%)
	Description	Yes-no-questions that start with the word "or".
	Examples	"Or was she kind of opposed to it?", "Or is it?"
aa1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	ba, bf, sd, sv
	Merged DAs	None
	Children	**aa2, b2, bh2, fc2, h2, na2, ng2**
aa2	Type	Leaf
	Associated DAs	Accept (71.6%), appreciation (11.8%), statement-non-opinion (7.2%), statement-opinion (6.3%)
	Description	Non-yes positive responses.
	Examples	"That's true/right", "Exactly", "I agree", "It is", "I know"
b2	Type	Leaf
	Associated DAs	Acknowledge (55.4%), statement-non-opinion (10.8%), yes-no-question (7.7%), other (7.7%), rhetorical-question (7.7%)
	Description	Certain acknowledges/agreements and filler utterances.
	Examples	"You know", "Well, all right", "Yeah, well"
bh2	Type	Leaf
	Associated DAs	Backchannel in question form (55.6%), reformulate (22.2%)
	Description	Utterances with an interrogative form that function as a backchannel.
	Examples	"Isn't that true", "Oh do you"
fc2	Type	Leaf
	Associated DAs	Conventional-closing (100%)
	Description	Positive responses that occurred during the closing sequence of the conversation.
	Examples	"I think so", "Absolutely", "It is"
h2	Type	Leaf
	Associated DAs	Hedge (50%), statement-opinion (14.3%)
	Description	Utterances that express uncertainty or a lack of knowledge.
	Examples	"I guess (so)", "Sort of"
na2	Type	Leaf
	Associated DAs	Affirmative non-yes answer (85.6%), statement-non-opinion (5.4%)
	Description	Affirmative answers to a yes-no-question that do not contain a form of "yes".
	Examples	"I do", "That's right", "Probably so", "Absolutely"
ng2	Type	Leaf
	Associated DAs	Negative non-no answer (84.8%), no answer (6.1%)
	Description	Negative answers to a yes-no-question that do not contain the word "no".
	Examples	"Not really", "Probably not", "Not quite"
ad1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, n-grams
	Excluded DAs	None
	Merged DAs	None
	Children	**^q2, ad2, co2, sd2**
^q2	Type	Leaf
	Associated DAs	Quotation (100%)
	Description	Utterances with an imperative form. The difference with ad2 is not really clear, so this node may not be very useful.
	Examples	"Go get her", "Let me explain"
ad2	Type	Leaf

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Associated DAs	Action directives (96.6%)
	Description	Utterances with an imperative form.
	Examples	"Go ahead", "Tell me", "Don't bother"
co2	Type	Leaf
	Associated DAs	Offers, options and commits (100%)
	Description	Utterances with an imperative form that contain "let me" in some way.
	Examples	"Let me ask you ..."
sd2	Type	Leaf
	Associated DAs	Statement-non-opinion (100%)
	Description	Utterances with an imperative form. The difference with ad2 is not really clear, so this node may not be very useful.
	Examples	"Go down", "Put them back"
b1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark
	Excluded DAs	b, bf, sd
	Merged DAs	b^m merged with ^2, am, na and ny merged with aa
	Children	^22, **^h2**, aa2-2, **ba2, bh2-2**, bk2, **nd2, nn2, o2, qy2, sv2**
^h2	Type	Leaf
	Associated DAs	Acknowledge (96.2%)
	Description	Only contains "um" utterances.
	Examples	"Um."
ba2	Type	Leaf
	Associated DAs	Acknowledge (45.9%), appreciation (40.5%), repeat-phrase (5.4%)
	Description	Short utterances that consist of words such as "uh", "oh" and "um", sometimes followed by an adjective.
	Examples	"Uh, oh", "Oh, um". "Huh, interesting", "Uh, strange"
bh2-2	Type	Leaf
	Associated DAs	Acknowledge (64.6%), backchannel in question form (31.4%)
	Description	Questions that function as acknowledges.
	Examples	"Oh, really.", "Is that right."
nd2	Type	Leaf
	Associated DAs	Acknowledge (36.8%), dispreferred answer (28.9%), statement-non-opinion (18.4%), hedge (7.9%)
	Description	Contains only "well" utterances.
	Examples	"Well."
nn2	Type	Leaf
	Associated DAs	Acknowledge (70.1%), no answer (20.2%), accept (8.2%)
	Description	Contains only "huh-uh" utterances.
	Examples	"Huh-uh"
o2	Type	Leaf
	Associated DAs	Other (44.3%), acknowledge (35.7%)
	Description	Acknowledges that signal the speaker is thinking about what to say next.
	Examples	"Okay, uh/um.", "Oh, well.", "Uh/um, all right."
qy2	Type	Leaf
	Associated DAs	Yes-no-question (65%), backchannel in question form (20%), acknowledge (10%)
	Description	Fragmentary utterances that form or seem to be the start of a question.
	Examples	"Or", "Do you"
sv2	Type	Leaf
	Associated DAs	
	Description	Mostly fragments and statements that contain "yes" and/or start with "and" or "but".
	Examples	"And", "But", "But in some cases, yes.", "And yes, the cost of living is super high."
^22	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark
	Excluded DAs	sd
	Merged DAs	b^m merged with ^2
	Children	**^23, b3**
^23	Type	Leaf

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Associated DAs	Repeat-phrase (24.9%), statement-non-opinion (21.4%), collaborative completion (16.8%), appreciation (5.9%)
	Description	Utterances that contain just a single word, usually a noun or adjective.
	Examples	"Beautiful", "Radio", "Glass", "One"
b3	Type	Leaf
	Associated DAs	Acknowledge (96.8%)
	Description	Acknowledging utterances such as "huh" and "ah".
	Examples	"Huh", "Um-hum", "Hm", "Ah"
aa2-2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA
	Excluded DAs	All but aa, ny, na, fc
	Merged DAs	None
	Children	**aa3, fc3, na3, ny3**
aa3	Type	Leaf
	Associated DAs	Acknowledge (82.2%), accept (15.3%)
	Description	Positive responses that do not follow after a yes-no-question.
	Examples	"Right", "Yes", "Yeah", "Uh-huh"
fc3	Type	Leaf
	Associated DAs	Conventional-closing (93.6%)
	Description	Positive responses that occur during the closing sequence of the conversation.
	Examples	""Right", "Yes", Yeah", "Uh-huh"
na3	Type	Leaf
	Associated DAs	Affirmative non-yes answer (72.4%), acknowledge (23.6%)
	Description	Non-yes positive responses that follow after a yes-no-question.
	Examples	"Right", "Sure"
ny3	Type	Leaf
	Associated DAs	Yes answer (53.4%), acknowledge (39.7%)
	Description	Positive, yes-type responses that follow after a yes-no-question.
	Examples	"Yes", "Yeah", "Uh-huh"
bk2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA
	Excluded DAs	All but bk and fc
	Merged DAs	None
	Children	**bk3, fc3-2**
bk3	Type	Leaf
	Associated DAs	Acknowledge (61.6%), response acknowledgement (20.2%), other (11.6%)
	Description	Acknowledges such as "Okay" and "I see".
	Examples	"Oh", "Okay", "I see", "Oh, uh-huh"
fc3	Type	Leaf
	Associated DAs	Conventional-closing (97.2%)
	Description	"Okay" utterances that occur during the closing sequence of the conversation.
	Examples	"Okay"
ba1	Type	Internal
	Clustering Method	Rules
	Rules	If the utterance contains a modal verb or a verb in the 3rd person singular present, assign it to sv2, else to ba2.
	Excluded DAs	^2, ^q, bf, qh
	Merged DAs	None
	Children	ba2, **sv2**
sv2	Type	Leaf
	Associated DAs	Appreciation (77.5%), statement-opinion (12.7%)
	Description	Statements that express some degree of appreciation for something.
	Examples	"That's awful.", "That sounds great.", "I'll bet.", "That would be wonderful."
ba2	Type	Internal
	Clustering Method	Classifier

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA
	Excluded DAs	^2, ^q, bf, qh
	Merged DAs	None
	Children	**aa3-2, b3-2, ba3, fc3-3, na3-2, nn3, sd3, sv3,**
aa3-2	Type	Leaf
	Associated DAs	Accept (63.2%), appreciation (23.7%), acknowledge (6.6%)
	Description	Positive responses to (mostly) opinions.
	Examples	"You bet.", "Oh, sure.", "There you go."
b3-2	Type	Leaf
	Associated DAs	Acknowledge (92.3%)
	Description	Utterances that mostly just consists of "oh". Perhaps not the most useful group.
	Examples	"Ooh.", "Oh, oh."
ba3	Type	Leaf
	Associated DAs	Appreciation (87.1%)
	Description	Mostly exclamations that express some degree of appreciation over something.
	Examples	"Wow", "Oh no", "Oh, my God.", "I understand.", "Unbelievable"
fc3-3	Type	Leaf
	Associated DAs	Conventional-closing (100%)
	Description	Appreciation-type utterances that occur during the closing sequence of the conversation.
	Examples	"Sounds fun.", "Great.", "No kidding."
na3-2	Type	Leaf
	Associated DAs	Affirmative non-yes answer (77.3%)
	Description	Affirmative non-yes answers to a yes-no-question
	Examples	"Oh sure.", "Oh, all the time.", "Oh constantly."
nn3	Type	Leaf
	Associated DAs	No answer (100%)
	Description	No-type answers in response to a yes-no-question.
	Examples	"Oh, no.", "Oh goodness no."
sd3	Type	Leaf
	Associated DAs	Statement-non-opinion (88.9%)
	Description	A bit of a junk class that contains short elliptical statements and appreciative utterances.
	Examples	"Oh, uh, eighty-six.", "Pretty good.", "Boy, what a difference."
sv3	Type	Leaf
	Associated DAs	Statement-opinion (87.1%), statement-non-opinion (9.7%)
	Description	The remaining appreciative statements that did not get caught by the rules of node ba1.
	Examples	"That was great.", "Oh, those are nice.", "Gosh, was that wonderful."
br1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	None
	Merged DAs	^2, b^m and qy^d merged with qy
	Children	**br2, qy2-2**
br2	Type	Leaf
	Associated DAs	Signal-non-understanding (90.7%)
	Description	Short utterances that indicate the speaker did not understand something.
	Examples	"Huh?", "Pardon?", "What?", "Excuse me?"
qy2-2	Type	Leaf
	Associated DAs	Declarative yes-no-question (40%), collaborative completion (22%), yes-no-question (16%), repeat-phrase (12%)
	Description	Short, elliptical questions.
	Examples	"Manchester?", "The university?", "Eighty-six?"
fc1	Type	Internal
	Clustering Method	Rules
	Rules	If the utterance contains "thank", assign it to ft2, else to fc2-2.
	Excluded DAs	None
	Merged DAs	None
	Children	fc2-2, **ft2**

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
ft2	Type	Leaf
	Associated DAs	Conventional-closing (60.5%), thanking (34.1%)
	Description	Utterances that express gratitude.
	Examples	"Thank you (for ...)", "Thanks (for ...)",
fc2-2	Type	Internal
	Clustering Method	Rules
	Rules	If the utterance contains "all right" or "alright", assign it to o3, else to fc3-4
	Excluded DAs	None
	Merged DAs	None
	Children	**fc3-4, o3**
fc3-4	Type	Leaf
	Associated DAs	Conventional-closing (93.2%)
	Description	The remaining closing utterances.
	Examples	"Bye", "It's been nice talking to you", "Take care"
o3	Type	Leaf
	Associated DAs	Conventional-closing (33.3%), acknowledge (24.8%), other (20.6%), accept (9.2%), response acknowledgement (8.3%)
	Description	Utterances that contain "all right"/"alright". It might be useful to subdivide this group further, because there are functional differences in how this type of utterance is used.
	Examples	"All right", "Alright"
qw1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sv
	Merged DAs	qw^d merged with qw, qy^d with qy
	Children	**br2-2, qo2, qw2, qy2-3, t12**
br2-2	Type	Leaf
	Associated DAs	Signal-non-understanding (66.7%), wh-question (23.3%)
	Description	Short wh-questions that indicate the speaker did not understand something.
	Examples	"What did you say?", "What's that?", "Did what?"
qo2	Type	Leaf
	Associated DAs	Open-question (100%)
	Description	Wh-questions that tend to be a bit more open, but overall the distinction with the regular wh-questions is not very clear.
	Examples	"What do you have planned for your yard?", "How does he like it?"
qw2	Type	Leaf
	Associated DAs	Wh-question (83.9%), rhetorical-question (6.3%)
	Description	Remaining wh-questions.
	Examples	"What grade do you teach?", "How old is he?", "Who wrote that?", "Which ones do you like?"
qy2-3	Type	Leaf
	Associated DAs	Yes-no-question (76.9%), declarative yes-no-question (23.1%)
	Description	Yes-no-questions that resemble or contain a wh-question.
	Examples	"When you're in the water?", "Where did you say you're at, Colorado?",
t12	Type	Leaf
	Associated DAs	Self-talk (95.5%)
	Description	Wh-questions that the speaker asks themselves and does not expect to be answered.
	Examples	"What is it", "What's the word", "Where did I put it"
qy1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	bh merged with ^g, qy^d with qy and sv with sd
	Children	**^g2, ad2-2, qw2-2, qy2-4, sd2-2**
^g2	Type	Leaf
	Associated DAs	Backchannel in question form (56.3%), yes-no-question (33.8%), tag-question (7.5%)
	Description	Tag-questions and other short yes-no-questions.
	Examples	"Are you?", "Don't they?", "Are you serious?", "Is it?"

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
ad2-2	Type	Leaf
	Associated DAs	Action-directive (92.3%)
	Description	Yes-no-questions that push the addressee to do something.
	Examples	"Could you hold on?", "Keep on watching those movies, huh?", "You want to start first?"
qw2-2	Type	Leaf
	Associated DAs	Wh-question (100%)
	Description	Wh-questions structured in less conventional ways.
	Examples	"If you had a choice of your car, what would you get?", "Wonder about them in what way?", "So, uh, I mean what was it for?"
qy2-4	Type	Leaf
	Associated DAs	Yes-no-question (85.9%), declarative yes-no-question (6.8%)
	Description	Remaining yes-no-questions
	Examples	"Is that correct?", "Do you have any pets?", "So you can't use oil on wood?"
sd2-2	Type	Leaf
	Associated DAs	Statement-non-opinion (54.3%), statement-opinion (43.5%)
	Description	Statements with an interrogative element. Usually through tag-questions or the use of fillers such as "you know?".
	Examples	"Jot them down, huh?", "As you know, I think it's more like the mess you make, isn't it?", "If you don't have bars on the window you're not even safe you know?"
sd1	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark
	Excluded DAs	^2, ^q, b^m, bf, bh, qh, sd, co, fa, o
	Merged DAs	am and ar merged with aa; ^h, nd, ng and no merged with na; qw^d with qw; qy^d with qy
	Children	aa2-3, ad2-3, **b2-2**, ba2-2, fc2-3, **fp2**, h2-2, na2-2, **qo2-2, qrr2**, qw2-3, qy2-5, sv2-2, **t12-2**
b2-2	Type	Leaf
	Associated DAs	Repeat-phrase (28.5%), acknowledge (22%), statement-non-opinion (17.2%), collaborative completion (14.5%), reformulate (7.5%)
	Description	Elliptical sentences.
	Examples	"Sesame seeds and bread crumbs.", "Changed directions.", "Oh, the weather.", "With computers."
fp2	Type	Leaf
	Associated DAs	Conventional-opening (54.1%), statement-non-opinion (28.2%), other (16.5%)
	Description	Utterances with which the speaker introduces themselves.
	Examples	"My name is ...", "I'm from ...", "I work for ..."
qo2-2	Type	Leaf
	Associated DAs	Open-question (92.3%)
	Description	Open wh-questions
	Examples	"What was it like living there.", "What about in New York.", "Where do you stand on ..."
qrr2	Type	Leaf
	Associated DAs	Or-clause (100%)
	Description	Yes-no-questions that start with "or".
	Examples	"Or was she doing it on her own.", "Or down here in Denison.", "Or building supplies place."
t12-2	Type	Leaf
	Associated DAs	Self-talk (72.7%), statement-non-opinion (14.5%), quotation (5.5%)
	Description	Wh-questions, often with the speaker as the subject. Usually the speaker is talking to themselves.
	Examples	"What was it.", "What else did I serve with that.", "What am I trying to say."
aa2-3	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sd, sv, ba

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Merged DAs	None
	Children	**aa3-3, am3, ar3**
aa3-3	Type	Leaf
	Associated DAs	Statement-non-opinion (64.8%), accept (25.2%)
	Description	Short to medium-length statements that usually have the speaker as subject and are in response to something said previously.
	Examples	"Well, we do too.", "So I'll have to go back and look at that.", "That's what I've heard.", "Well if I were closer I might."
am3	Type	Leaf
	Associated DAs	Maybe (65.5%), statement-non-opinion (27.6%)
	Description	Statements with some element of uncertainty in them.
	Examples	"Something like that.", "Well, it sort of is.", "Uh, I guess not."
ar3	Type	Leaf
	Associated DAs	Statement-non-opinion (49.4%), reject (44.6%)
	Description	Statements that contain a negation or other elements that indicate a contradiction with what was previously said..
	Examples	"I don't know about that,", "Well, actually I'm from California.", "I frequently disagree with his commentaries,",
ad2-3	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	None
	Children	**ad3, sd3-2**
ad3	Type	Leaf
	Associated DAs	Action-directive (88.9%), statement-non-opinion (6.8%)
	Description	Statements that tell the addressee to do something.
	Examples	"Use about half the sugar.", "I'll let you go first.", "So don't say she's small, just say she's perfect."
sd3-2	Type	Leaf
	Associated DAs	Statement-non-opinion (88.2%)
	Description	The remaining statements with some elements from action-directives.
	Examples	"Just make good friends.", "Uh, oh okay you name it.", "I don't happen to have one at home."
ba2-2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	None
	Children	**ba3-3, sd3-3**
ba3-3	Type	Leaf
	Associated DAs	Appreciation (100%)
	Description	Statements that express some degree of appreciation. Not very distinct from sd3-3, so the division may not be that useful.
	Examples	"I understand what you're saying.", "I don't blame you.", "That is pretty big though."
sd3-3	Type	Leaf
	Associated DAs	Statement-non-opinion (70%), appreciation (20.1%)
	Description	Statements that express some degree of appreciation. Not very distinct from ba3-3, so the division may not be that useful.
	Examples	"I know what you mean.", "I hate that show.", "Oh, now that was fabulous.", "Sounds like my grandchildren."
fc2-3	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	None
	Children	**fc3-5, sd3-4**
fc3-5	Type	Leaf
	Associated DAs	Conventional-closing (99.7%)

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Description	Statements that occur during the closing sequence of the conversation.
	Examples	"I have got to go.", "I gathered you might be a teacher.", "I hope you have a nice day.", "We had no damage to our house or anything."
sd3-4	Type	Leaf
	Associated DAs	Statement-non-opinion (94.4%), conventional-closing (5.4%)
	Description	Mostly statements that start with "well" and/or contain "I've".
	Examples	"Well I've got two.", "Well, I enjoy fiddling around.", "I've got to have somebody to compete with."
h2-2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	None
	Children	**h3, sd3-5**
h3	Type	Leaf
	Associated DAs	Hedge (100%)
	Description	Statements that contain "don't know". The distinction with sd3-5 is not very clear, so this division may not be that useful.
	Examples	"You know, but, um, I don't know.", "And I don't know about your part of the country.", "I don't know if that's a good thing or a bad thing."
sd3-5	Type	Leaf
	Associated DAs	Statement-non-opinion (85.5%), hedge (11.4%)
	Description	Statements that contain "don't know" or other negations.
	Examples	"I don't know about education.", "I can't remember.", "I'm not sure what they're called now."
na2-2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sd, sv, ba
	Merged DAs	None
	Children	**^h3, na3-3, nd3, ng3, no3**
^h3	Type	Leaf
	Associated DAs	Statement-non-opinion (83.8%), ^h (5.1%)
	Description	Short to medium-length statements, often with the speaker as the subject.
	Examples	"I'm trying to think.", "Well, seeing as how I'm a musician, I like all kinds of music.", "I use just a wood sealer."
na3-3	Type	Leaf
	Associated DAs	Affirmative non-yes answer (49.1%), statement-non-opinion (46.2%)
	Description	Statements that form a non-yes/no answer to a yes-no-question.
	Examples	"Um, we got two cats.", "I did.", "I'm a sports fan.", "Well, we've been married for five and a half years."
nd3	Type	Leaf
	Associated DAs	Dispreferred answer (71.6%), statement-non-opinion (28.4%)
	Description	Statements that usually start with "well" and form a non-yes/no answer to a yes-no-question.
	Examples	"Well, I don't have solutions to the problems.", "Well, it's, a lot of fun, at the moment.", "Well, actually, you have to put them in there unsmashed."
ng3	Type	Leaf
	Associated DAs	Negative non-no answer (64%), statement-non-opinion (30.3%)
	Description	Negated statements that form a non-no answer to a yes-no-question.
	Examples	"Not really.", "I haven't seen it in years.", "It's not my favorite.", "I don't have any children."
no3	Type	Leaf
	Associated DAs	Other answer (88.9%), statement-non-opinion (1.1%)

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Description	Other non-yes/no answers to yes-no-questions.
	Examples	"I have no idea.", "It really depends on who shows up.", "Uh, I don't know the exact numbers.", "That will work out just fine."
qw2-3	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sv
	Merged DAs	qw^d merged with qw; qy^d with qy
	Children	**qw3, sd3-6**
qw3	Type	Leaf
	Associated DAs	Wh-question (47.6%), declarative wh-question (46.4%)
	Description	Wh-questions in less conventional forms.
	Examples	"Well, I'm still puzzled though, what is the argument.", "I'm not sure where you're from.", "You know, what's the trade off there."
sd3-6	Type	Leaf
	Associated DAs	Statement-non-opinion (90%)
	Description	Statements that often contain a wh-word or "don't know".
	Examples	"I do not know what they are called.", "Well, uh, when I was in college.", "I wonder why.", "My undergraduate degree was not in what my master's was."
qy2-5	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sv
	Merged DAs	qw^d merged with qw; qy^d with qy
	Children	**qy3, sd3-7**
qy3	Type	Leaf
	Associated DAs	Declarative yes-no-question (55.9%), yes-no-question (36.6%)
	Description	Yes-no-questions.
	Examples	"Uh, do you cook for yourself.", "Um, have you been singing a long time.", "I don't know if you've tried it.", "Oh, so they don't go camping with you."
sd3-7	Type	Leaf
	Associated DAs	Statement-non-opinion (58.9%), declarative yes-no-question (19.9%), yes-no-question (6.4%), reformulate (5.4%)
	Description	Statements, sometimes with interrogative elements.
	Examples	"So you like a variety.", "My mother did not like cats.", "Had to mess with my phone here."
sv2-2	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	None
	Children	**o3-2, sd3-8**
o3-2	Type	Leaf
	Associated DAs	Statement-non-opinion (82.8%), statement-opinion (13.4%)
	Description	Medium to long statements. The distinction with sd3-8 is not really clear, so this division may not be all that useful.
	Examples	"And it had broken loose enough to where if it got hot", "And so I bought a ninety, um, because I really liked my eighty-eight.", "We always lived away from our family and relatives while the kids were growing up."
sd3-8	Type	Leaf
	Associated DAs	Statement-non-opinion (98.8%)
	Description	Medium to long statements. The distinction with o3-2 is not really clear, so this division may not be all that useful.
	Examples	"I mean he started off as a stray.", "But then I just thought the food was over priced for what it was.", "They put a lot of pressure on him from the outside and from the inside."
sv1	Type	Internal
	Clustering Method	Classifier

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b, b^m, bf, qh, sv, o
	Merged DAs	am and ar merged with aa; ^h, nd, ng and no with na; qw^d with qw; qy^d with qy
	Children	**aa2-4, ad2-4, ba2-3, fc2-4, h2-3**, na2-3, **qw2-4**, qy2-6, **sd2-3**
aa2-4	Type	Leaf
	Associated DAs	Statement-opinion (57.9%), accept (27%)
	Description	A group of loosely related statements that sometimes have an affirmative function.
	Examples	"It sounds like it.", "I would think so.", "It sounds like Iowa or something.", "Nobody could figure it out."
ad2-4	Type	Leaf
	Associated DAs	Action-directive (63.9%), statement-opinion (27.8%)
	Description	Statements that directly target the addressee through the use of "you".
	Examples	"You really should go to Europe.", "Well, it might be a good time for you to start a tradition.", "But you have to look out for yourself you know.", "Well, you would be interested in it then."
ba2-3	Type	Leaf
	Associated DAs	Statement-opinion (58%), appreciation (32.5%)
	Description	Statements that express some degree of appreciation for something.
	Examples	"That's scary.", "Um. Sounds good.", "I think that's really important.", "Which is probably pretty nice."
fc2-4	Type	Leaf
	Associated DAs	Conventional-closing (95.3%)
	Description	Opinions expressed during the closing sequence of the conversation.
	Examples	"Well, I think that covers it.", "It's sort of funny", "You couldn't get a job to save yourself over there."
h2-3	Type	Leaf
	Associated DAs	Hedge (77.3%), statement-opinion (22.7%)
	Description	Statements that function as hedges.
	Examples	"Let's put it that way.", "Uh, that may be wrong.", "I don't think"
qw2-4	Type	Leaf
	Associated DAs	Wh-question (64.3%), declarative wh-question (28.6%)
	Description	Wh-questions that ask for an opinion on something.
	Examples	"But what do you think", "Well, from your point of view, how would you feel about actually sending someone that's, you know, means something to you to one of those homes.",
sd2-3	Type	Leaf
	Associated DAs	Statement-opinion (75.4%), statement-non-opinion (20.4%)
	Description	The remaining statements, both opinion and non-opinion, although they often contain a personal viewpoint.
	Examples	"So that's interesting.", "And I think that would be just the perfect family car.", "It seems to be made out of something different."
na2-3	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh, sv
	Merged DAs	None
	Children	**^h3-2, na3-4, nd3-2, ng3-2, no3-2**
^h3-2	Type	Leaf
	Associated DAs	Hold before answer (71.9%), statement-opinion (28.1%)
	Description	Statements that precede an answer.
	Examples	"Trying to think of the name of it.", "That's a good question.", "There's different ways to do it."
na3-4	Type	Leaf
	Associated DAs	Affirmative non-yes answer (57.9%), statement-opinion (35.5%)
	Description	Affirmative non-yes answers to a yes-no-question that often contain "I think".

Continued on next page

Table C.1 – continued from previous page

Node	Characteristics	
	Examples	"Pretty much.", "I think it's a very good thing.", "Oh, excellent, excellent special effects."
nd3-2	Type	Leaf
	Associated DAs	Dispreferred answer (61.5%), statement-opinion (38.5%)
	Description	Answers to yes-no-questions that generally start with "well" and often contain "I think" and "actually".
	Examples	"Well I think it made parts of it a lot easier.", "Actually, it would be worth it.", "Well, it really isn't too big yet."
ng3-2	Type	Leaf
	Associated DAs	Negative non-no answer (86.2%), statement-opinion (10.3%)
	Description	Negative non-no answers to yes-no-questions.
	Examples	"I don't think so.", "Oh, you can't,", "Not that I can think of."
no3-2	Type	Leaf
	Associated DAs	Other answer (81.3%), statement-opinion (18.8%)
	Description	Answers to yes-no-questions that do not fit the other groups.
	Examples	"Well, maybe it would.", "Well, uh, it's really both.", "Uh, well, it would depend on when you go."
qy2-6	Type	Internal
	Clustering Method	Classifier
	Features	First nine words/POS tags, length, presence of question mark, previous speaker and DA, n-grams, POS n-grams
	Excluded DAs	^2, ^q, b^m, bf, qh
	Merged DAs	qw^d merged with qw; qy^d with qy
	Children	**qy3-2, sv3-2**
qy3-2	Type	Leaf
	Associated DAs	Yes-no-question (48.6%), declarative yes-no-question (45.8%)
	Description	Yes-no-questions that ask for an opinion on something.
	Examples	"I mean, do you think things are going to change.", "Isn't that a good feeling.", "But wouldn't it be wonderful."
sv3-2	Type	Leaf
	Associated DAs	Statement-opinion (59.9%), declarative yes-no-question (27%), reformulate (5.6%)
	Description	Statements that are often loosely interpretable as (declarative) yes-no-questions.
	Examples	"I'm guessing you probably do.", "Well, you must have a relatively clean conscience then.", "Oh, you guys sound pretty self sufficient.", "You want to do something for them."

Bibliography

Adolphs, Peter et al. (2008). "Some Fine Points of Hybrid Natural Language Parsing". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 1380–1387.

Andernach, Toine, Mannes Poel, and Etto Salomons (1997). "Finding classes of dialogue utterances with kohonen networks". In: *Workshop Notes of the ECML / MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, pp. 85–94.

Anderson, Anne H et al. (1991). "The HCRC map task corpus". In: *Language and speech* 34.4, pp. 351–366.

Ang, J., Yang Liu, and E. Shriberg (2005). "Automatic Dialog Act Segmentation and Classification in Multiparty Meetings". In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* Vol. 1. IEEE, pp. 1061–1064.

Austin, John Langshaw (1962). *How to do things with words*. Oxford University Press.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe (1998). "The Berkeley FrameNet Project". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. Montreal, Quebec, Canada: Association for Computational Linguistics, pp. 86–90.

Boyer, Kristy Elizabeth et al. (2010). "Dialogue Act Modeling in a Complex Task-oriented Domain". In: *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL '10)*. Tokyo, Japan: Association for Computational Linguistics, pp. 297–305.

Cambria, Erik and Bebo White (2014). "Jumping NLP Curves: A Review of Natural Language Processing Research". In: *IEEE Computational Intelligence Magazine* 9.2, pp. 48–57.

Carletta, Jean et al. (1997). "The Reliability of a Dialogue Structure Coding Scheme". In: *Computational Linguistics* 23.1, pp. 13–31.

Chiticariu, Laura, Yunyao Li, and Frederick R. Reiss (2013). "Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!" In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 827–832.

Core, Mark G and James Allen (1997). "Coding dialogs with the DAMSL annotation scheme". In: *AAAI fall symposium on communicative action in humans and machines*. AAAI, pp. 28–35.

Das, Dipanjan et al. (2014). "Frame-semantic parsing". In: *Computational Linguistics* 40.1, pp. 9–56.

Dignum, Virginia and Frank Dignum (2015). "Contextualized Planning Using Social Practices". In: *Coordination, Organizations, Institutions, and Norms in Agent Systems X*. Springer, pp. 36–52.

Fernandez, Raul and Rosalind W Picard (2002). "Dialog act classification from prosodic features using support vector machines". In: *Speech Prosody 2002, International Conference*, pp. 291–294.

Ferreira, Fernanda and Nikole D Patson (2007). "The 'Good Enough' Approach to Language Comprehension". In: *Language and Linguistics Compass* 1.1-2, pp. 71–83.

Forsythand, E. N. and C. H. Martell (2007). "Lexical and Discourse Analysis of Online Chat Dialog". In: *International Conference on Semantic Computing (ICSC 2007)*. IEEE, pp. 19–26.

Fromkin, Victoria, Robert Rodman, and Nina Hyams (2013). *An introduction to language*. tenth. Cengage Learning.

Gambäck, Björn, Fredrik Olsson, and Oscar Täckström (2011). "Active learning for dialogue act classification". In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, pp. 1329–1332.

Gisladottir, Rosa S et al. (2012). "Speech act recognition in conversation: experimental evidence". In: *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, pp. 1596–1601.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). "SWITCHBOARD: telephone speech corpus for research and development". In: *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. Vol. 1. IEEE, pp. 517–520.

Janin, A. et al. (2003). "The ICSI Meeting Corpus". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*. Vol. 1. IEEE, pp. I–364–I–367.

Ji, Gang and J. Bilmes (2005). "Dialog Act Tagging Using Graphical Models". In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* Vol. 1. IEEE, pp. 33–36.

Jia, Jiyou (2004). "The study of the application of a web-based chatbot system on the teaching of foreign languages". In: *Proceedings of Society for Information Technology & Teacher Education International Conference*. Atlanta, GA, USA: Association for the Advancement of Computing in Education (AACE), pp. 1201–1207.

Jurafsky, Daniel and James H. Martin (2008). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. second. Prentice Hall.

Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin (2010). "Classifying Dialogue Acts in One-on-one Live Chats". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 862–871.

Kingsbury, Paul and Martha Palmer (2002). "From TreeBank to PropBank." In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Spain: European Language Resources Association (ELRA), pp. 1989–1993.

Kipper-Schuler, Karin (2005). "VerbNet: A broad-coverage, comprehensive verb lexicon". PhD thesis. University of Pennsylvania.

Klüwer, Tina, Hans Uszkoreit, and Feiyu Xu (2010). "Using Syntactic and Semantic Based Relations for Dialogue Act Recognition". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*

(COLING '10). Beijing, China: Association for Computational Linguistics, pp. 570–578.

Larsson, Staffan and David R. Traum (2000). "Information state and dialogue management in the TRINDI dialogue move engine toolkit". In: *Natural language engineering* 6.3&4, pp. 323–340.

Margolis, Anna, Karen Livescu, and Mari Ostendorf (2010). "Domain adaptation with unlabeled data for dialog act tagging". In: *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, pp. 45–52.

Màrquez, Lluís et al. (2008). "Semantic role labeling: an introduction to the special issue". In: *Computational linguistics* 34.2, pp. 145–159.

McTear, Michael F (2002). "Spoken dialogue technology: enabling the conversational user interface". In: *ACM Computing Surveys (CSUR)* 34.1, pp. 90–169.

Moldovan, Cristian, Vasile Rus, and Arthur C Graesser (2011). "Automated Speech Act Classification For Online Chat." In: *22nd Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2011)*. MAICS, pp. 23–29.

Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman (2011). "Natural language processing: an introduction". In: *Journal of the American Medical Informatics Association* 18.5, pp. 544–551.

Omuya, Adinoyi, Vinodkumar Prabhakaran, and Owen Rambow (2013). "Improving the Quality of Minority Class Identification in Dialog Act Tagging". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, pp. 802–807.

Petukhova, Volha and Harry Bunt (2011). "Incremental Dialogue Act Understanding". In: *Proceedings of the Ninth International Conference on Computational Semantics (IWCS '11)*. Association for Computational Linguistics, pp. 235–244.

Rus, Vasile et al. (2012). "Automated Discovery of Speech Act Categories in Educational Games." In: *International Educational Data Mining Society*.

Samei, Borhan et al. (2014). "Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings". In: Springer. Chap. Context-Based Speech Act Classification in Intelligent Tutoring Systems, pp. 236–241.

Searle, John R (1976). "A classification of illocutionary acts". In: *Language in society* 5.1, pp. 1–23.

Serafin, Riccardo and Barbara Di Eugenio (2004). "FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification". In: *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics (ACL '04)*. Association for Computational Linguistics.

Shriberg, Elizabeth et al. (2004). "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus". In: *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*. Cambridge, Massachusetts, USA: Association for Computational Linguistics, pp. 97–100.

Sridhar, Vivek Kumar Rangarajan, Srinivas Bangalore, and Shrikanth Narayanan (2009). "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging". In: *Computer Speech & Language* 23.4, pp. 407–422.

Stolcke, Andreas et al. (2000). "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". In: *Computational Linguististics* 26.3, pp. 339–373.

Surendran, Dinoj and Gina-Anne Levow (2006). "Dialog act tagging with support vector machines and hidden Markov models." In: *INTERSPEECH-2006*, pp. 1950–1953.

Tavafi, Maryam et al. (2013). "Dialogue act recognition in synchronous and asynchronous conversations". In: *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013)*. Vol. 13. Association for Computational Linguistics, pp. 117–121.

Verbree, D., R. Rienks, and D. Heylen (2006). "Dialogue-act tagging using smart feature selection; results on multiple corpora". In: *2006 IEEE Spoken Language Technology Workshop*. IEEE, pp. 70–73.

Webb, Nick and Michael Ferguson (2010). "Automatic Extraction of Cue Phrases for Cross-corpus Dialogue Act Classification". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*. Association for Computational Linguistics, pp. 1310–1317.

Webb, Nick, Mark Hepple, and Yorick Wilks (2005). "Dialogue act classification based on intra-utterance features". In: *Proceedings of the AAAI Workshop on Spoken Language Understanding*. AAAI.

Wu, Tianhao et al. (2005). "Posting Act Tagging Using Transformation-Based Learning". In: *Foundations of Data Mining and knowledge Discovery*. Ed. by Tsau Young Lin et al. Springer, pp. 319–331.