# Detection of Agreement and Disagreement: An investigation of linguistic coordination and conversational features

Maria Alexandropoulou

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington
2014

Committee:
Gina-Anne Levow
Michael Tjalve

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Detection of Agreement and Disagreement: An investigation of linguistic coordination and conversational features

Maria Alexandropoulou

Chair of the Supervisory Committee:

Gina-Anne Levow, Assistant Professor

Department of Linguistics

The focus of this thesis is detection of agreement and disagreement in multiparty conversations using existing transcripts from the ICSI corpus. We use an unsupervised lexicon-based method to create our baseline and then follow a supervised approach to study the effect on performance of different feature sets. The feature sets we are interested in are the following: a. lexicosyntactic, b. Dialog act tag-based features, c. Linguistic style coordination features and d. Conversational features (baseline, individual and non-individual ones). The results enabled us to study the presence of coordination in agreeing and disagreeing statements and showed that the performance can improve when adding further features on top of the lexicosyntactic ones. Additionally, we saw that non-individual conversational features can improve the performance of agreement detection while individual conversational ones appear to have a similar effect on disagreement detection.

# Contents

# 1. Introduction

## 1.1. General Discussion

The research area this thesis focuses on is Opinion Mining and Sentiment Analysis. The availability of subjective and opinion-rich content in the World Wide Web has been increasing (personal blogs, online reviews etc). The analysis of subjective content can be very useful either for discovering opinions (e.g. political views) or because the views and experience of others can be useful for making decisions (e.g. product reviews and recommendations). Based on the above, the need for computational systems that can leverage these abundant online resources and provide people with the information they need has emerged and triggered a significant amount of research in this area.

The area of Opinion Mining and Sentiment Analysis includes several problems that researchers have been working on (Pang & Lee 2008). Discovering "Sentiment Polarity and Degrees of Positivity" (Pang & Lee 2008) involves processing a piece of text where opinions about a specific subject are expressed and the goal is determining the general sentiment towards the subject (positive/negative) and discovering the position of this text in the area between the two polarities. While the previous problems assume that the piece of text they are examining contains opinions related to a specific subject, "Subjectivity Detection and Opinion Identification problem" (Pang & Lee 2008) tries to determine whether the input contains subjective content or not, as well as to isolate the subjective portions. In some scenarios we do not know in advance the topic of the text we are processing, which is the case, for example, of text that includes opinions on many different subjects. In these cases, the problem of "Joint Topic-Sentiment Analysis" (Pang & Lee 2008) arises where both the type of the sentiment as well as the target of the sentiment need to be discovered. Discovering "Viewpoints and Perspectives" (Pang & Lee 2008) is another issue in the area, which analyzes sentiment in politically oriented text and tries to determine the general attitude and ideological stance expressed by it. The difference is that in this case, the text is not about specific targets

for which opinions are trying to be extracted, but cues related to the general ideology of the text are being looked for. Finally, discovering "Other Non-Factual Information in Text" (Pang & Lee 2008) has been a research problem and it involves analysis of emotions like anger, disgust, fear, happiness, humor recognition, genre identification etc.

## 1.2. Motivation

The current thesis focuses on the detection of agreement and disagreement in multiparty conversations. Literature was researched for various techniques used in sentiment analysis and a subset of them was employed for solving the problem of agreement/disagreement detection. In our study we used a subset of the meeting recordings collected and transcribed by ICSI which had been annotated for agreement/disagreement (Hillard, Ostendorf & Shriberg 2003).

As outlined in (Wilson 2008) applications such as meeting browsers and meeting assistants can greatly benefit from sentiment analysis research in spoken multiparty conversations. Apart from the extraction of objective information like the list of topics discussed or the items assigned to different people, subjective information can be extremely useful as well. Decision detection could benefit from it while a summary of what opinion was supported by whom is also interesting. Our task of agreement/disagreement detection aims at contributing to the general task of performing sentiment analysis in meetings.

The problem of agreement/disagreement detection, as it happens with similar problems of opinion mining, is not a trivial one that can be solved just by relying on lexical cues. In the following examples you can see that the lexical cue "yeah" can be encountered when expressing agreement, disagreement or when back-channeling [1] :

[1] Quoting from page 48 in (Dhillon, Bhagat, Carvey & Shriberg 2004): "Utterances which function as backchannels are not made by the speaker who has the floor. Instead, backchannels are utterances made in the background that simply indicate that

*(the examples have been taken from the manually annotated agreement/disagreement corpus described in (Hillard, Ostendorf & Shriberg 2003))*

**Example 1:**

…

```
Speaker c3: okay so that's good i- i think in the sense that i think andreas meant
the question right
Speaker c1: that's that's good yeah because the overall rate is
Speaker cA: yeah
Speaker c2: hmph
Speaker cB: statistical
Speaker c1: yeah
Speaker cB: yep
Speker cA: other- otherwise you'd get double counts here and there and then it
would be harder
Speaker c4: yeah
Speaker c1: uh but yeah <- disagreement
Speaker c3: yeah
Speaker c4: yeah
Speaker c8: i should also say i did a simplifying uh count in that if a was
speaking b overlapped with a and then a came back again and overlapped with b
again i i didn't count that as a three person overlap i counted that as a two
person overlap and it was a being overlapped with by d
Speaker cB: uhhuh
Speaker c8: because the idea was the first speaker had the floor and the second
person started speaking and then the f- the first person reasserted the floor kind
of thing these are simplifying assumptions didn't happen very often there may be
like three overlaps affected that way in the whole thing
Speaker c1: yeah
```

…

**Example 2:**

…

```
Speaker c1: well not to mention the fact that i would be
Speaker c2: yeah yeah
Speaker c1: hesitant certainly to take anyone under eighteen probably even anyone
under twenty one so
Speaker c3: oh you ageist
Speaker c1: what's that
Speaker c4: age-ism
Speaker c1: well ageist the eighteen is because of the consent form we'd have to
get find their parent to sign for them
Speaker c9: right yeah
```

---

a listener is following along or at least is yielding the illusion that he is paying attention. When uttering backchannels, a speaker is not speaking directly to anyone in particular or even to anyone at all.

Common backchannels include the following: "uhhuh", "okay", "right", "oh", "yes", "yeah", "oh yeah", "uh yeah", "huh", "sure", "hm.". “

```
Speaker c2: yeah ageist yeah yeah  <- agreement
 …
```

**Example 3:**

```
…
Speaker c3: there are there are different types and within those types like as
jose was saying that sounded like a backchannel overlap meaning the kind that's a
friendly encouragement like uhhuh great yeah
Speaker c4: yeah <- back-channel
Speaker c3: and it doesn't take you don't take the floor um but some of those as
you showed i think can be discriminated by the duration of the overlap
Speaker c4: yeah
 …
```

As will be mentioned later in the literature survey section, many different techniques have been employed in the area of sentiment analysis apart from the ones using subjectivity lexicons. One example is the use of word n-grams, character n-grams and phoneme n-grams in (Raaijmakers & Wilson 2008). However, the use of a subjectivity lexicon is going to be our baseline. Additionally, we will evaluate feature sets that have been used for subjectivity and polarity analysis by other researchers to see their effect on agreement/disagreement detection (for example we will use lexicosyntactic and conversational features that have been used in other studies e.g. (Murray & Carenini 2011)). The promising results for polarity analysis in (Murray & Carenini 2011) provided us with the motivation to use conversational features while at the same time generated the interest to extend these features and to examine how the deviation from a speaker's average conversational style could influence the results. Furthermore, (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012) led us to examine the relationship between linguistic style coordination and expression of agreement/disagreement. Quoting from pages 1-2 of (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012): "Language coordination is a phenomenon in which people tend to unconsciously mimic the choices of function-word classes made by the people they are communicating with; roughly speaking, if you are communicating with someone who uses a lot of articles — or prepositions, or personal pronouns — then you will tend to increase your usage of these types of words as well, even if you don't

consciously realize it.”. Last but not least, the corpus we used was a subset of the ICSI Meeting Recorder Dialog Act (MRDA) corpus that had already been annotated with dialog act tags (Shriberg, Dhillon, Bhagat, Ang & Carvey 2004). Therefore, we decided to examine if leveraging the existing annotations could contribute to our agreement/disagreement detection task.

Our results showed that there is a relationship between linguistic style coordination and agreement/disagreement. More specifically, the probability of coordination appears to be higher in disagreeing spurts. Also, dialog act tags seemed to be informative and contribute to our task as well. However, using lexicosyntactic features by themselves provides a very strong baseline that can be improved to some extent by the aforementioned coordination and dialog tag based features.

# 2. Literature survey

## 2.1. Introduction to opinion mining and sentiment analysis

(Pang & Lee 2008) gives an overview of the research done in the field of Opinion Mining and Sentiment Analysis. It summarizes the major problems in the field as well as the techniques that have been employed to deal with those issues. Several supervised learning approaches have been used. Diverse linguistic features are used in these algorithms, including:

a. Term presence and/or term frequency
b. Position of certain tokens, higher order n-grams, construction of sentiment lexicons and techniques that define relationships among lexical features for polarity classification.
c. Use of part of speech information of words
d. Syntax information
e. Identifying use of negation since it can lead to polarity reversal.
f. Use of topic-oriented features

Also, unsupervised learning methods have been used e.g. to create lexicons to identify polarity in text or generate labeled data. Some of them are based on assumptions regarding the relationship between the use of frequent or rare words and the expression of an opinion or sentiment.

The above survey served as an introduction to this research area and gave a general overview of the breadth of features and methods used. Also, in our research we used among others the feature categories described in a-c above. Last but not least, our baseline is created using an existing subjectivity lexicon and applying a simple unsupervised method for agreement/disagreement detection.

(Hu & Liu 2004) created a system for automatically extracting information from reviews of a product and summarizing the results. The method they followed consisted of three steps: 1. Determining the aspects of a product that the customers have talked about, 2. For each aspect, discovering the positive or negative sentences referring to it and 3. Summarizing the information extracted. Adjectives were the only words that were taken into account for deciding if a sentence is subjective or not and for determining its polarity. An existing polarity/subjectivity dictionary was not used but WordNet was leveraged instead. More specifically, an initial list of seed adjectives with known polarities was constructed. For any adjective they encountered, they searched within WordNet to determine if the synsets (synset: set of synonyms for a certain sense of a word) of the adjective and its antonym contains any adjectives in the seed list. This way the polarity of the adjective was inferred and it was added to the seed list. For predicting the polarity of a sentence the positive and negative adjectives were counted and whichever of the two classes prevailed determined the polarity of the sentence. For our baseline experiments and in order to determine if a sentence expresses agreement/disagreement we followed a similar method except for the fact that we used an existing subjectivity/polarity dictionary.

## 2.2. Approaches to subjectivity recognition in conversational speech

(Wilson 2008) outlines an annotation scheme used for subjectivity and polarity analysis for the AMI[2] corpus (multiparty conversation corpus). This paper does not describe any specific machine learning method used since it refers to the first stage of the process which is annotation of the data. However, it helped understanding the work performed on this corpus which is described in subsequent papers.

[2] AMI: Augmented Multiparty Interaction

(Raaijmakers & Wilson 2008) perform subjectivity identification in the AMI corpus while comparing different features used for classification. The features used are word n-grams, character n-grams and phoneme n-grams all coming from either the reference transcription or from automatic speech recognition. Six experiments are run (one for each feature set). The performance is always higher when the reference transcriptions are used compared to the ASR transcriptions. Additionally, both the phoneme and character n-grams yield higher recall than the word n-grams (apart from the case of automatically recognized phoneme n-grams). The best results in terms of recall and F1 score were given by the character n-grams from the reference transcriptions

(Raaijmakers, Truong & Wilson 2008) perform subjectivity recognition and polarity classification using two types of features: lexical and acoustic. The method is applied to the AMI corpus and the features used were: character, word and phoneme n-grams, as well as pitch, energy and the distribution of energy in the long-term averaged spectrum (The last three types of features will be referenced as prosodic features from now on. Also the first three types of features were extracted from manually created transcriptions). Four classifiers are trained, one with each one of the following features: word n-grams, phoneme n-grams, character n-grams, prosodic features. The results of the four single source classifiers are combined using a linear interpolation strategy. Several experiments were performed and the results showed that character n-grams outperform prosodic features, word n-grams and phoneme n-grams in subjectivity recognition and polarity classification. Prosodic information seems to be the least informative features while character-level information is very important. A combination of prosodic, word-level, character-level and phoneme-level information gives the best performance for subjectivity recognition, while for polarity classification this happens through the combination of words, characters and phonemes.

(Murray & Carenini 2011) perform polarity and subjectivity analysis in written (e-mails) and spoken (AMI meeting corpus) conversations. They focus on four classification tasks:

a. Classification of subjective utterances

b. Classification of all subjective phenomena including subjective questions

c. Classification of just positive-subjective utterances

d. Classification of just negative-subjective utterances

The set of features they experimented with are the following:

a. Varying instantiation trigrams

A varying instantiation trigram consists of a mixture of words and POS tags and in the context of the aforementioned survey it is a pattern for recognizing positive and negative subjective sentences. A specific word trigram gives eight varying instantiation trigrams (patterns). First of all, the authors collected positive/negative subjective patterns and subjective/non-subjective patterns (one pattern set for each of the four classes). They used the annotated corpus and extracted varying instantiation trigrams. Then these trigrams were ranked based on the probability P(relevance | trigram) for each class. They used a probability threshold to keep the most important patterns. Apart from the already annotated corpora, they additionally learned patterns using an unsupervised method. They extracted subjective patterns by using a corpus consisting of blog posts which were assumed to be all subjective. The non-subjective patterns were extracted from online newswire data and they were assumed to be all objective. Similarly to the approach followed with the annotated corpus, they extracted subjective and non-subjective patterns, ranked them using the probability described above and filtered them using again a cut-off probability.

The way they leveraged the above patterns for feature generation, was by splitting the pattern set for each class into bins based on the probability value (as defined in the previous paragraph) that was

calculated for each pattern. Then the features generated for each sentence consisted of the number of patterns it contained per bin.

b. Raw Feature set

The raw feature set includes the following features that were extracted from each sentence:

1. Character trigrams
2. Word bigrams
3. POS bigrams
4. Word pairs (skip word bigrams with any number of words in-between)
5. POS pairs (skip POS bigrams with any number of part of speech tags in-between)
6. Varying instantiation bigrams

c. Conversational features

To better explain the conversational features they used they give the following two definitions:

A. $S_{prob}(t) = max_S p(S|t)$ where $p(S|t)$ expresses the probability of participant S given the word t. The motivation behind this is that certain words are more likely to be used by some speakers than by others.

B. $T_{prob}(t) = max_T p(T|t)$ where $p(T|t)$ expresses the probability of turn T given the word t.

For each sentence, the following features were extracted:

| Feature ID | Description |
|---|---|
| **MXS** | max *Sprob* score |
| **MNS** | mean *Sprob* score |
| **SMS** | sum of *Sprob* scores |
| **MXT** | max *Tprob* score |
| **MNT** | mean *Tprob* score |
| **SMT** | sum of *Tprob* scores |
| **TLOC** | position in turn |
| **CLOC** | position in conv. |
| **SLEN** | word count, globally normalized |
| **SLEN2** | word count, locally normalized |
| **TPOS1** | time from beg. of conv. to turn |
| **TPOS2** | time from turn to end of conv. |
| **DOM** | participant dominance in words |
| **COS1** | cosine of conv. splits, w/ *Sprob* |
| **COS2** | cosine of conv. splits, w/ *Tprob* |
| **PENT** | entropy of conv. up to sentence |
| **SENT** | entropy of conv. after the sentence |
| **THISENT** | entropy of current sentence |
| **PPAU** | time btwn. current and prior turn |
| **SPAU** | time btwn. current and next turn |
| **BEGAUTH** | is first participant (0/1) |
| **CWS** | rough ClueWordScore (cohesion) |
| **CENT1** | cos. of sentence & conv., w/ *Sprob* |
| **CENT2** | cos. of sentence & conv., w/ *Tprob* |

*Table 1: Table taken from Murray & Carenini 2011 – page 11*

It was shown that by using a set of lexicosyntactic and conversational features high performance is achieved without the need for features specific to a particular conversational modality (e.g. prosodic features for speech or headers for the e-mail conversations). The experiments demonstrate the importance of using conversational features.

(Wang & Liu 2012) describe two unsupervised methods for subjectivity analysis that they tried on data from three different domains: movies, news articles and meeting dialogs. The data for the last one came from the AMI corpus. Their focus is sentence level classification but no context information is used for

each sentence, i.e. it is not assumed that sentences in the data are consecutive. The iterative learning methods used were: self-training and calibrated EM (expectation-maximization).

### 2.2.1. Opinion frames and discourse relationships

(Ruppenhofer, Somasundaran & Wiebe 2008a) suggest the use of opinion frames in order to express discourse-level relationships between opinions. An opinion frame associates two text spans expressing opinions and the targets they refer to. An opinion can be positive or negative and it can be perceived either as sentiment expression or argument. Given the above as well as the fact that the target of two opinions can be either the same or alternative, there are 32 types of opinion frames. The benefit of the above approach is that discourse information can be used to identify opinion frames and these results can later be used in order to classify individual opinions. The above paper describes the annotation scheme and how it was evaluated when used in the AMI corpus, although it does not perform any experiments identifying these frames automatically and using them for opinion recognition.

(Ruppenhofer, Somasundaran & Wiebe 2008b) build on top of the opinion frame approach described in the previous paragraph. The paper uses a machine learning approach in order to determine if two text spans that express opinions participate in any kind of opinion frame (two opinions create an opinion frame when their targets are the same or when they are mutually exclusive). The AMI corpus is used.

(Somasundaran, Namata, Wiebe & Getoor 2009) try to prove the usefulness of discourse-level information as modelled using the opinion frame analysis described above. This research considers that opinion frames are known in advance and focuses on different approaches to perform polarity analysis leveraging this information. The two different methods used are (a) iterative collective classification as well as (b) integer linear programming. Additionally, a combination of the above methods is used since the first algorithm seems to perform better for opinions not belonging to opinion frames while the second one works better for the remaining ones. The AMI corpus was used for this study as well.

(Somasundaran, Namata, Getoor & Wiebe 2009) also use the opinion frame model to perform polarity classification. For this task they use Collective Classification Framework which consists of the following three classifiers: an instance polarity classifier (IPC), a target-link classifier (TLC) and a frame-link classifier (FLC). The first one classifies a text span as being positive, negative or neutral. The second one determines if two text spans have same or alternative targets while the third one decides if two spans participate in any frame and if this frame is reinforcing or non-reinforcing. A frame is reinforcing if the two opinions it includes reinforce each other, while it is non-reinforcing in the opposite case. The AMI corpus was used for this study.

The above papers presented in detail several techniques and feature sets used for subjectivity and polarity analysis using conversational features. This gave us some insight into the different methods applied in the area of sentiment analysis and as a result we ended up leveraging some of these methods and features for our agreement/disagreement analysis task. More specifically we used a supervised approach while our feature set included some of the lexicosyntactic and conversational features mentioned in the above papers, especially in Murray & Carenini 2011. At the same time, we extended our feature set by adding features based on linguistic coordination and dialog act annotations as well as per speaker conversational features.

## 2.3. The ICSI Meeting Recorder Corpus: Recognizing Agreement and Disagreement

(Shriberg, Dhillon, Bhagat, Ang & Carvey 2004) discuss the introduction of the ICSI Meeting Recorder Dialog Act (MRDA) corpus and the annotation scheme they used. Their effort included marking the boundaries of DA segments, annotating them and determining the adjacency pairs. The ICSI Meeting corpus includes 75 meetings and 53 speakers while there are 6 speakers per meeting on average.

(Hillard, Ostendorf & Shriberg 2003) employ different methods for automatically detecting agreement and disagreement in a subset of the ICSI Meeting Recorder corpus. They hand-labeled part of it for

agreement/disagreement and they used prosodic features as well as features based on word types, word counts and n-gram scores. The prosodic features include pause, fundamental frequency and duration. Also, in order to construct their feature set they created bigram language models (both word and part-of-speech class language models) for each one of the four classes (Pos, Neg, Other, BackChan). For each sentence and using the above models they calculated the probability that the sentence belongs to each one of the classes and this information was included in their features. To create the language models they used two approaches: 1. A supervised one using the manually annotated data and 2. An unsupervised one that starts using a dictionary approach for classification and then iteratively leads to the creation of the language models for each class. The unsupervised approach made it possible to leverage a lot of unlabeled data and resulted in satisfactory results (even when working with ASR transcripts). Additionally, prosodic and word-based features led to similar performance for the ASR transcripts.

The hand-labeled portion of the data mentioned in the previous paragraph was also used during the experiments we performed for agreement-disagreement detection. Additionally, the existing Dialog Act annotations were leveraged in order to construct our feature set as will be explained later.

## 2.4. Coordination and Entrainment

(Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012) studies the relationship between linguistic style coordination and power relationships in group conversations (written and spoken language). Linguistic style is defined based on a speaker's frequency of usage of the following functional categories: articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers. When the linguistic style of a speaker's reply is similar to the linguistic style of the speaker they are replying to and when it is deviating from their personal style, then linguistic style coordination is present. The ultimate goal of this study is to test the following hypotheses:

1. People in general coordinate more towards high-powered people than towards low-powered people.
2. High-powered people coordinate less than low-powered people towards their targets.
3. People have a baseline coordination level which is determined by personal characteristics.

The data used for this study consists of discussions between Wikipedia editors as well as arguments taking place before the U.S. Supreme Court. The results validated the above hypotheses for both sets of data. Additionally, the performance of different feature sets was measured regarding the task of predicting the status of a speaker given the transcript of what they said. The feature sets compared were the following and were extracted from the set of replies of speaker y to speaker x and vice versa:

1. For each function word class, binary features are constructed that express if x coordinates to y more than y to x.
2. The frequency of each functional category and the average utterance length in each set of replies.
3. The frequency of each word in each set of replies.

The experiments showed that only feature set 1 was able to give higher performance than the baseline for both corpora. As a result, it is the best one when working with cross-domain data. However, when examining data within a certain domain the other two feature sets can lead to better performance per domain. This is because certain lexical cues can reveal the status and role of the speaker e.g. in the Supreme Court context it is very common for lawyers to start their sentence with the phrase “Your honor”. Nevertheless, this method fails to generalize across domains.

The conclusions drawn from these experiments served as motivation to use linguistic coordination-related features for the current thesis where the main goal is agreement/disagreement detection. As will be mentioned later, we introduced a way to measure the degree of coordination between two consecutive turns. Specific definitions will be given in the “Methodology section”.

# 3. Methodology

## 3.1. On automatic recognition of agreement and disagreement

In this work, we are studying the detection of agreement and disagreement in multiparty conversations. Our goal is to determine if an utterance in the conversation expresses agreement or disagreement or if it is a back-channel. We are going to approach the issue as a supervised classification problem, and we will study the effect of different feature sets on the performance of the algorithm.

For our study, we used the agreement/disagreement corpus which is a small subset of the ICSI meeting recordings (Morgan et al 2001) that consists of 75 meetings. More specifically the portion that has been manually annotated consists of only the first 450 spurts of four meetings (Hillard, Ostendorf & Shriberg 2003). The annotators who labeled this data were viewing the transcripts at the same time as they were listening to the corresponding speech. These meetings were recorded at ICSI (Berkeley), and they were a subset of the Meeting Recorder project weekly meetings. The average number of speakers per meeting is 6, there are more male than female speakers per meeting (on average 33% of the participants of a meeting are female) and more native English speakers than non-native ones (on average 16% of the participants of a meeting are non-native English speakers).

For our experiments, we used the transcripts of the meetings. To begin with, we provide some definitions that will be useful to better understand the remainder of the analysis:

- **Turn**: A turn consists of the group of consecutive utterances that one person speaks before another one starts speaking. In this case, the second person initiates a new turn. To determine the boundaries of a turn, we relied on the meeting transcripts.

  See for example the turns in the following snippet from the meeting:

```
Speaker c3: oh that's optional                          <- 1st turn
Speaker c2: no that's okay                              <- 2nd turn
Speaker c0: okay so                                     <- 3rd turn
Speaker c3: you know                                    <- 4th turn
Speaker c0: this time the form
Speaker c0: discussion should be very short right       <- 5th turn
Speaker c2: it also should be later                     <- 6th turn
```

- **Spurt**: This is a period of speech by one speaker that has no pauses of greater than half a second. As will be seen later, the turn a spurt which belongs to can be seen as context for this spurt, and this logic is going to be used during the design of some of the classification features we chose. A spurt is also the unit that was already annotated for agreement/disagreement in ICSI corpus.

  For example each line in the turn example above is one spurt as defined in the corpus.

- **Dialog Act**: This is an utterance or portion of an utterance that expresses some specific function. For different types of dialog act, look at the tags described in Appendix B. A spurt can consist of many dialog acts or it can be a subset of a dialog act.

The following table explains more about the difference between dialog act and spurt. To better understand the table we will give the definitions of the general part of the dialog act tags you see there since the general tag is what we end up leveraging.

s: the dialog act is a statement

b: the dialog act is a backchannel

fg: the dialog act is a floor grabber

Each dialog act tag contains at least one general tag and may contain one or more specific tags, depending on the type of the utterance. The more specific tags are appended to the general tag using the symbol "^". For explanation of the specific tags found in the following examples see Appendix 9.2.

| **Snippet split in dialog acts.**<br><br>**The format is: Speaker ID, dialog act tag, content of dialog act** | **Same snippet split in spurts.**<br><br>**The format is: Speaker ID, content of spurt** |
|---|---|
| c2, s, and so you take that<br><br>c2, s, and then he's - he's uh measuring at the frame level<br><br>c2, s.%--, still at the frame level of what<br><br>c3, b, right<br><br>c2, s, and then - and then just uh normalizing with that larger amount<br><br>c2, s^rt, um and - but one thing he was pointing out is when he - he looked at a bunch of examples in log domain it is actually pretty hard to see the change<br><br>c2, s^rt, and you can sort of see that because of j- - of just putting it on the board that if you sort of have log x plus log x that's the log of x plus the log of two<br><br>c4, b, yep<br><br>c3, fg\|s^cs, yeah \| maybe it's not log distributed<br><br>c4, b, huh<br><br>c4, b, yeah<br><br>c2, s, and it's just you know it – it diminishes the effect of having two of them | c2, and so you take that and then he's he's uh measuring at the frame level still at the frame level of what and then and then just uh normalizing with that larger amount<br><br>c3, right<br><br>c2, um and but one thing he was pointing out is when he he looked at a bunch of examples in log domain it is actually pretty hard to see the change and you can sort of see that because of j- of just putting it on the board that if you sort of have log x plus log x that's the log of x plus the log of two and it's just you know it it diminishes the effect of having two of them<br><br>c4, yep<br><br>c3, yeah maybe it's not log distributed<br><br>c4, huh yeah |

*Table 2: Example snippet from one meeting that has been split into dialog acts (first column) as well as spurts (second column)*

The minimum unit of focus in our study (and as a result the unit that was the target of our classification task) was a spurt. The reason behind this choice is that the corpus we used for this study had already been manually annotated using spurts as units. Each spurt could either express agreement (its label is: “Pos”) or disagreement (its label is: “Neg”) or it could have been used as a backchannel (its label is: “BackChan”).

A spurt not belonging to any of the above three categories falls into the "Other" category (its label is: "Other"). The distribution of the four classes in the corpus is the following:

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 1114 |
| BackChan | 405 |

*Table 3: Distribution of the four classes in the corpus (Pos, Neg, Other, BackChan)*

## 3.2. Classification Model and Features

The way we decided to model our data was by splitting a meeting in turns and by using a linear chain CRF to model the sequence of spurts within a turn. This way we assumed relationships among adjacent spurts in a turn.ss For our experiments we followed an 8-fold cross validation approach, where in each classification task a different portion of the data was used as test data while the remaining was used as training data for this iteration.

For the agreement/disagreement/backchannel detection task we used four types of features for each spurt:

1. Word unigrams, bigrams
2. Lexicosyntactic features, influenced by (Murray & Carenini 2011)
3. Conversational features
4. Features based on existing Dialog Act annotations
5. Coordination features, influenced by (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012)

### 3.2.1. Word unigrams and bigrams

This is a very basic feature set that has been used in various sentiment analysis tasks. We will use it as an additional baseline and examine how the addition of more features influences the results.

### 3.2.2. Lexicosyntactic Features

This set of features was highly influenced by the lexicosyntactic features in (Murray & Carenini 2011). The above research focused on the subjectivity and polarity detection and we decided to leverage it for our agreement/disagreement detection work. A subset of the features in the above research was used:

1. Part of speech (POS) bigrams
2. Word bigrams
3. Part of speech skip bigrams (consecutive pairs of POS with only one POS in between)
4. Word skip bigrams (consecutive pairs of words with only one word in between)

Additionally, we also used unigrams as a feature since it seemed appropriate for our study (single words like “yes” or “no” in many cases can be good enough to indicate existence of agreement or disagreement).

For example for the spurt “This looks good” we generated the following lexicosyntactic features:

```
WordUnigrams.this WordUnigrams.looks WordUnigrams.good POSBigrams.DT.VBZ
POSBigrams.VBZ.JJ WordBigrams.this.looks WordBigrams.looks.good POSSkipBigrams.1.DT.JJ
WordSkipBigrams.1.this.good
```

### 3.2.3. Conversational features

A set of conversational features was also implemented for our experiments. Our initial hypothesis was that features such as the length of an utterance and the maximum word length can improve the performance of agreement/disagreement detection. This was based on the initial hypothesis that a

speaker uses longer words and utterances when disagreeing. This set can be logically divided into 3 subsets. Some of the features in the first two sets are also seen in (Murray & Carenini 2011).

1. *Baseline conversational features*

   These include the following:

   a. The position of the spurt within a turn.

   b. The position in the meeting of the turn the spurt belongs to.

   c. The length of the spurt (number of words) normalized by the max spurt length in the current turn.

   d. Whether the current spurt is the first one in an interrupting turn

   e. Whether the turn the current spurt belongs to is the first turn of a speaker in the meeting

   For example for the fourth spurt of one of the meetings "This looks good" that also constitutes a turn, we have the following features (the turn is not an interrupting one and it is not the first one in the meeting for the current speaker):

   ```
   PositionOfDA@0.0 PositionOfTurn@0.00980392156862745 DALengthNormMaxDAinTurn@1.0
   ```

2. *Individual conversational features*

   These features attempt to express the fact that each speaker has a personal style and way of speaking. As a result, the way the current spurt diverges from that style is reflected on them.

   a. The number of spurts in the current turn divided by the maximum number of spurts in a turn for the current speaker in the meeting.

   b. The length of the current spurt (number of words) divided by the maximum length spurt for the current speaker in the meeting

c. The time distance of the current turn from the previous turn divided by the maximum time distance among turns for the current speaker in the meeting

d. The average word length in the current spurt divided by the maximum average word length in all spurts for the current speaker in the meeting.

e. The word rate in the current turn divided by the maximum turn word rate for the current speaker in the meeting.

f. If the current turn is an interrupting turn, we include the percentage of the turns of the current speaker that interrupted the speaker before.

So for the same example spurt “This looks good”, the set of features will be the following:

```
PerSpeakerTurnLength@0.25 PerSpeakerDALength@0.03260869565217391
PerSpeakerTimeFromePrevTurn@0.06560150375939862
PerSpeakerMeanWordLengthInDA@0.6666666666666666
PerSpeakerWordRateInTurn@0.4000000000005457
```

3. *Non-individual conversational features*

These features are not specific to a speaker and are in correspondence to the conversational features mentioned above:

a. The number of spurts in the current turn.

b. The number of words in the current spurt.

c. The time distance from the previous turn.

d. The average word length (number of characters) in the current spurt.

e. The word rate (number of words in the turn divided by the duration of the turn) in the current turn.

f. Whether the current turn interrupted the previous turn (this information was extracted from the annotations found in the agreement/disagreement corpus based on the presence or absence of the tag <Interrupt> in the first spurt of the turn).

For the same example, the features extracted are:

```
TurnLength@1 DALength@3 TimeFromePrevTurn@0.6980000000000004
MeanWordLengthInDA@4.333333333333333 WordRateInTurn@4.0
```

### 3.2.4. Features based on existing Dialog Act annotations

As was discussed in the literature review section, the corpus that we used for our study, apart from having been annotated for agreement/disagreement, had also been split into annotated Dialog Acts (Shriberg, Dhillon, Bhagat, Ang & Carvey 2004). Due to the fact that dialog acts constitute an alternative representation of utterances and since they are speech segments with specific functionality, we decided to leverage these existing annotations and incorporate them in our feature set in order to study their effect in the agreement/disagreement detection task. Since agreement/disagreement annotations had been applied to a spurt and a spurt was the target of our classifier, we had to map the Dialog Act annotations to each spurt. Each spurt could either be the same as a Dialog Act or a subset of it or it could span over a number of dialog acts. For that reason, we detected all the dialog act tags that spanned over one spurt and we generated unigrams and bigrams out of them keeping only the general part of each tag. These unigrams and bigrams of dialog act tags were added to the features used for one spurt.

So for example the spurt “so uh so the uh the new procedural change that just got suggested which i think is a good idea is that um we do the digit recordings at the end” will have the following features:

```
daUnigrams.fg daUnigrams.s daBigrams.fg.s
```

### 3.2.5. Coordination features

The linguistic coordination features were inspired by the study in (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012) where the relationship between the status of a person in a group and the degree of linguistic style coordination is examined. Quoting from the above paper (end of page 1, beginning of page 2): “Language coordination is a phenomenon in which people tend to unconsciously mimic the choices of function-word classes made by the people they are communicating with; roughly speaking, if you are communicating with someone who uses a lot of articles — or prepositions, or personal pronouns — then you will tend to increase your usage of these types of words as well, even if you don’t consciously realize it.”. In the same paper, it is stated that “the need to convince someone who disagrees with you creates a form of dependence” and it is shown that these dependents tend to coordinate more towards the people they depend on. This conclusion led us to hypothesize that the degree of coordination between two turns could assist the agreement/disagreement detection task, and more specifically that the strong presence of linguistic style coordination could be linked to disagreeing utterances. To leverage the fact that coordination could assist our task, we define a score to quantify the linguistic style coordination between one turn and the one before. For each one of the following classes of words (articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers), we calculate this score as following: if words of a specific category exist in the current turn and if the previous turn does not contain any word of the same category, then the score is zero. On the other hand, if words of the same category are present in the preceding turn, then the score will be one (1) minus the proportion of the current speaker’s total number of turns during which he/she uses words from the specific category (regardless of the preceding turns). The above is summarized in the following equation:

$$S_C = \begin{cases} 0 & , if\ current\ turn\ does\ not\ contain\ words\ of\ category\ C \\ & or \\ & if\ it\ contains\ words\ of\ category\ C \\ & but\ the\ preceding\ turn\ doesn\ ot\ contain\ any \\ 1 - \dfrac{\substack{number\ of\ turns\ where\ current\ speakers \\ uses\ words\ of\ category\ C}}{\substack{total\ number\ of\ turns \\ of\ the\ current\ speaker}}, & Otherwise \end{cases}$$

*Equation 1: Coordination score for class C (where C is one of the classes mentioned above e.g. articles, auxiliary verbs, conjunctions etc).*

Note that the above categories of words that are used to express the linguistic style of a speaker are the same as the ones used in (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012). The above paper, however, tried to express the coordination among two speakers and then generalized by defining the coordination among groups of people with different status. What we are trying to do in this thesis is to come up with a measure of the degree of coordination between two specific turns. We are not interested in drawing conclusions regarding how a speaker generally responds to another speaker. So, if words of a certain class are shared between two turns we just subtract the score we described above, and we use this quantity as a measure that shows to what extent there is a coordination relationship among the two turns.

In order to determine which words belong to a specific category, we used the LIWC dictionary (Pennebaker, Booth & Francis 2007). The final coordination feature set that we extract for a specific spurt is one that describes the turn this spurt belongs to and includes (for each class in the set above) the linguistic coordination score.

For example for any spurt within the turn “that is weird it's like when it's been sitting for a long time or something” (it contains a single spurt in this case), the previous turn is “this looks good” and the set of coordination features will be:

```
ARTICLES@0.0          AUX_VERBS@0.0          CONJUNCTIONS@0.0          FREQ_ADVERBS@0.0
IMP_PRONOUNS@0.33333333333333337 PERS_PRONOUNS@0.0 PREPOSITIONS@0.0 QUANTIFIERS@0.0
```

## 3.3. Baseline generation

To generate a baseline to which we compared our results, we implemented a vocabulary-based approach similar to (Hu & Liu 2004). The method was used for polarity detection in that paper and we employed it for our agreement/disagreement study. We used a subjectivity dictionary and for each spurt we counted the number of instances of positive and negative words. If the positive words were more than the negative ones, then the instance was labeled as expressing agreement and vice versa. We used the MPQA subjectivity dictionary (Wilson, Wiebe &Hoffmann 2005) for that purpose.

## 4. Algorithms, implementation, etc.

As mentioned earlier, the problem of annotating the spurts of a turn for agreement/disagreement was modeled using linear chain CRF. For training and testing of the model we used the CRF implementation of Mallet software (McCallum 2002). Specifically, this software provides a class called SimpleTagger, which we used after we modified it for continuous feature support. (Source: SImpleTaggerContinuous Features.java)

The training and test data provided to the above part of the implementation was created by a java component that we wrote and which split the data in order to perform n-fold cross validation. (Source: Runner.java)

For the lexicosyntactic and dialog act-based features, we wrote a java component that performed feature selection using chi-square metric for ranking. (Source: SelectionHelper.java)

In order to extract the lexicosyntactic features from a spurt we used some of the Stanford NLP group tools (The Stanford Natural Language Group 2010) and particularly their tokenizer as well as their POS tagger.

For extracting the coordination-based features, we used the LIWC dictionaries (Pennebaker, Booth & Francis 2007) which provided us with a set of words belonging to a function-word class we were interested in (articles, auxiliary verbs, conjunctions, high-frequency adverbs, impersonal pronouns, personal pronouns, prepositions, and quantifiers). (Source: CoordinationHelper.java)

In order to extract the dialog act based features (Shriberg, Dhillon, Bhagat, Ang & Carvey 2004), we built a system that searched for the dialog act tags within a spurt. Then it generated unigrams and bigrams of dialog act tags as mentioned in the “Methodology” section. This task was particularly challenging because even though the sections we used of the two corpora were supposed to have the same transcripts, certain

words or filler sounds had been transcribed differently. Since the differences were not known in advance and did not follow any pattern, we had to discover and then hard code these difference for each meeting. (Source: DATagMatcher.java, InputVectorGenerator.java)

As far as the conversational features are concerned, we extracted them from the meeting transcript files using java classes we wrote. The biggest challenge here was to determine the beginning and ending time of each turn. The agreement/disagreement corpus did not include this information so we had to go back again to the dialog act corpus, map the turns there and extract the timing information. The differences between the two transcripts made this task challenging. (Source: DialogAct.java, Turn.java, Speaker.java, MeetingParser.java, GlobalStatistics.java, InputTextConverter.java)

To make our system easily extensible to other corpora, we converted the meeting transcript information to xml format so that the next stages of our system's pipeline did not need to know about the initial file format. (Source: InputTextConverter.java)

In order to generate our baseline, we used the dictionary based approach described in the methodology section. The dictionary we used was from (Wilson, Wiebe &Hoffmann 2005) and the classification algorithm was similar to the one used in (Hu & Liu 2004) for polarity classification. For that part of the code we used the stemmer from the Stanford NLP Group tools mentioned earlier (Source: BaselineRunner.java, LexiconApproachHelper.java)

# 5. Experiments

## 5.1. Creating our baseline

We created our baseline as described in the "Methodology" section. This method labels a sentence as "Pos," "Neg," or "Neutral" based on how many positive and/or negative words it contains as defined in the MPQA subjectivity lexicon. The presence of more positive than negative words results in the sentence being labeled as "Positive", while in the opposite case it is marked as "Negative". When the number of positive words is equal to the number of negative words, it is labeled as "Neutral". We generated the baseline using two different methods: 1. Using only the strongly subjective words from the dictionary and 2. Using all the subjective words.

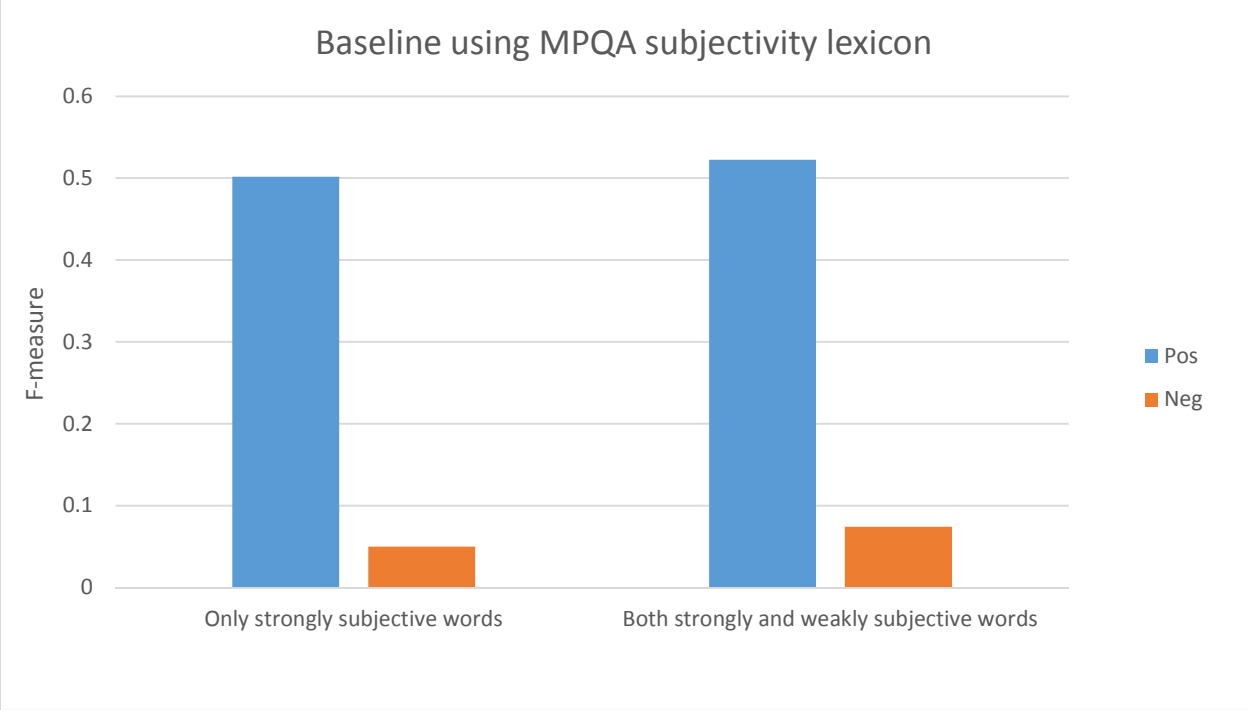



*Chart 1: Baseline measurements using the MPQA subjectivity lexicon*

Looking at the chart above, we can see that the two experiments give similar results to the one using both the strongly and weakly subjective words giving a slightly better F-measure. Moving forward this is

going to be our baseline. It is worth mentioning that the subjectivity lexicon approach results in very poor performance for disagreement detection compared to agreement detection.

## 5.2. Experiments with a shared feature set

We started by performing a set of experiments that all use the same baseline feature set: unigrams, bigrams and baseline conversational features (see "Methodology" section above for analysis of each type of feature set). These runs were:

a. CRF Order=0: We used all of the instances in the annotated agreement/disagreement corpus. For each turn we performed training/testing using CRF with order equal to 0 (i.e. for each spurt we did not take into account the previous spurt in the turn)
b. CRF Order=1: This type of run is similar to the above except for the order of the CRF which is 1,i.e. for each spurt we assume a relationship with the previous one in the turn.
c. Downsampling: The annotated corpus does not contain spurts uniformly distributed over the 4 classes ( Pos (Positive), Neg (Negative), Other, BackChan (backchannel) ):

   Its composition is as follows:

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 1114 |
| BackChan | 405 |

*Table 4: Initial composition of the agreement/disagreement corpus*

Due to the above observations, we downsampled "Other" and "BackChan" classes (by randomly deciding to include or exclude certain instances) and the new distribution was the following.

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 135 |
| BackChan | 116 |

*Table 5: Composition of the agreement/disagreement corpus after downsampling "Other" and "BackChan" classes*

Since after the downsampling certain spurts have been removed and the remaining ones are not necessarily consecutive, we used CRF with order equal to 0.

The results for the above types of runs are shown in the following chart:

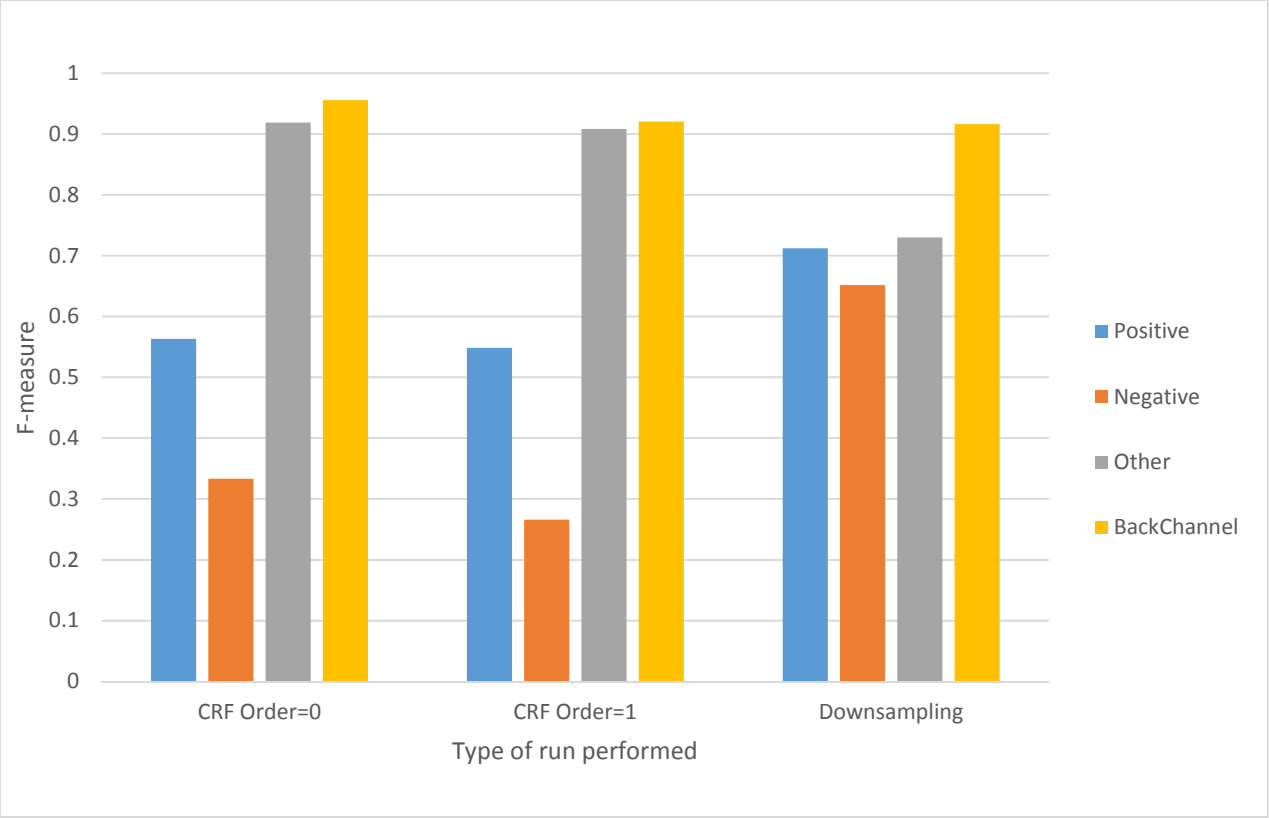



*Chart 2: Behavior of the F-measure for all classes and for different types of runs. The features used for each run are: unigrams, bigrams and baseline conversational (as defined in the "Methodology" section).*

Based on the above results, we can see that CRF with order equal to 0 performs better than CRF with order 1 for all classes. When we downsample our data so that all classes are almost equally represented, the results for "Pos" and "Neg" classes improve significantly compared to the other runs while there is some insignificant deterioration for "BackChan" class. For the remaining experiments in this thesis we followed the downsampling approach (that uses CRF with order equal to 0).

### *5.3. Comparing various feature sets*

Using the downsampling approach mentioned above, we carried out a series of experiments in order to compare the performance of various feature sets. The results are shown in the following chart.

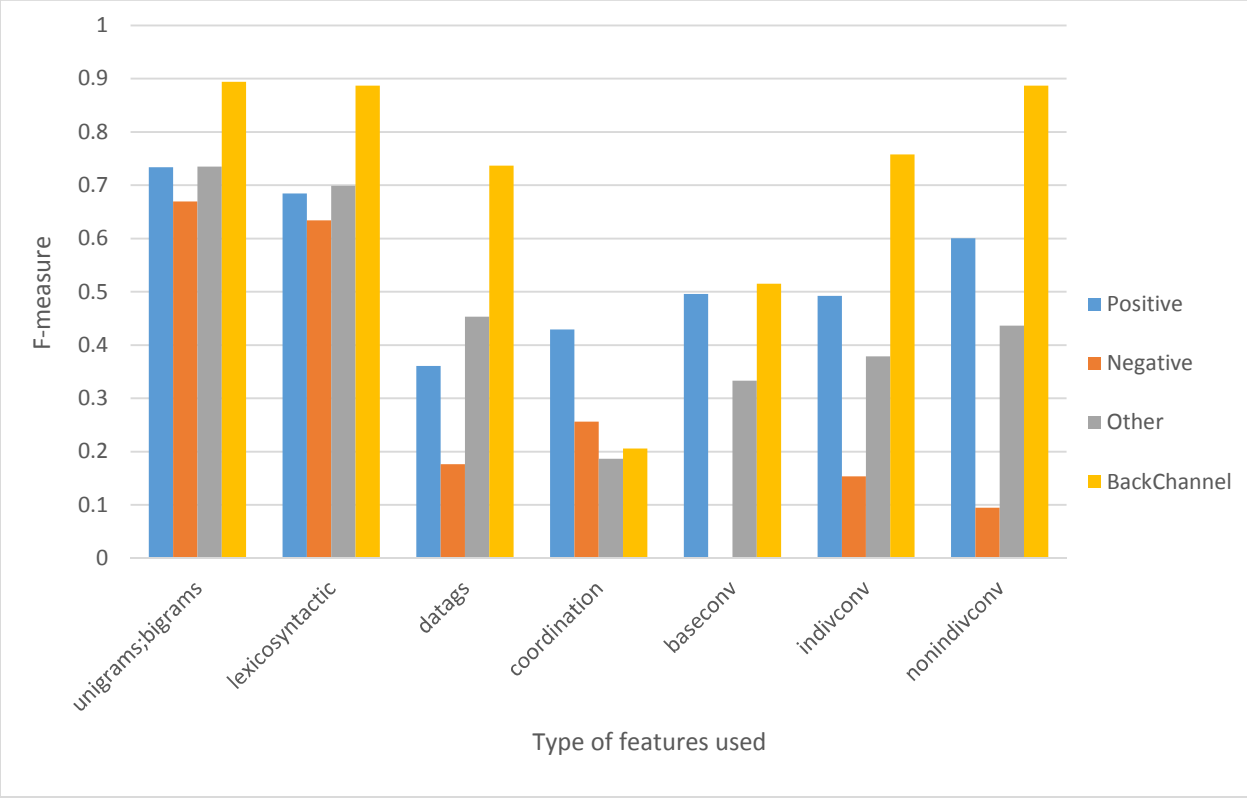



*Chart 3: F-measure for each class when using different feature sets. The features shown on the horizontal axis from left to right correspond to the following feature sets that have been discussed in the "Methodology" section: 1. Unigrams, bigrams; 2. Lexicosyntactic features; 3. Features based on dialog act annotations; 4. Coordination features; 5. Baseline conversational features; 6. Individual conversational feature; 7. Non-individual conversational features.*

We can see that the use of unigrams/bigrams and lexicosyntactic features gives the best results for all classes of interest ("Pos", "Neg", "BackChan"). Additionally, these two feature sets have similar performance. Therefore the use of unigrams/bigrams gives very good results and at the same time leads to a much smaller feature set compared to all lexicosyntactic features. Comparing the experiments that use dialog act-based and coordination-based features, the use of the latter improves the performance for "Pos" and "Neg" classes while it gives significantly worse results for "BackChan." Additionally, both of these feature sets appear to have poorer performance for "Pos" than each one of the conversational feature sets while they perform better for "Neg." Looking at the results of conversational features, there are no trends that are shared among the various classes so we will talk about each one of them individually. Non-individual conversational features perform better than baseline conversational and individual conversational features which have about the same performance when compared to each other. On the other hand, individual conversational features perform better than non-individual ones for "Neg" class, while the second ones perform better than baseline conversational. The performance for "BackChan" is the worst when baseline conversational features are used, but it improves significantly with the individual conversational features and shows further increase in the experiments with non-individual conversational features.

Next we combined unigrams/bigrams with other features to see the effect that the additional features had:

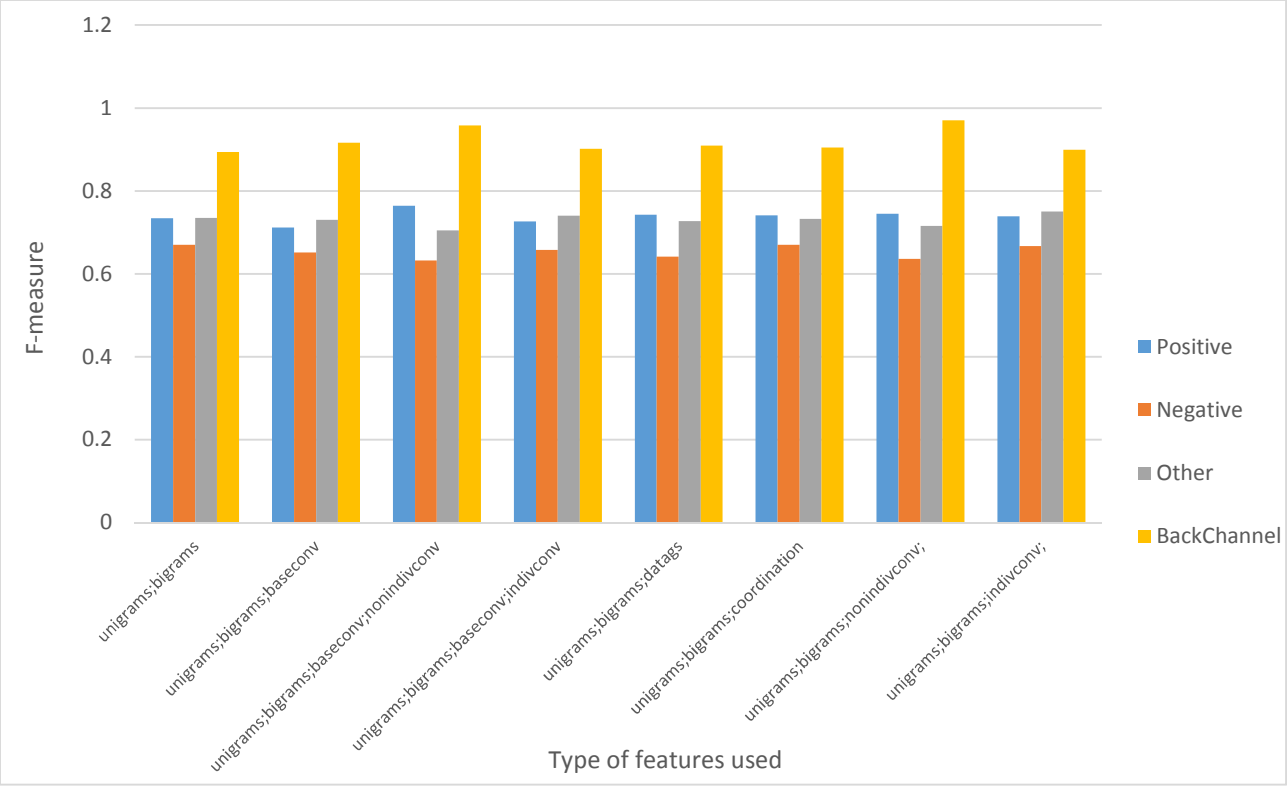



*Chart 4: F-measure for each class when the combination of unigrams/bigrams and other features is used.*

Based on the above chart, we see that the addition of baseline conversational features causes a slight deterioration in the performance for “Pos” and “Neg” while it helps the “BackChan” classes. When adding both non-individual and baseline conversational features to our baseline, we have an increase in F-measure for “Pos” and “BackChan” while the F-measure for “Neg” decreases. On the other hand, the addition of individual and baseline conversational features slightly deteriorates the performance for “Pos”, “Neg” and causes a small improvement for “BackChan” class. Adding Dialog Act-based features gives better performance for “Pos” and “BackChan” and worse for “Neg”, while coordination features improve the results for “Pos” and “BackChan” and do not affect “Neg” class. Studying the individual and non-individual conversational features independently from the base conversational features, we see that the non-individual ones improve the performance of unigrams/bigrams for “Pos” and “BackChan” while for “Neg” class the performance is worse. Individual conversational features leave the performance of unigrams/bigrams almost unchanged. In reality it is insignificantly worse.

The results are summarized in the following table:

| Class | Baseline F-measure | Best F-measure | Feature set whose addition caused the best F-measure | Worst F-measure | Feature set whose addition caused the worst F-measure |
|---|---|---|---|---|---|
| Pos | 0.733945 | 0.764526 | Baseline and non-individual conversational features | 0.712074 | Baseline conversational features |
| Neg | 0.669811 | 0.669811 | Coordination-based features (but even without the additional feature set the baseline has the same performance) | 0.632558 | Baseline and non-individual conversational features |
| BackChan | 0.894118 | 0.970464 | Non-individual conversational features | - | The addition of any of the feature sets improved the performance for BackChan class |

*Table 6: Summary of the effect that different feature sets have on F-measure when added to the "unigrams/bigrams" feature set*

Next we repeated the same measurements as in Chart 3 and Chart 4, but by using only the "Pos" and "Neg" instances during training and testing. The results were the following:

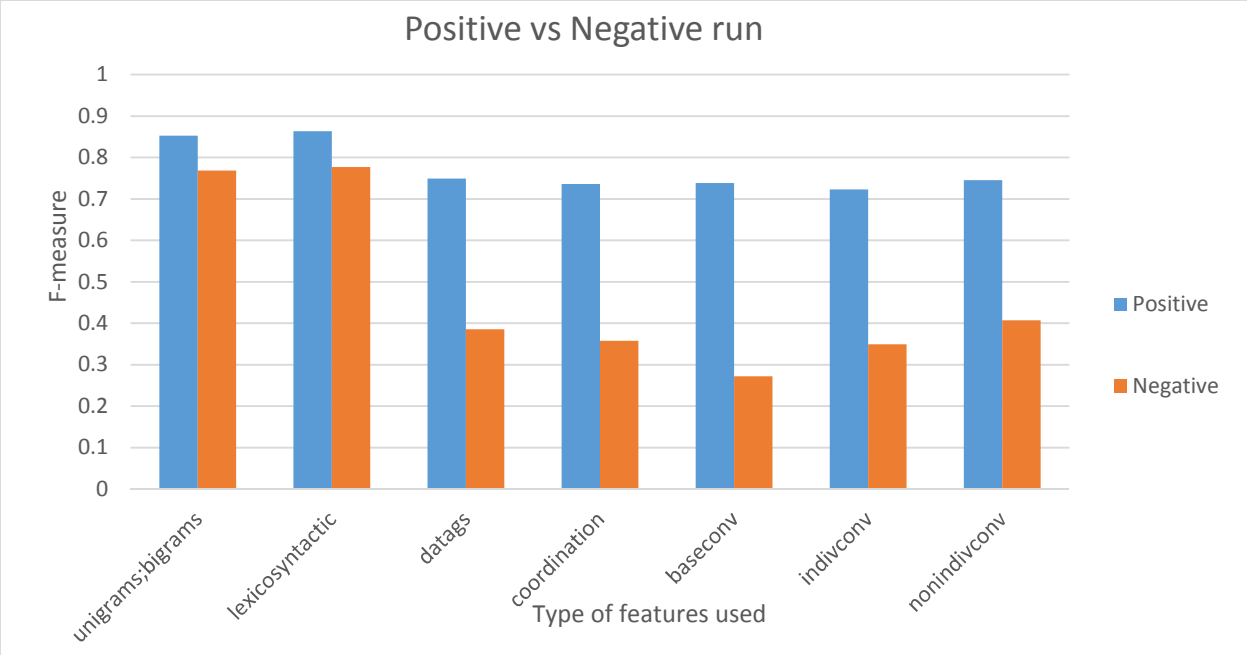



*Chart 5: F-measure for each class when using different feature sets. For these runs only Pos and Neg instances were used. The features shown on the horizontal axis from left to right correspond to the following feature sets that have been discussed in the "Methodology" section: 1. Unigrams/bigrams, 2. Lexicosyntactic features, 3. Features based on dialog act existing annotations, 4. Coordination features, 5. Baseline conversational features, 6. Individual conversational features, 7. Non-individual conversational features.*

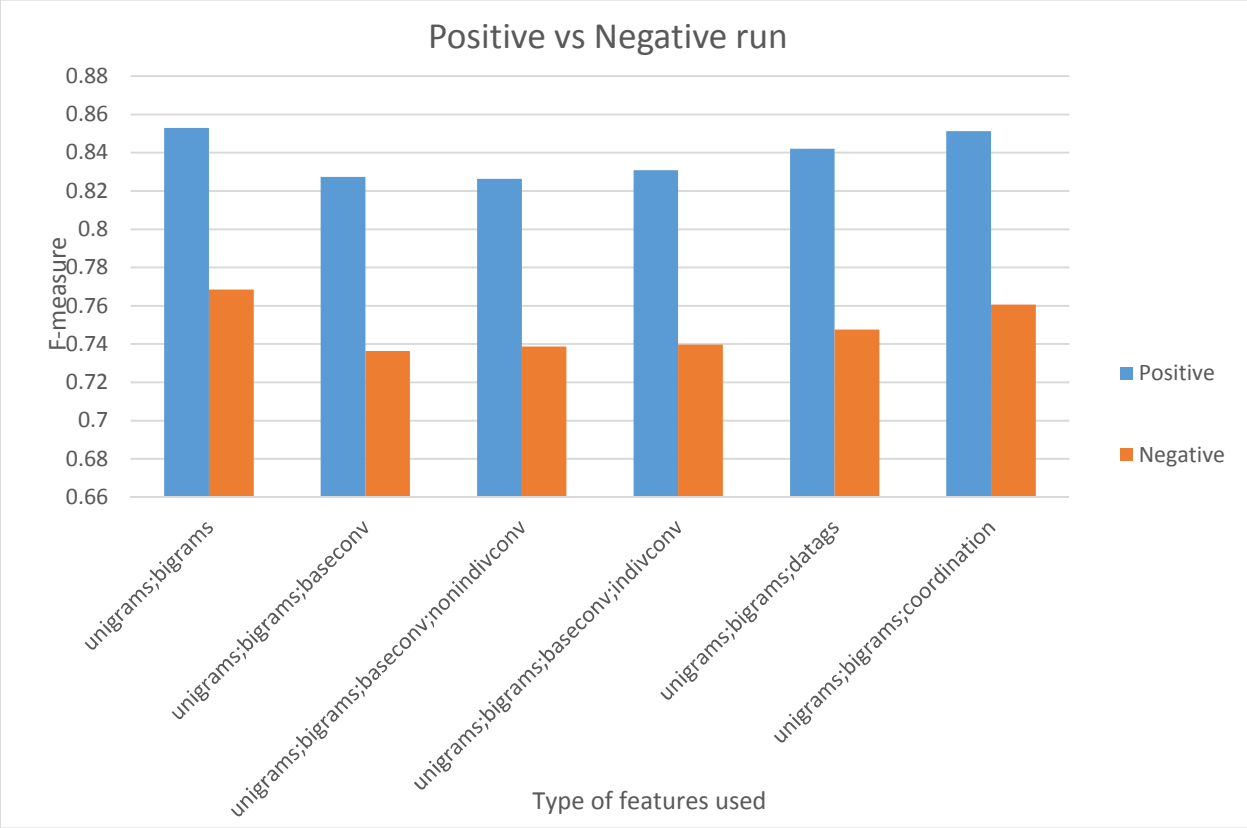


*Chart 6: F-measure for each class when the combination of unigrams/bigrams and other features is used. For these runs only Pos and Neg instances were used.*

Based on the previous two charts we can see that performance has improved compared to the 4-class task for all of the feature sets. This result did not surprise us since we had fewer instances belonging to only two categories and the noise from the "BackChan" and "Other" instances was absent. Another thing to notice is that the addition of other feature sets to unigrams/bigrams did not lead to better performance.

### 5.3.1. Experiments when performing feature selection for the "unigrams/bigrams" feature set

Unigrams/bigrams constitute a big feature set and the features it includes are numerically dominant compared to the other feature sets we are studying. Since we wanted to more closely investigate the effect of the various feature sets, we decided to use only a subset of the unigrams/bigrams in our training data. Thus, the effect of the features added on top of them would be more clearly seen. This change was implemented by keeping the most frequent unigrams/bigrams seen in our training data. This was done for each fold and the same number of features was kept for all. We made the decision to keep the 15 most frequent unigrams/bigrams from each fold. The features we kept had to be those that would lead to decent performance: not too low since this way any addition could blindly lead to improvement and not too high either so that the additional features can have some impact. The following curve shows the performance for different numbers of most frequent unigrams/bigrams kept:

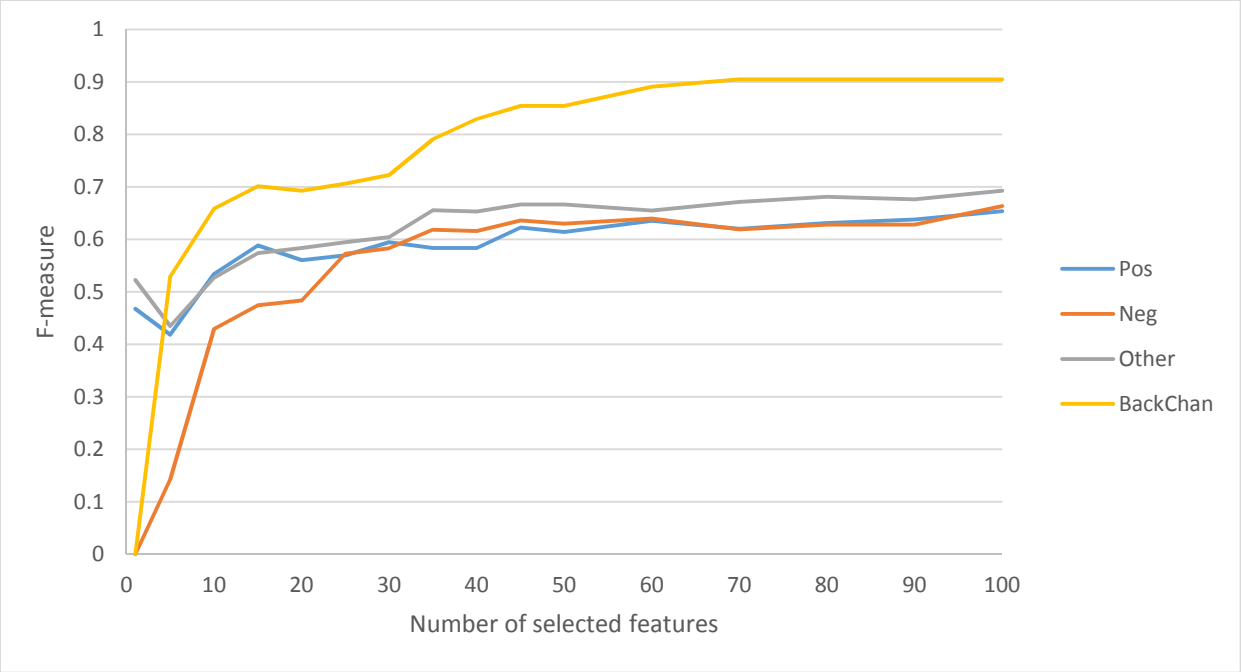



*Chart 7: Changes in F-measure for each class as the number of unigrams/bigrams selected gets different values*

Regarding the 15 unigrams/bigrams that were selected, we merged them for all folds (this is why there are more than 15) and we present them in the following table:

| |
|---|
| WordUnigrams.yeah |
| WordUnigrams.you |
| WordUnigrams.the |
| WordUnigrams.to |
| WordUnigrams.that |
| WordUnigrams.'s |
| WordUnigrams.of |
| WordUnigrams.a |
| WordUnigrams.and |
| WordUnigrams.but |
| WordUnigrams.i |
| WordUnigrams.it |
| WordUnigrams.so |
| WordUnigrams.right |
| WordUnigrams.is |
| WordUnigrams.uh |
| WordUnigrams.n't |

*Table 7: Merged set of unigrams/bigrams features that were used by the various folds*

The results of the runs using various feature set on top of a subset of the unigrams/bigrams are shown next:

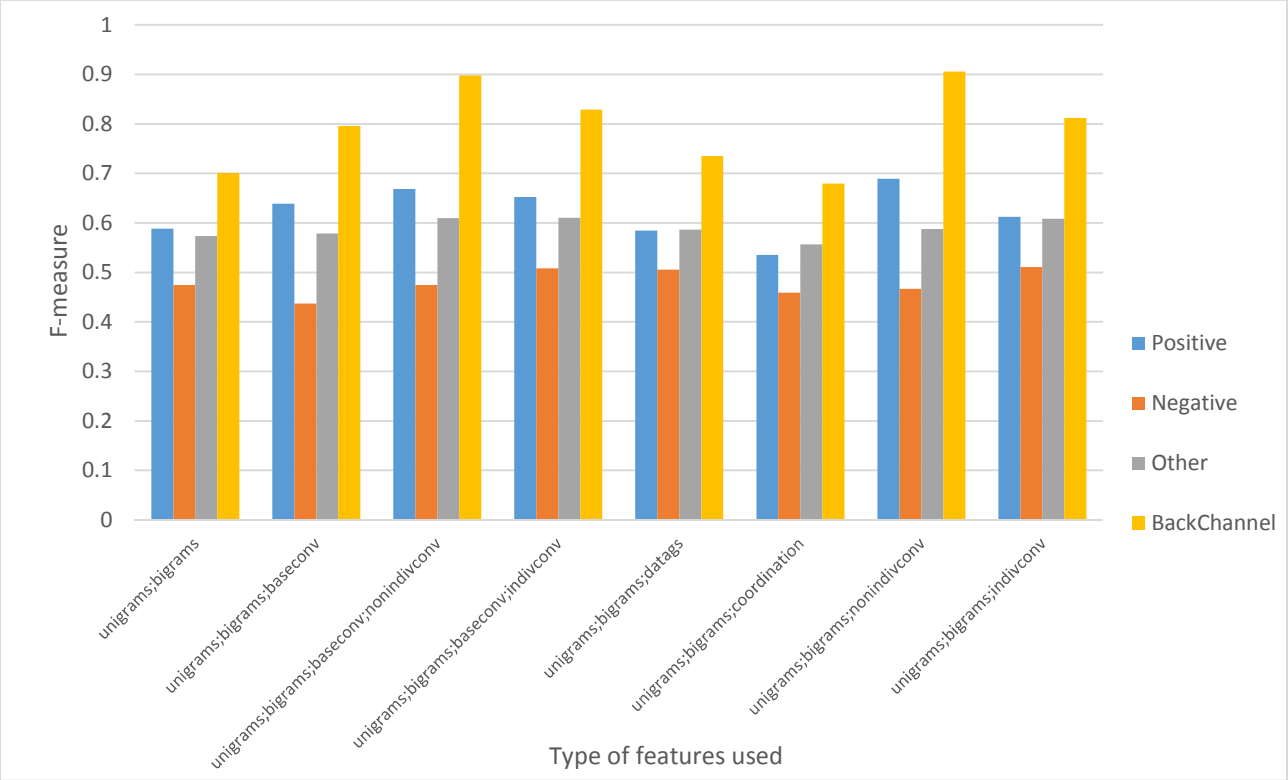



*Chart 8: F-measure for each class when the combination of unigrams/bigrams and other features is used and in the case where only the 15 most frequent unigrams/bigrams are kept in each fold.*

Based on the above chart, we can sort (from best to worst) the various feature sets based on the effect they have on F-measure when added to unigrams/bigrams.

| Pos | Neg | BackChan |
|---|---|---|
| 1. Non-individual conversational features | 1. Individual conversational features | 1. Non-individual conversational features |

| | | |
|---|---|---|
| 2. Baseline and non-individual conversational features<br>3. Baseline and individual conversational features<br>4. Baseline conversational<br>5. Individual conversational features<br>6. DA tag-based features<br>7. Coordination-based features | 2. Baseline and individual conversational features<br>3. DA tag-based features<br>4. Baseline and non-individual conversational features<br>5. Non-individual conversational features<br>6. Coordination-based features<br>7. Baseline conversational | 2. Baseline and non-individual conversational features<br>3. Baseline and individual conversational features<br>4. Individual conversational features<br>5. Baseline conversational<br>6. DA tag-based features<br>7. Coordination-based features |

*Table 8: All features sets are sorted from best to worst based on the effect they have on the F-measure of each class*

Looking at the information in the above chart, we can see that the non-individual conversational features are quite relevant when it comes to recognizing "Pos" and "BackChan" in the corpus. On the other hand, individual conversational features appear to improve the detection of "Neg" instances.

## 5.4. Experiments with lexicosyntactic features

Next we studied the lexicosyntactic features which are a superset of unigrams/bigrams as mentioned in the "Methodology" section. We decided to focus on one of the folds of our corpus and examine the effect of the number of features selected on F-measure. For this purpose we ranked the features based on their chi Square value and for each run we kept a certain number of the features at the beginning of the list (ordered by decreasing chi Square value). Please note that this experiment is different from the one we

performed in section 5.3.1. In that section, we selected a subset of the unigrams/bigrams (vs. all lexicosyntactic features) keeping the most frequent features among them (vs. the ones with the best Chi Square value) and also the focus there was to study the effect of various other feature sets added on top of them. In this experiment, we want to see the effect of building the set of active features by gradually adding features ranked by their Chi Square value. Additionally, the features with the highest Chi Square values are presented. The results are shown in the following chart:

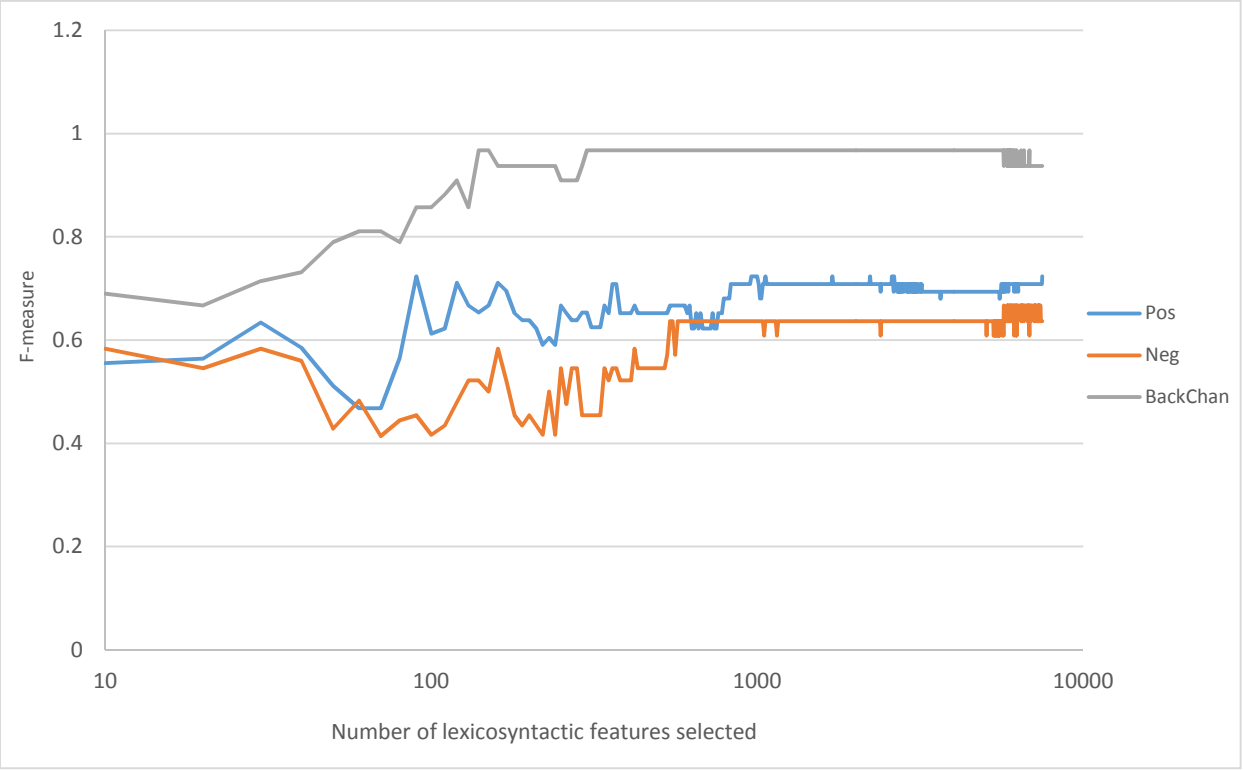



*Chart 9: F-measure per class when only lexicosyntactic features are used*

At the beginning, and as we add features to our active feature set, performance seems to be improving. Later as we keep adding features, F-measure reaches an upper bound and it remains within a very small range of values. The above behavior was expected: the most important features improve performance and the less important features have a less significant impact on performance.

Taking a look at the ranked list of lexicosyntactic features we can determine which features seem to be more useful for our classification task. Here is the list of the top 30 features:

| | |
|---|---|
| 1. WordUnigrams.but | 16. POSSkipBigrams.1.IN.NN |
| 2. WordUnigrams.yeah | 17. POSBigrams.IN.PRP |
| 3. WordUnigrams.well | 18. WordUnigrams.in |
| 4. WordUnigrams.the | 19. POSBigrams.DT.NNS |
| 5. POSBigrams.DT.NN | 20. WordUnigrams.'s |
| 6. WordUnigrams.uhhuh | 21. WordUnigrams.it |
| 7. WordUnigrams.i | 22. WordUnigrams.no |
| 8. WordUnigrams.and | 23. POSBigrams.DT.JJ |
| 9. WordUnigrams.you | 24. POSSkipBigrams.1.DT.NN |
| 10. POSBigrams.PRP.VBP | 25. WordBigrams.but.i |
| 11. POSBigrams.IN.DT | 26. POSBigrams.NN.CC |
| 12. WordUnigrams.that | 27. POSSkipBigrams.1.NN.IN |
| 13. POSBigrams.NN.IN | 28. POSSkipBigrams.1.IN.VBP |
| 14. WordUnigrams.of | 29. WordUnigrams.uh |
| 15. POSSkipBigrams.1.DT.IN | 30. POSBigrams.JJ.NN |

*Table 9: Top 30 features in the list of lexicosyntactic features ordered by decreasing Chi Square value*

## 5.5. Experiments with features based on dialog act tags

In earlier charts, we showed the results of the runs where dialog act based features were used. In order to better understand the available data, we studied the distribution of these features.

The distribution is shown in the following charts for each class:

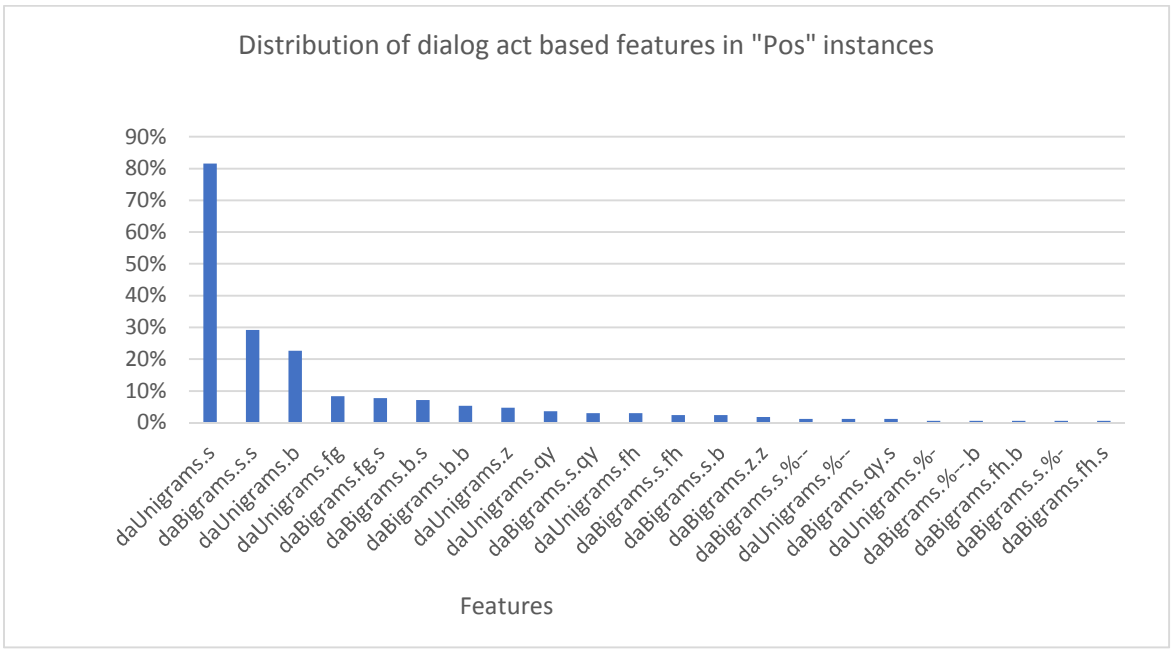



*Chart 10 (a)*

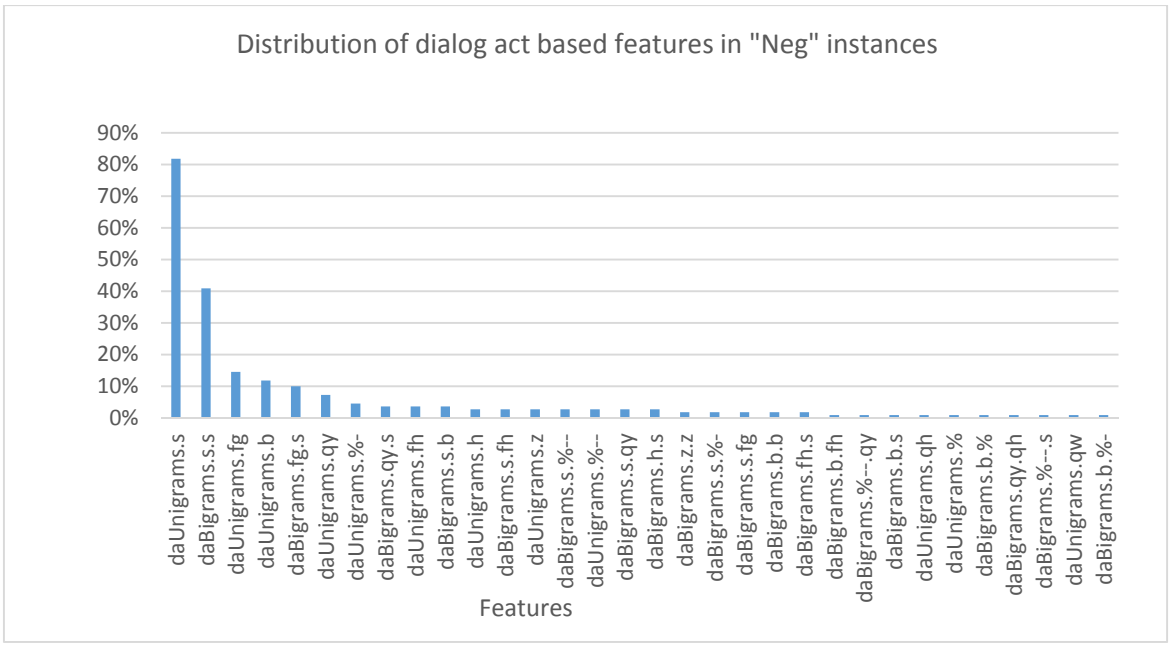



*Chart 10 (b)*

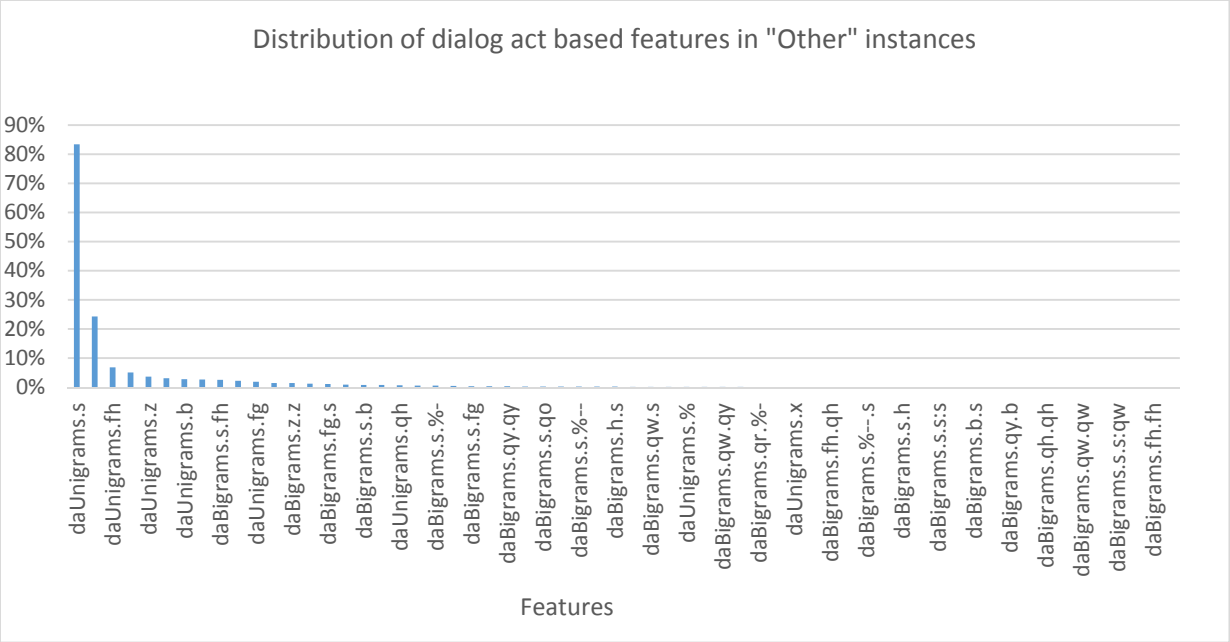



Chart 10 (c)

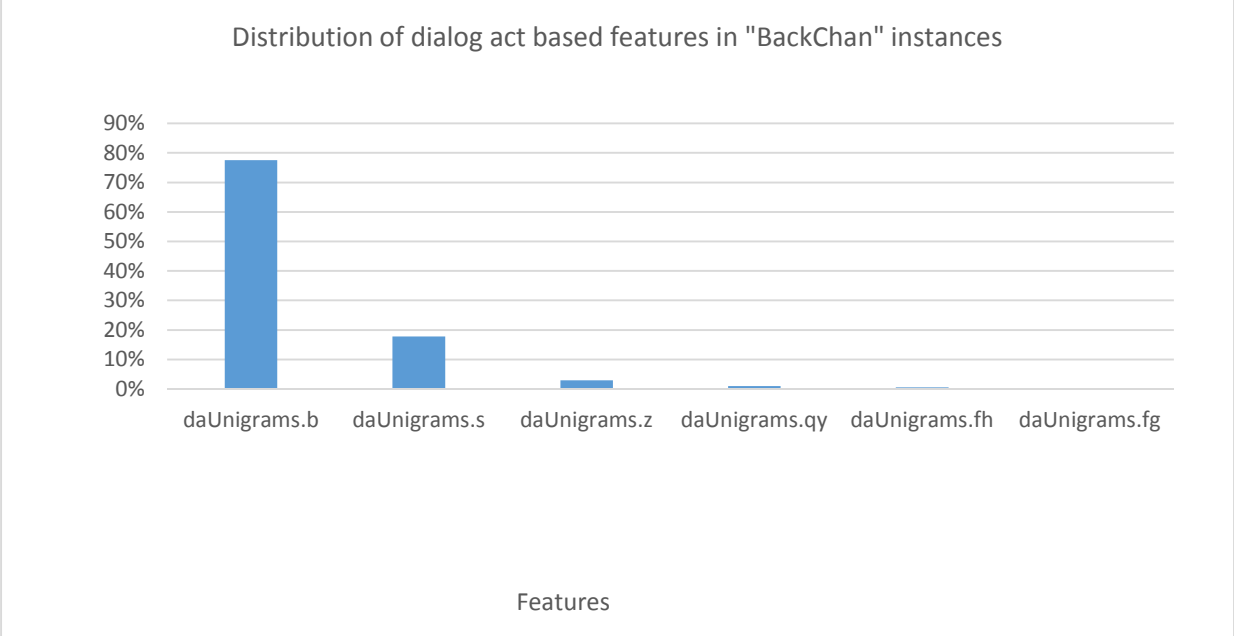



Chart 10 (d)

*Chart 10: Percentage of spurts per class that contain each one of the dialog act based features. The features have been sorted based on the frequency. The horizontal axis shows the dialog act based features which are unigrams and bigrams of the general part of the dialog act parts found in the MRDA corpus. The feature daUnigrams.s refers to the presence of the general part "s" while the feature daBigrams.s.fh refers to two consecutive general parts (the first one is "s" and the second one is "fh")* (*Appendix contains an explanation of what the feature symbols mean)*

For back-channels, we notice the back-channel DA tag is present for most of the instances. Also, back-channel DA tags are more frequent in “Pos” instances compared to “Neg” ones while the opposite happens with floor grabbers. Yes/No question DA tags are found more frequently in negative than positive statements while the same happens with consecutive statements (daBigrams.s.s).

As we did with the lexicosyntactic feature set, we focused again on one fold, ranked the features in the training data using the Chi Square values and performed feature selection based on them. The results of the feature selection experiments are shown in the following chart. The vertical axis shows what feature gets added in the set of selected features for the current run. The horizontal axis demonstrates the F-measure value for each class.

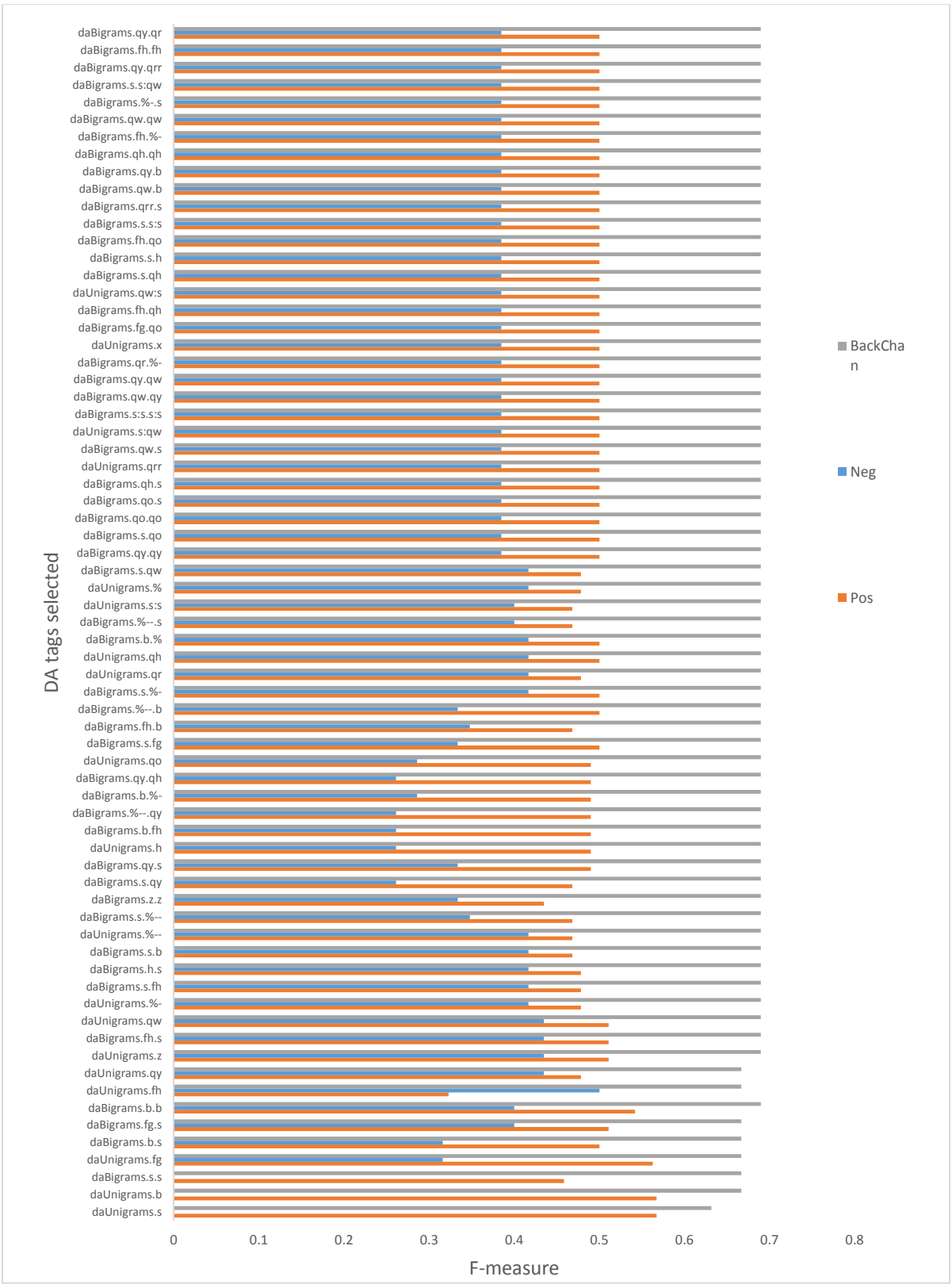



*Chart 11: F-measure per class when only dialog act-based features are used. Each position on the vertical axis implies that for this specific run, the selected features are the ones preceding them on the axis and additionally the label found in this position*

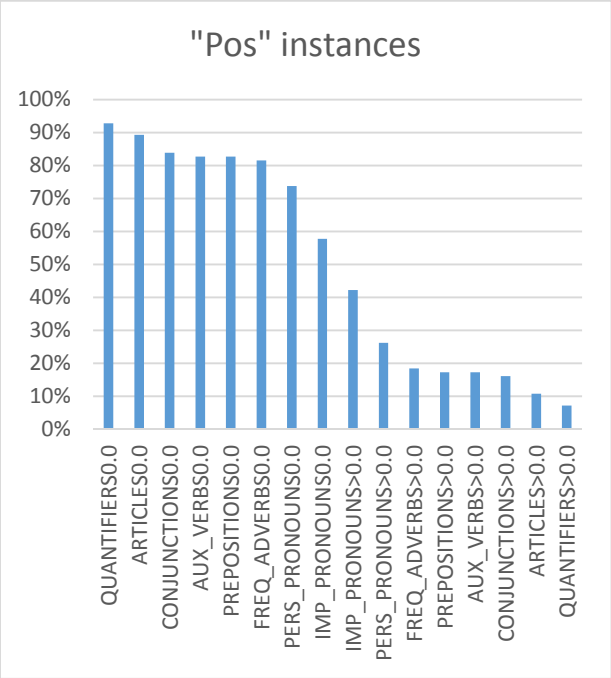
The top 30 features in the ranked list are shown in the following table:

| | |
|---|---|
| 1. daUnigrams.s | 16. daBigrams.s.b |
| 2. daUnigrams.b | 17. daUnigrams.%-- |
| 3. daBigrams.s.s | 18. daBigrams.s.%-- |
| 4. daUnigrams.fg | 19. daBigrams.z.z |
| 5. daBigrams.b.s | 20. daBigrams.s.qy |
| 6. daBigrams.fg.s | 21. daBigrams.qy.s |
| 7. daBigrams.b.b | 22. daUnigrams.h |
| 8. daUnigrams.fh | 23. daBigrams.b.fh |
| 9. daUnigrams.qy | 24. daBigrams.%--.qy |
| 10. daUnigrams.z | 25. daBigrams.b.%- |
| 11. daBigrams.fh.s | 26. daBigrams.qy.qh |
| 12. daUnigrams.qw | 27. daUnigrams.qo |
| 13. daUnigrams.%- | 28. daBigrams.s.fg |
| 14. daBigrams.s.fh | 29. daBigrams.fh.b |
| 15. daBigrams.h.s | 30. daBigrams.%--.b |

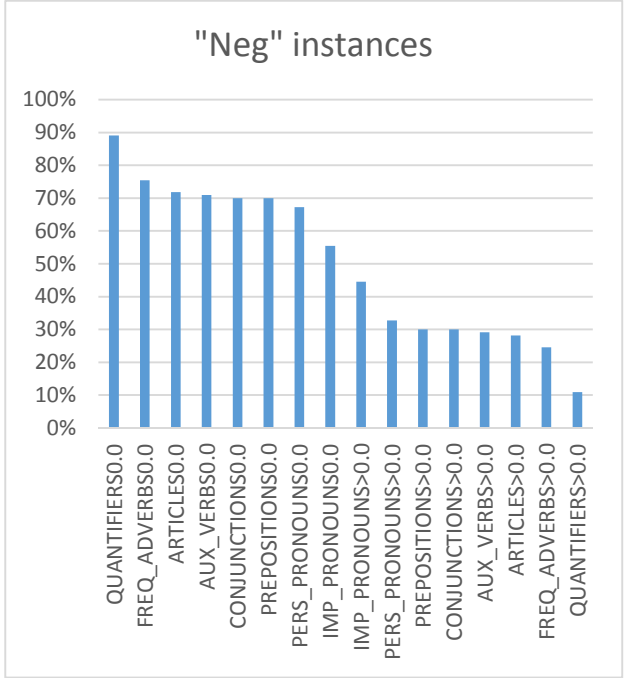
*Table 10: Top 30 features in the list of dialog act tag-based features ordered by decreasing Chi Square value*

## 5.6. Experiments with features based on linguistic style coordination

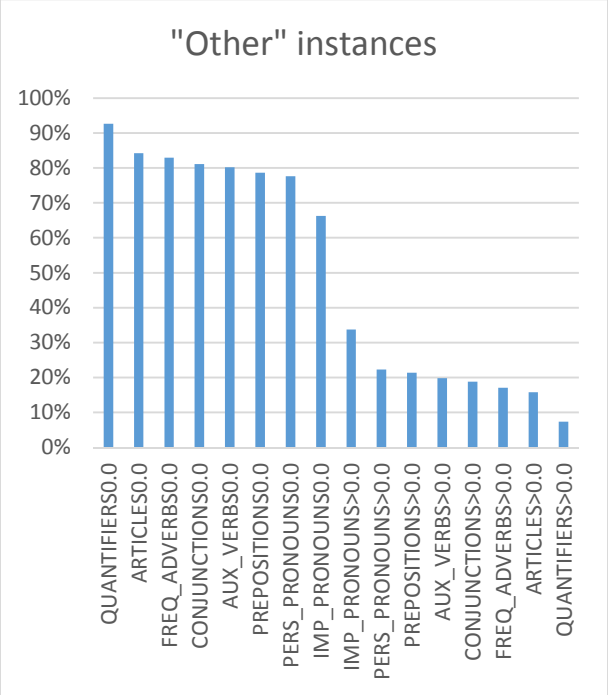
For this type of features, we analyzed the distribution of various coordination properties among instances of different classes. The results are the following:



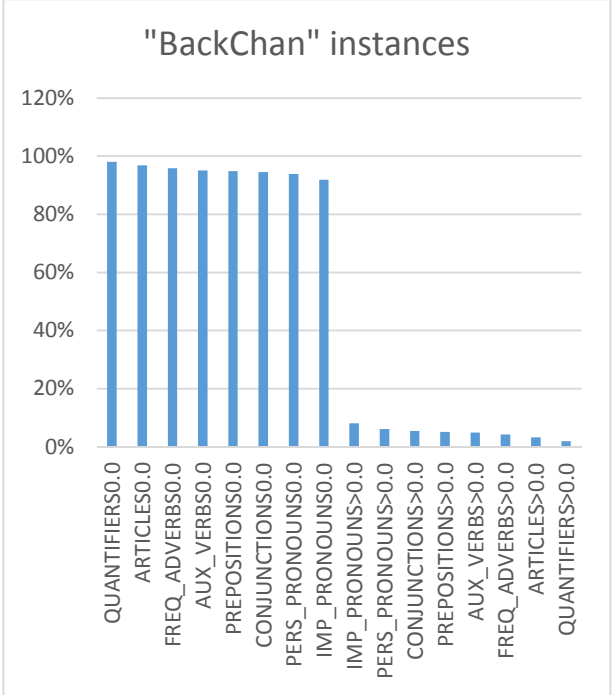

(a)




(b)




(c)




(d)

*Chart 12: Percentage of spurts per class that have each one of the coordination based properties shown in the horizontal axis. The properties have been sorted based on the frequency value of each one of them. The properties in the horizontal axis should be interpreted as in the following examples,e.g. a spurt that has the property QUANTIFIERS0.0 is a turn that contains words of the LIWC category QUANTIFIERS without seeing words of the same category in the previous turn. Thus the linguistic style coordination score for this category is 0.0. However, a property of type QUANTIFIERS>0 means that the score of coordination is greater than 0.0.*

For each one of the lexical categories, the percentage of instances that have positive coordination score is always higher for negative instances compared to positive ones. Similarly, for each category the percentage of instances with zero coordination score is always higher for the positive instances compared to the negative ones. This might be indicative of more frequent presence of coordination when expressing disagreement vs agreement and it is discussed further in the next section. Regarding back-channels, we can see that in general they do not coordinate linguistically with the previous turn. This is expected since back-channels are usually expressed through very specific lexical cues (such as “yeah”, “right”, “uhhuh”, “ok”, etc.).

# 6. Discussion

## 6.1. Feature set comparisons

In Chart 3 and Chart 4, we presented the results that are obtained when different feature sets are used. An obvious observation is that it seems easier to recognize agreement compared to disagreement. Our interpretation is that agreement is more likely to be expressed directly, while when people disagree they have the tendency to be less straightforward and use more words,for instance, trying to be polite or justify their opposing point of view. Compare the following examples from the corpus:

```
Agreement: this looks good
```

```
Disagreement: but the thing is i think that these people are of high enough level in their in their language proficiency that and i'm not objecting to accents i i'm i'm just thinking that we have to think at a at a higher level view could we have a language model a a grammar a grammar basically
```

Looking at Chart 1, we can see that the lexicon-based approach gives good performance. Comparing it with Chart 2 that presents the performance of each individual feature set, we can see that all feature sets give higher performance than the baseline for “Neg” class. However, when it comes to “Pos” class the performance improves only when using one of the following: a. unigrams/bigrams, b. lexicosyntactic features, c. non-individual conversational features. The other feature sets either keep the “Pos” performance almost equal to the baseline (baseconv, indivconv) or cause a more significant drop (datags, coordination). Looking at Chart 3, we can see that when combining various feature sets the performance for both classes is always better than the baseline.

Among the individual feature sets shown in Chart 3, the full set of lexicosyntactic features as well as just unigrams/bigrams seem to give the highest F-measure for detection of agreement, disagreement and back channeling (agreement: ~0.7, disagreement: ~0.64, backchannels: ~0.9). Also, coordination-based

features are giving a decent F-measure of ~0.42 for agreement and ~0.26 for disagreement. The corresponding F-measure values when using dialog act-based features are ~0.36 and ~0.18. As a result, both coordination and dialog act features appear to contain information that could assist our agreement/disagreement detection task. Standalone conversational features are not performing that well with disagreement detection (F-measure < 0.16), while their performance with agreement spurts seems to be much better (F-measure: ~0.55).

Looking at Chart 4 which uses the unigrams/bigrams as a baseline, we notice that adding certain features to that baseline feature set improves performance for agreement and backchannel detection while the one for disagreement detection is almost the same. More specifically F-measure for "Pos" class improves when adding one of the following feature sets on top of unigrams/bigrams:

1. Baseline conversational and non-individual conversational features;
2. Dialog act tag–based features
3. Coordination-based features
4. Non-individual conversational featureso
5. Individual conversational features

All feature sets added to unigrams/bigrams boosted back-channel performance. It is also worth mentioning that the measurements that used only a selected number of unigrams/bigrams in order to enable us to study the effect of the remaining feature sets, showed that individual conversational features seem to be improving the performance for "Neg" class while non-individual conversational features were associated with improvements in "Pos" and "BackChan" classes.

In the following tables, we see the confusion matrices for the feature sets presented in Chart 3 and Chart 4.

| | unigrams; bigrams | | | | Lexicosyntactic | | | | Datags | | | | coordination | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
| P | 71.43 | 7.14 | 12.5 | 8.93 | 67.26 | 8.33 | 13.1 | 11.31 | 36.31 | 15.48 | 9.52 | 38.69 | 64.88 | 8.93 | 18.45 | 7.74 |
| N | 19.09 | 64.55 | 0.91 | 15.45 | 22.73 | 59.09 | 0.91 | 17.27 | 53.64 | 13.64 | 6.36 | 26.36 | 61.82 | 19.09 | 7.27 | 11.82 |
| BC | 1.72 | 0 | 98.28 | 0 | 1.72 | 0 | 98.28 | 0 | 8.62 | 1.72 | 72.41 | 17.24 | 73.28 | 5.17 | 18.1 | 3.45 |
| O | 11.85 | 14.07 | 2.22 | 71.85 | 16.3 | 11.85 | 2.96 | 68.89 | 29.63 | 12.59 | 3.7 | 54.07 | 57.78 | 8.89 | 20.74 | 12.59 |

| | Baseconv | | | | Indivconv | | | | nonindivconv | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC O | O |
| P | 54.76 | 3.57 | 25.6 | 16.07 | 55.95 | 10.12 | 17.86 | 16.07 | 73.81 | 1.79 | 8.33 | 16.07 |
| N | 49.09 | 0 | 34.55 | 16.36 | 48.18 | 10.91 | 11.82 | 29.09 | 52.73 | 5.45 | 5.45 | 36.36 |
| BC | 9.48 | 0 | 73.28 | 17.24 | 6.9 | 0 | 90.52 | 2.59 | 0 | 0 | 98.28 | 1.72 |
| O | 34.07 | 0.74 | 35.56 | 29.63 | 43.7 | 12.59 | 9.63 | 34.07 | 46.67 | 5.93 | 5.19 | 42.22 |

*Table 11: Confusion matrix for the experiments shown in Chart 3. The value in each cell expresses the percentage (%) of the total number of. Interpretation of symbols: P : Pos class, N : Neg class, BC : BackChan class, O: Other*

| | unigrams; bigrams | | | | unigrams; bigrams; baseconv | | | | unigrams; bigrams; baseconv; nonindivconv | | | | unigrams; bigrams; baseconv; indivconv | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
| P | 71.43 | 7.14 | 12.5 | 8.93 | 68.45 | 11.9 | 10.12 | 9.52 | 74.4 | 11.31 | 2.38 | 11.9 | 70.24 | 9.52 | 12.5 | 7.74 |
| N | 19.09 | 64.55 | 0.91 | 15.45 | 19.09 | 65.45 | 0.91 | 14.55 | 15.45 | 61.82 | 0.91 | 21.82 | 19.09 | 64.55 | 0.91 | 15.45 |
| BC | 1.72 | 0 | 98.28 | 0 | 0.86 | 0 | 99.14 | 0 | 0.86 | 0 | 98.28 | 0.86 | 0.86 | 0 | 99.14 | 0 |
| O | 11.85 | 14.07 | 2.22 | 71.85 | 13.33 | 14.07 | 1.48 | 71.11 | 11.85 | 13.33 | 2.22 | 72.59 | 12.59 | 14.07 | 1.48 | 71.85 |

| | unigrams; bigrams; datags | | | | unigrams; bigrams; coordination | | | | unigrams; bigrams; nonindivconv | | | | unigrams; bigrams; indivconv | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
| P | 72.02 | 9.52 | 10.12 | 8.33 | 73.21 | 7.74 | 10.71 | 8.33 | 73.81 | 11.9 | 1.79 | 12.5 | 71.43 | 10.12 | 11.9 | 6.55 |
| N | 18.18 | 62.73 | 1.82 | 17.27 | 19.09 | 64.55 | 0.91 | 15.45 | 20.91 | 62.73 | 0 | 16.36 | 20 | 66.36 | 0.91 | 12.73 |
| BC | 0.86 | 0 | 99.14 | 0 | 1.72 | 0 | 98.28 | 0 | 0.86 | 0 | 99.14 | 0 | 0 | 0 | 100 | 0 |
| O | 11.85 | 14.81 | 2.22 | 71.11 | 13.33 | 13.33 | 2.22 | 71.11 | 12.59 | 13.33 | 2.22 | 71.85 | 11.11 | 14.07 | 3.7 | 71.11 |

*Table 12: Confusion matrix for the experiments shown in Chart 4. The value in each cell expresses the percentage (%) of the total number of instances. Interpretation of symbols: P : Pos class, N : Neg class, BC : BackChan class, O: Other*

Given Table 12, we can see that the use of coordination features by themselves seems to favor “Pos” class, since the most instances of any class (“Pos”, “Neg”, “BackChan”) seemed to be labeled as “Pos”. Apart

from this feature set, all the remaining ones seem to be having quite good performance classifying back-channeling spurts. However, all feature sets in the same table, except for unigrams/bigrams and lexicosyntactic, also appear to have the tendency to classify negative spurts as "Pos". Using lexicosyntactic or unigrams/bigrams features seem to be the only cases where the majority of the spurts of each class are labeled correctly.

The following tables show the results of Charts 3 and 4 in terms of overall accuracy and agreement/disagreement confusion and recovery. Similar performance measures had been used in (Hillard, Ostendorf & Shriberg 2003) and additionally "Pos" and "BackChan" have been merged to calculate these measures. (Hillard, Ostendorf & Shriberg 2003) experimented with unsupervised and supervised methods on the same corpus as this thesis used. Comparing our approach to the above research, there are many differences in the methodology followed to solve the problem. There are also differences in the number of folds we used during execution as well as the way we used to balance off the corpus in order to make sure that all classes are almost equally represented in the corpus. Despite those differences, our results are in accordance with those from the aforementioned paper.

| | **Accuracy** | **A/D Confusion** | **A/D Recovery** |
|---|---|---|---|
| **word unigrams; word bigrams** | 0.803403 | 0.085 | 0.83248731 |
| **lexicosyntactic** | 0.773157 | 0.100502513 | 0.822916667 |
| **datags** | 0.489603 | 0.274853801 | 0.472081218 |
| **coordination** | 0.536862 | 0.201244813 | 0.677664975 |
| **baseconv** | 0.512287 | 0.231132075 | 0.586294416 |
| **indivconv** | 0.557656 | 0.197149644 | 0.631979695 |
| **nonindivconv** | 0.599244 | 0.166253102 | 0.654822335 |

*Table 13: Performance of each feature set in terms of Accuracy/Confusion/Recovery*

| | Accuracy | A/D Confusion | A/D Recovery |
|---|---|---|---|
| **word unigrams; word bigrams** | 0.803403 | 0.085 | 0.83248731 |
| **word unigrams; word bigrams; baseconv** | 0.786389 | 0.104738155 | 0.812182741 |
| **Word unigrams; word bigrams; baseconv; nonindivconv** | 0.775047 | 0.095854922 | 0.791878173 |
| **word unigrams;word bigrams; baseconv;indivconv** | 0.799622 | 0.094527363 | 0.827411168 |
| **word unigrams; word bigrams; datags** | 0.79206 | 0.095 | 0.819796954 |
| **word unigrams; word bigrams; coordination** | 0.801512 | 0.087064677 | 0.83248731 |

*Table 14: Performance combinations of feature sets in terms of Accuracy/Confusion/Recovery*

The best results from the experiments in (Hillard, Ostendorf & Shriberg 2003) on the same corpus are the following:

| Overall Accuracy | A/D Confusion | A/D Recovery |
|---|---|---|
| 0.82 | 0.02 | 0.87 |

*Table 15: Results from the experiments in (Hillard, Ostendorf & Shriberg 2003) on the same corpus*

## 6.2. Discussion about dialog act tags

Looking at Chart 10, our first observation was that less than 80% of the spurts labeled as "BackChan" included the dialog act tag for backchannels. After looking into the data more closely, we found examples labeled as "BackChan" that do not contain the dialog act tag "b" and so we attributed the above observation to the fact that the definition of the tag "b" in MRDA corpus happens to be narrower compared to the "BackChan" class definition in the agreement/disagreement corpus. Some examples are presented here:

Example 1:

…

Speaker c9: like the front-end meeting and maybe a networking group meeting

Speaker c1: right

**Speaker c1: yep <- BackChan spurt that contains the dialog act tag s^aa**

…

Example 2:

…

Speaker c1: yeah we - we've only had three .
Speaker c1: so
**Speaker c9: okay <- BackChan spurt that contains the dialog act tag s^bk**
…

Example 3:

…
Speaker c3: i don't know if he's talked with you yet .
**Speaker c4: yeah <- BackChan spurt that contains the dialog act tag fg**
Speaker c3: but in sort of honing in on these different types
Speaker c4: i don't consi- - now i don't consider that possibility
…

Based on the same chart (Chart 10) we noticed the frequent presence of backchannel dialog act tags ("b" tags) in "Pos" (in 23% of the "Pos" instances) and "Neg" (in 12% of the "Neg" instances) spurts. This does not happen in the case of "Other" instances (2.8%). We would expect to find backchanneling spurts in the places where the speaker expresses that they are following along and paying attention. This can justify their presence in "Pos" spurts. To understand better the numbers we got, we looked for instances where there is agreement/disagreement when a "b" tag is encountered. Again, the conclusion we drew is that the definition of "b" tag in MRDA corpus is not identical to the "BackChan" tag used in the agreement/disagreement corpus.

Example of use of "b" dialog act tag in a "Pos" spurt:
…
Speaker c4: this is the worst case scenario
Speaker c2: yeah yeah **<- spurt labeled as "Pos" that is made of 2 dialog act tags "b"**

…

First example of use of “b” dialog act tag in a “Neg” spurt:

…
Speaker c3: yeah maybe it's not log distributed
Speaker c4: huh yeah **<- spurt labeled as “Neg” that is made of 2 dialog act tags “b”**

…

Second example of use of “b” dialog act tag in a “Neg” spurt:

…
Speaker c8: they're not double counted
Speaker c4: yeah
Speaker c1: ah but yeah **<- spurt labeled as “Neg” that is made of one dialog act tag “b”**
…

One final, yet unsurprising, observation in Chart 10 is that in both “Pos” and “Neg” spurts we notice a relatively high frequency of “fg” tags and (“fg”,”s”) tag pairs. This does not show up in the instances that belong to the remaining classes. Examples like the following are very common in conversations:

Example 1
“Neg” spurt: “yeah but - but what she's saying is which is right is le-”
(“fg” followed by “s” tag)

Example 2
”Pos” spurt: “right because if energy doesn’t matter there like i don’t think this is true but what if” (“fg” followed by two “s” tags).

## 6.3. Linguistic Style Coordination

Next we will look at some of the results from the experiments where only linguistic coordination features were used (Chart 14). For each one of the lexical categories the percentage of instances that have positive coordination score is always higher for negative instances compared to positive ones. Similarly, for each category the percentage of instances with zero coordination score is always higher for the positive instances compared to the negative ones. The fact that the instances labeled to be back-channeling spurts appear to have low percentages of coordination was something that we expected since backchannels

usually consist of certain lexical cues that don't necessarily have a linguistic relationship to the previous turn. The difference in the percentages for positive and negative instances shows that it is more likely for negative instances to belong to a turn that coordinates linguistically with the previous turn.

Based on the above, our first assumption was that the spurts expressing disagreement could be negating what was said in the preceding turn while expressing rejection. This would mean that a big part of the task of detecting if a turn disagrees with the previous one would be straightforward and done by looking for this type of negation. So, we took a closer look at the negative instances to detect the different ways disagreement can be expressed that would justify some kind of parallelism between the current and the preceding turn:

Pattern 1: Repeating part of the previous turn while at the same time negating it:

**Speaker c3:** well but see i find it interesting even if it wasn't any more because since we were dealing with this full duplex sort of thing in switchboard where it was just all separated out we just everything was just nice so that so the issue is in in a situation where th- that's

**Speaker c1:** so uhhuh well it's *not* really *nice* it depends what you're doing so if you were actually

Pattern 2: Repeating part of the previous turn while at the same time using the opposite word or a word expressing a contrasting meaning:

**Speaker c3:** s- so i- it *would be statistically incorrect* to conclude from this that adam talked too much or something
**Speaker cB:** no no actually that *would be* actually *statistically correct* but

Pattern 3: Repeating part of the spurt in the previous turn, but only in order to contradict it:

**Speaker c1**: well you you you don't have to study everybody individually but just simple case and the one that has the lot of data associated with it

**Speaker c3**: well to study the simplest case

Pattern 4: There are sentences that might have any of the features described in 1-3 but at the same time they maintain similar discourse relationships among the sentences they use to express an opinion:

**Speaker c2:** it- it would be uh uh very difficult to to put uh a lot of uh head phones uh in different people when you have to to record only with uh this kind of uh d- device

**Speaker c1:** yeah but i think if we if we want to just record with the tabletop microphones that's easy

Of course disagreement is also expressed by using short words or phrases like "No", "I do not think", "I feel the opposite need", by starting a spurt with "But", "Well", "Yeah but" or by just expressing an opinion that is different to what was expressed previously. However, at this point we are trying to detect the cases where the disagreeing turn and the preceding turn demonstrate some kind of parallelism. The frequency of patterns 1-4 in the agreement/disagreement corpus is as following:

| Pattern | Percentage of Negative Instances |
|---|---|
| Pattern 1 | 7.2% |
| Pattern 2 | 5.5% |
| Pattern 3 | 3.6% |
| Pattern 4 | 1.8% |

*Table 16: Percentage of negative instances that follow the disagreement patterns that exhibit some parallelism among a disagreeing turn and the preceding one.*

Based on the examples we saw in the corpus, it seems that when speakers disagree with the previous speaker they are likely to repeat part of the statement they disagree with. The impression we get is that this happens because people are trying to be more convincing and clear by addressing one by one the points brought up by the speaker they disagree with. Also - and for the same reason - they are likely to repeat the same statement replacing only a couple of words needed to invert the meaning; by doing this they emphasize their objection and state their point. The above conclusions about more frequent

presence of linguistic style coordination in a disagreeing turn seem to be in agreement with the conclusions in (Danescu-Niculescu-Mizil, Lee, Pang & Kleinberg 2012) where it is stated that “the need to convince someone who disagrees with you creates a form of dependence” and where it is shown that these dependents tend to coordinate more towards the people they depend on.

# 7. Conclusions and future work

We saw that the task of detecting agreement/disagreement in a multiparty conversation is not a trivial task that could be easily performed by using a subjectivity dictionary. The performance of this approach is easily beaten by many of the feature sets we use in our approach. Additionally, the use of bigrams/unigrams as well as more lexicosyntactic features has very good performance for our task. However, these results can further improve by adding more feature sets, especially for "Pos" and "BackChan" classes, as we discussed in the "Discussion" section.

Regarding future research, we assumed for the coordination experiments that one turn might or might not demonstrate linguistic style coordination with the preceding turn only. In reality, one turn expressing a certain viewpoint might be followed by two or more turns of different speakers expressing either agreement or disagreement with that turn. In this case, we will need to examine the coordination relationships between both of these turns and the first turn. In other words, examining a non-linear chain model for the turns in a meeting is something that is considered for future research.

Additionally, the study of conversational features needs to be continued. We saw that individual and non-individual features are promising and can improve the performance of lexicosyntactic features. The understanding of the contribution of each one of the conversational features separately can lead to better solutions to the problem of agreement and disagreement detection. Furthermore, since the agreement/disagreement corpus was only a subset of the MRDA corpus, potential annotation of the whole corpus and collection of data for each speaker over longer periods of time could also result in more in depth conclusions regarding the contribution of the per-individual conversational features.

# 8. References


Danescu-Niculescu-Mizil, C., Lee, L., Pang, B. & Kleinberg, J. (2012). Echoes of Power: Language Effects and Power Differences in Social Interaction. *Proceedings of WWW 2012, pp 699-708, Lyon, France, April 16–20, 2012*

Dhillon, R., Bhagat, S., Carvey, H., & Shriberg, E. (2004). Meeting Recorder Project: Dialog Act Labeling Guide. *ICSI Technical Report TR-04-002, International Computer Science Institute.*

Hillard, D., Ostendorf, M. & Shriberg, E. (2003). Detection of Agreement vs. Disagreement in Meetings: Training With Unlabeled Data. *Proceedings of HLT-NAACL 2003, pp 34-36, Edmonton, Canada, May 27 – June 1, 2003*

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 168-177, Seattle, WA, USA, August 22 - 25, 2004*

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." *http://mallet.cs.umass.edu. 2002.*

Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., Stolcke, A. (2001). The meeting project at ICSI. *Proc. Conf. on Human Language Technology*, *pages 246–252, March 2001.*

Murray, G. & Carenini, G. (2011). Subjectivity detection in spoken and written conversations. *Journal of Natural Language Engineering, 17, pp 397-418 doi:10.1017/ S1351324910000264*

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*. 2(1-2), 1-135. doi: 10.1561/1500000001

Pennebaker, J. W., Booth, R. J., and Francis, M. E (2007). Linguistic inquiry and word count (LIWC): A computerized text analysis program. *http://www.liwc.net/, 2007*

Raaijmakers, S., Truong, K. & Wilson, T. (2008). Multimodal Subjectivity Analysis of Multiparty Conversation. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 466-474, Honolulu, October 2008.*

Raaijmakers, S. & Wilson, T. (2008). Comparing word, character, and phoneme n-grams for subjective utterance recognition. *Proceedings of INTERSPEECH 2008 conference, pages 1614-1617, Brisbane, Australia, September 2008.*

Ruppenhofer, J., Somasundaran,S. & Wiebe, J. (2008a). Discourse Level Opinion Interpretation. *International Conference on Computational Linguistics, pages 801-808, Manchester, August 2008.*

Ruppenhofer, J., Somasundaran,S. & Wiebe, J. (2008b). Discourse Level Opinion Relations: An Annotation Study, *SIGdial Workshop on Discourse and Dialogue, pages 129-137, Columbus, Ohio, June 2008.*

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J. & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proceedings of 5th SIGdial workshop on discourse and dialogue at HLT-NAACL 2004, pp 97 – 100, Boston, Massachusetts, USA, May 2 -7, 2004.*

Somasundaran, S., Namata, G., Getoor, L. & Wiebe, J. (2009). Opinion Graphs for Polarity and Discourse Classification. *TextGraphs-4: Graph-based Methods for Natural Language Processing, pages 66-74, Singapore, August 2009.*

Somasundaran, S., Namata, G., Wiebe, J. & Getoor, L. (2009). Supervised and Unsupervised Methods in Employing Discourse Relations for Improving Opinion Polarity Classification. *Conference on Empirical Methods in Natural Language Processing, pages 170-179, Singapore, August 2009*

The Stanford Natural Language Group (2010). Stanford CoreNLP: A Suite of Core NLP Tools (Version 3.2.0) [Software*]. Available from http://nlp.stanford.edu/software/corenlp.shtml*

Wang, D. & Liu, Y. (2012). A cross-corpus study of subjectivity identification using unsupervised learning. Journal of *Natural Language Engineering, 18, pp 375-397 doi:10.1017/S1351324911000234*

Wilson, T., Wiebe, J. and Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proc. of HLT-EMNLP-2005.*

Wilson, T. (2008). Annotating Subjective Content in Meetings. *Proceedings from the Language Resources and Evaluation Conference (LREC-2008).*

# 9. Appendix

## 9.1. Explanation of symbols used for feature representation

In this part you will find an explanation of various symbols used to describe features used in this thesis.

**Dialog act tag-based features:**

*daUnigrams.<dialog-act-tag>* : The general part of a dialog act tag

*daBigrams.<dialog-act-tag>.<dialog-act-tag>*: A pair of the general part of two consecutive dialog act tags

**Lexicosyntactic features:**

WordUnigrams.<word>: Word unigram

WordBigrams.<word>.<word>: Word bigram

WordSkipBigrams.1.<word>.<word> : Word bigrams consisting of two inconsecutive words (only one word exists between them)

POSBigrams.<pos-tag>.<pos-tag>: Bigrams consisting of POS tags

POSSkipBigrams.1. .<pos-tag>.<pos-tag>: Bigrams consisting of two inconsecutive POS tags (only one tag exists between them)

## 9.2. Explanation of various dialog act tags

This table is copied from the manual for the MRDA corpus annotation (Dhillon, Bhagat, Carvey & Shriberg 2004) and its intention is to help the reader understand the symbols mentioned in this thesis.

**Group 1: Statements**

- s Statement

**Group 2: Questions**

- qy Y/N Question
- qw Wh-Question
- qr Or Question
- qrr Or Clause After Y/N Question
- qo Open-ended Question
- qh Rhetorical Question

**Group 3: Floor Mechanisms**

- fg Floor Grabber
- fh Floor Holder
- h Hold

**Group 4: Backchannels and Acknowledgements**

- b Backchannel
- bk Acknowledgement
- ba Assessment/Appreciation
- bh Rhetorical Question Backchannel

**Group 5: Responses**

**Positive**

- aa Accept
- aap Partial Accept
- na Affirmative Answer

**Negative**

- ar Reject
- arp Partial Reject
- nd Dispreferred Answer
- ng Negative Answer

**Uncertain**

- am Maybe
- no No Knowledge

**Group 6: Action Motivators**

- co Command
- cs Suggestion
- cc Commitment

**Group 7: Checks**

- f "Follow Me"
- br Repetition Request
- bu Understanding Check

**Group 8: Restated Information**

**Repetition**

- r Repeat
- m Mimic
- bs Summary

**Correction**

- bc Correct Misspeaking
- bsc Self-Correct Misspeaking

**Group 9: Supportive Functions**

- df Defending/Explanation
- e Elaboration
- 2 Collaborative Completion

**Group 10: Politeness Mechanisms**

- bd Downplayer
- by Sympathy
- fa Apology
- ft Thanks
- fw Welcome

**Group 11: Further Descriptions**

- fe Exclamation
- t About-Task
- tc Topic Change
- j Joke
- t1 Self Talk
- t3 Third Party Talk
- d Declarative Question
- g Tag Question
- rt Rising Tone

**Group 12: Disruption Forms**

- % Indecipherable
- %- Interrupted
- %-- Abandoned
- x Nonspeech

**Group 13: Nonlabeled**

- z Nonlabeled

s

## 9.3. Tables

| Feature ID | Description |
|---|---|
| **MXS** | max *Sprob* score |
| **MNS** | mean *Sprob* score |
| **SMS** | sum of *Sprob* scores |
| **MXT** | max *Tprob* score |
| **MNT** | mean *Tprob* score |
| **SMT** | sum of *Tprob* scores |
| **TLOC** | position in turn |
| **CLOC** | position in conv. |
| **SLEN** | word count, globally normalized |
| **SLEN2** | word count, locally normalized |
| **TPOS1** | time from beg. of conv. to turn |
| **TPOS2** | time from turn to end of conv. |
| **DOM** | participant dominance in words |
| **COS1** | cosine of conv. splits, w/ *Sprob* |
| **COS2** | cosine of conv. splits, w/ *Tprob* |
| **PENT** | entropy of conv. up to sentence |
| **SENT** | entropy of conv. after the sentence |
| **THISENT** | entropy of current sentence |
| **PPAU** | time btwn. current and prior turn |
| **SPAU** | time btwn. current and next turn |
| **BEGAUTH** | is first participant (0/1) |
| **CWS** | rough ClueWordScore (cohesion) |
| **CENT1** | cos. of sentence & conv., w/ *Sprob* |
| **CENT2** | cos. of sentence & conv., w/ *Tprob* |

*Table 1: Table taken from Murray & Carenini 2011 – page 11*

| | |
|---|---|
| c2, s, and so you take that<br><br>c2, s, and then he's - he's uh measuring at the frame level<br><br>c2, s.%--, still at the frame level of what<br><br>c3, b, right<br><br>c2, s, and then - and then just uh normalizing with that larger amount<br><br>c2, s^rt, um and - but one thing he was pointing out is when he - he looked at a bunch of examples in log domain it is actually pretty hard to see the change | c2, and so you take that and then he's he's uh measuring at the frame level still at the frame level of what and then and then just uh normalizing with that larger amount<br><br>c3, right<br><br>c2, um and but one thing he was pointing out is when he he looked at a bunch of examples in log domain it is actually pretty hard to see the change and you can sort of see that because of j- of just putting it on the board that if you sort of have log x plus log x that's the log of x plus the log of two and it's just you know it it diminishes the effect of having two of them |

| | |
|---|---|
| c2, s^rt, and you can sort of see that because of j- - of just putting it on the board that if you sort of have log x plus log x that's the log of x plus the log of two<br><br>c4, b, yep<br><br>c3, fg\|s^cs, yeah \| maybe it's not log distributed<br><br>c4, b, huh<br><br>c4, b, yeah<br><br>c2, s, and it's just you know it – it diminishes the effect of having two of them | c4, yep<br><br>c3, yeah maybe it's not log distributed<br><br>c4, huh yeah |

*Table 2: Example snippet from one meeting that has been split into dialog acts (first column) as well as spurts (second column)*

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 1114 |
| BackChan | 405 |

*Table 3: Distribution of the four classes in the corpus (Pos, Neg, Other, BackChan)*

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 1114 |
| BackChan | 405 |

*Table 4: Initial composition of the agreement/disagreement corpus*

| Class Name | Number of Instances |
|---|---|
| Pos | 168 |
| Neg | 110 |
| Other | 135 |
| BackChan | 116 |

*Table 5: Composition of the agreement/disagreement corpus after downsampling "Other" and "BackChan" classes*

| Class | Baseline F-measure | Best F-measure | Feature set whose addition caused the best F-measure | Worst F-measure | Feature set whose addition caused the worst F-measure |
|---|---|---|---|---|---|
| Pos | 0.733945 | 0.764526 | Baseline and non-individual conversational features | 0.712074 | Baseline conversational features |
| Neg | 0.669811 | 0.669811 | Coordination-based features (but even without the additional feature set the baseline has the same performance) | 0.632558 | Baseline and non-individual conversational features |
| BackChan | 0.894118 | 0.970464 | Non-individual conversational features | - | The addition of any of the feature sets improved the performance for BackChan class |

*Table 6: Summary of the effect that different feature sets have on F-measure when added to the "unigrams/bigrams" feature set*

| |
|---|
| WordUnigrams.yeah |
| WordUnigrams.you |
| WordUnigrams.the |
| WordUnigrams.to |
| WordUnigrams.that |
| WordUnigrams.'s |
| WordUnigrams.of |
| WordUnigrams.a |
| WordUnigrams.and |
| WordUnigrams.but |
| WordUnigrams.i |
| WordUnigrams.it |
| WordUnigrams.so |
| WordUnigrams.right |
| WordUnigrams.is |
| WordUnigrams.uh |

| WordUnigrams.n't |
|---|

*Table 7: Merged set of unigrams/bigrams features that were used by the various folds*

| Pos | Neg | BackChan |
|---|---|---|
| 1. Non-individual conversational features<br>2. Baseline and non-individual conversational features<br>3. Baseline and individual conversational features<br>4. Baseline conversational<br>5. Individual conversational features<br>6. DA tag-based features<br>7. Coordination-based features | 1. Individual conversational features<br>2. Baseline and individual conversational features<br>3. DA tag-based features<br>4. Baseline and non-individual conversational features<br>5. Non-individual conversational features<br>6. Coordination-based features<br>7. Baseline conversational | 1. Non-individual conversational features<br>2. Baseline and non-individual conversational features<br>3. Baseline and individual conversational features<br>4. Individual conversational features<br>5. Baseline conversational<br>6. DA tag-based features<br>7. Coordination-based features |

*Table 8: All features sets are sorted from best to worst based on the effect they have on the F-measure of each class*

| | |
|---|---|
| 1. WordUnigrams.but<br>2. WordUnigrams.yeah<br>3. WordUnigrams.well<br>4. WordUnigrams.the | 16. POSSkipBigrams.1.IN.NN<br>17. POSBigrams.IN.PRP<br>18. WordUnigrams.in<br>19. POSBigrams.DT.NNS |

| | |
|---|---|
| 5. POSBigrams.DT.NN | 20. WordUnigrams.'s |
| 6. WordUnigrams.uhhuh | 21. WordUnigrams.it |
| 7. WordUnigrams.i | 22. WordUnigrams.no |
| 8. WordUnigrams.and | 23. POSBigrams.DT.JJ |
| 9. WordUnigrams.you | 24. POSSkipBigrams.1.DT.NN |
| 10. POSBigrams.PRP.VBP | 25. WordBigrams.but.i |
| 11. POSBigrams.IN.DT | 26. POSBigrams.NN.CC |
| 12. WordUnigrams.that | 27. POSSkipBigrams.1.NN.IN |
| 13. POSBigrams.NN.IN | 28. POSSkipBigrams.1.IN.VBP |
| 14. WordUnigrams.of | 29. WordUnigrams.uh |
| 15. POSSkipBigrams.1.DT.IN | 30. POSBigrams.JJ.NN |

*Table 9: Top 30 features in the list of lexicosyntactic features ordered by decreasing Chi Square value*

| | |
|---|---|
| 1. daUnigrams.s | 16. daBigrams.s.b |
| 2. daUnigrams.b | 17. daUnigrams.%-- |
| 3. daBigrams.s.s | 18. daBigrams.s.%-- |
| 4. daUnigrams.fg | 19. daBigrams.z.z |
| 5. daBigrams.b.s | 20. daBigrams.s.qy |
| 6. daBigrams.fg.s | 21. daBigrams.qy.s |
| 7. daBigrams.b.b | 22. daUnigrams.h |
| 8. daUnigrams.fh | 23. daBigrams.b.fh |
| 9. daUnigrams.qy | 24. daBigrams.%--.qy |
| 10. daUnigrams.z | 25. daBigrams.b.%- |
| 11. daBigrams.fh.s | 26. daBigrams.qy.qh |
| 12. daUnigrams.qw | 27. daUnigrams.qo |
| 13. daUnigrams.%- | 28. daBigrams.s.fg |
| 14. daBigrams.s.fh | 29. daBigrams.fh.b |
| 15. daBigrams.h.s | 30. daBigrams.%--.b |

*Table 10: Top 30 features in the list of dialog act tag-based features ordered by decreasing Chi Square value*

| | unigrams; bigrams | | | | Lexicosyntactic | | | | Datags | | | | coordination | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
| P | 71.43 | 7.14 | 12.5 | 8.93 | 67.26 | 8.33 | 13.1 | 11.31 | 36.31 | 15.48 | 9.52 | 38.69 | 64.88 | 8.93 | 18.45 | 7.74 |
| N | 19.09 | 64.55 | 0.91 | 15.45 | 22.73 | 59.09 | 0.91 | 17.27 | 53.64 | 13.64 | 6.36 | 26.36 | 61.82 | 19.09 | 7.27 | 11.82 |
| BC | 1.72 | 0 | 98.28 | 0 | 1.72 | 0 | 98.28 | 0 | 8.62 | 1.72 | 72.41 | 17.24 | 73.28 | 5.17 | 18.1 | 3.45 |
| O | 11.85 | 14.07 | 2.22 | 71.85 | 16.3 | 11.85 | 2.96 | 68.89 | 29.63 | 12.59 | 3.7 | 54.07 | 57.78 | 8.89 | 20.74 | 12.59 |

| | Baseconv | | | | Indivconv | | | | nonindivconv | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC O | O |
| P | 54.76 | 3.57 | 25.6 | 16.07 | 55.95 | 10.12 | 17.86 | 16.07 | 73.81 | 1.79 | 8.33 | 16.07 |
| N | 49.09 | 0 | 34.55 | 16.36 | 48.18 | 10.91 | 11.82 | 29.09 | 52.73 | 5.45 | 5.45 | 36.36 |
| BC | 9.48 | 0 | 73.28 | 17.24 | 6.9 | 0 | 90.52 | 2.59 | 0 | 0 | 98.28 | 1.72 |
| O | 34.07 | 0.74 | 35.56 | 29.63 | 43.7 | 12.59 | 9.63 | 34.07 | 46.67 | 5.93 | 5.19 | 42.22 |

*Table 11: Confusion matrix for the experiments shown in Chart 3. The value in each cell expresses the percentage (%) of the total number of. Interpretation of symbols: P : Pos class, N : Neg class, BC : BackChan class, O: Other*

| | unigrams; bigrams | unigrams; bigrams; baseconv | unigrams; bigrams; baseconv; nonindivconv | unigrams; bigrams; baseconv; indivconv |
|---|---|---|---|---|

| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 71.43 | 7.14 | 12.5 | 8.93 | 68.45 | 11.9 | 10.12 | 9.52 | 74.4 | 11.31 | 2.38 | 11.9 | 70.24 | 9.52 | 12.5 | 7.74 |
| N | 19.09 | 64.55 | 0.91 | 15.45 | 19.09 | 65.45 | 0.91 | 14.55 | 15.45 | 61.82 | 0.91 | 21.82 | 19.09 | 64.55 | 0.91 | 15.45 |
| BC | 1.72 | 0 | 98.28 | 0 | 0.86 | 0 | 99.14 | 0 | 0.86 | 0 | 98.28 | 0.86 | 0.86 | 0 | 99.14 | 0 |
| O | 11.85 | 14.07 | 2.22 | 71.85 | 13.33 | 14.07 | 1.48 | 71.11 | 11.85 | 13.33 | 2.22 | 72.59 | 12.59 | 14.07 | 1.48 | 71.85 |

| | unigrams; bigrams; datags | | | | unigrams; bigrams; coordination | | | | unigrams; bigrams; nonindivconv | | | | unigrams; bigrams; indivconv | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | N | BC | O | P | N | BC | O | P | N | BC | O | P | N | BC | O |
| P | 72.02 | 9.52 | 10.12 | 8.33 | 73.21 | 7.74 | 10.71 | 8.33 | 73.81 | 11.9 | 1.79 | 12.5 | 71.43 | 10.12 | 11.9 | 6.55 |
| N | 18.18 | 62.73 | 1.82 | 17.27 | 19.09 | 64.55 | 0.91 | 15.45 | 20.91 | 62.73 | 0 | 16.36 | 20 | 66.36 | 0.91 | 12.73 |
| BC | 0.86 | 0 | 99.14 | 0 | 1.72 | 0 | 98.28 | 0 | 0.86 | 0 | 99.14 | 0 | 0 | 0 | 100 | 0 |
| O | 11.85 | 14.81 | 2.22 | 71.11 | 13.33 | 13.33 | 2.22 | 71.11 | 12.59 | 13.33 | 2.22 | 71.85 | 11.11 | 14.07 | 3.7 | 71.11 |

*Table 12: Confusion matrix for the experiments shown in Chart 4. The value in each cell expresses the percentage (%) of the total number of instances. Interpretation of symbols: P : Pos class, N : Neg class, BC : BackChan class, O: Other*

| | **Accuracy** | **A/D Confusion** | **A/D Recovery** |
|---|---|---|---|
| **word unigrams; word bigrams** | 0.803403 | 0.085 | 0.83248731 |
| **lexicosyntactic** | 0.773157 | 0.100502513 | 0.822916667 |
| **datags** | 0.489603 | 0.274853801 | 0.472081218 |
| **coordination** | 0.536862 | 0.201244813 | 0.677664975 |
| **baseconv** | 0.512287 | 0.231132075 | 0.586294416 |
| **indivconv** | 0.557656 | 0.197149644 | 0.631979695 |
| **nonindivconv** | 0.599244 | 0.166253102 | 0.654822335 |

*Table 13: Performance of each feature set in terms of Accuracy/Confusion/Recovery*

| | **Accuracy** | **A/D Confusion** | **A/D Recovery** |
|---|---|---|---|
| **word unigrams; word bigrams** | 0.803403 | 0.085 | 0.83248731 |
| **word unigrams; word bigrams; baseconv** | 0.786389 | 0.104738155 | 0.812182741 |
| **Word unigrams; word bigrams; baseconv; nonindivconv** | 0.775047 | 0.095854922 | 0.791878173 |
| **word unigrams;word bigrams; baseconv;indivconv** | 0.799622 | 0.094527363 | 0.827411168 |
| **word unigrams; word bigrams; datags** | 0.79206 | 0.095 | 0.819796954 |
| **word unigrams; word bigrams; coordination** | 0.801512 | 0.087064677 | 0.83248731 |

*Table 14: Performance combinations of feature sets in terms of Accuracy/Confusion/Recovery*

| Overall Accuracy | A/D Confusion | A/D Recovery |
|---|---|---|
| 0.82 | 0.02 | 0.87 |

*Table 15: Results from the experiments in (Hillard, Ostendorf & Shriberg 2003) on the same corpus*

| **Pattern** | **Percentage of Negative Instances** |
|---|---|
| Pattern 1 | 7.2% |
| Pattern 2 | 5.5% |
| Pattern 3 | 3.6% |
| Pattern 4 | 1.8% |

*Table 16: Percentage of negative instances that follow the disagreement patterns that exhibit some parallelism among a disagreeing turn and the preceding one.*

## 9.4. Charts

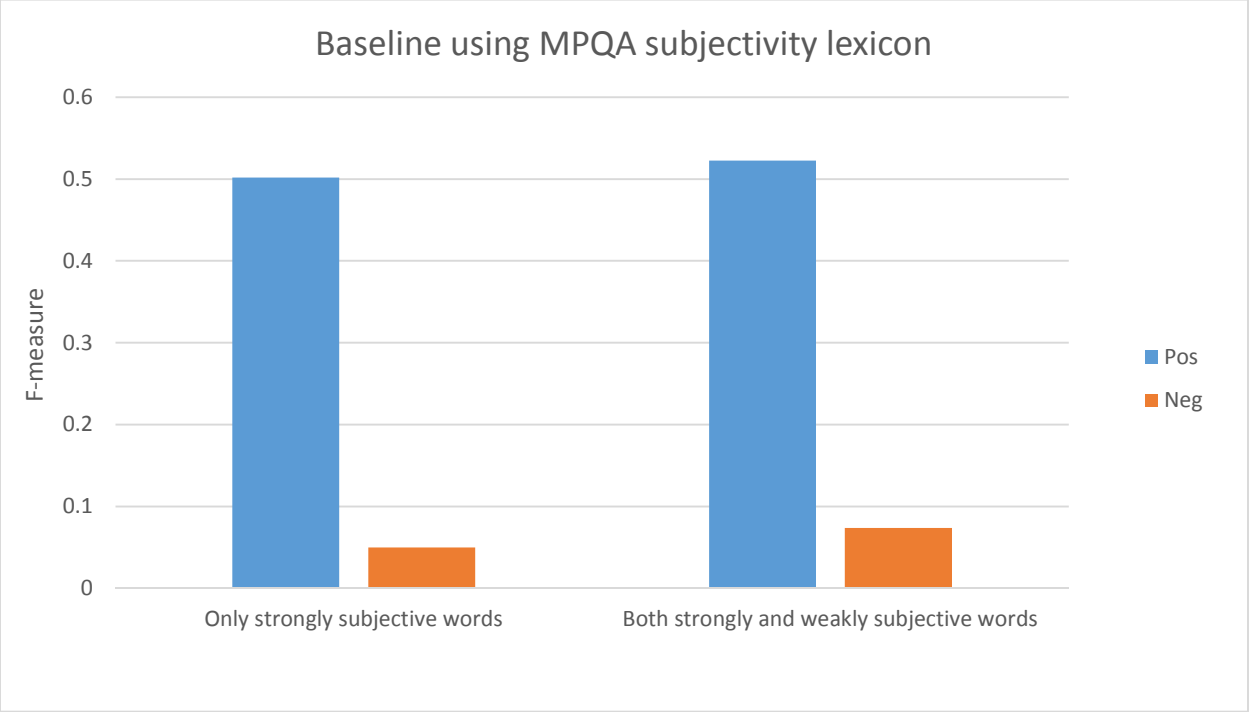



*Chart 1: Baseline measurements using the MPQA subjectivity lexicon*

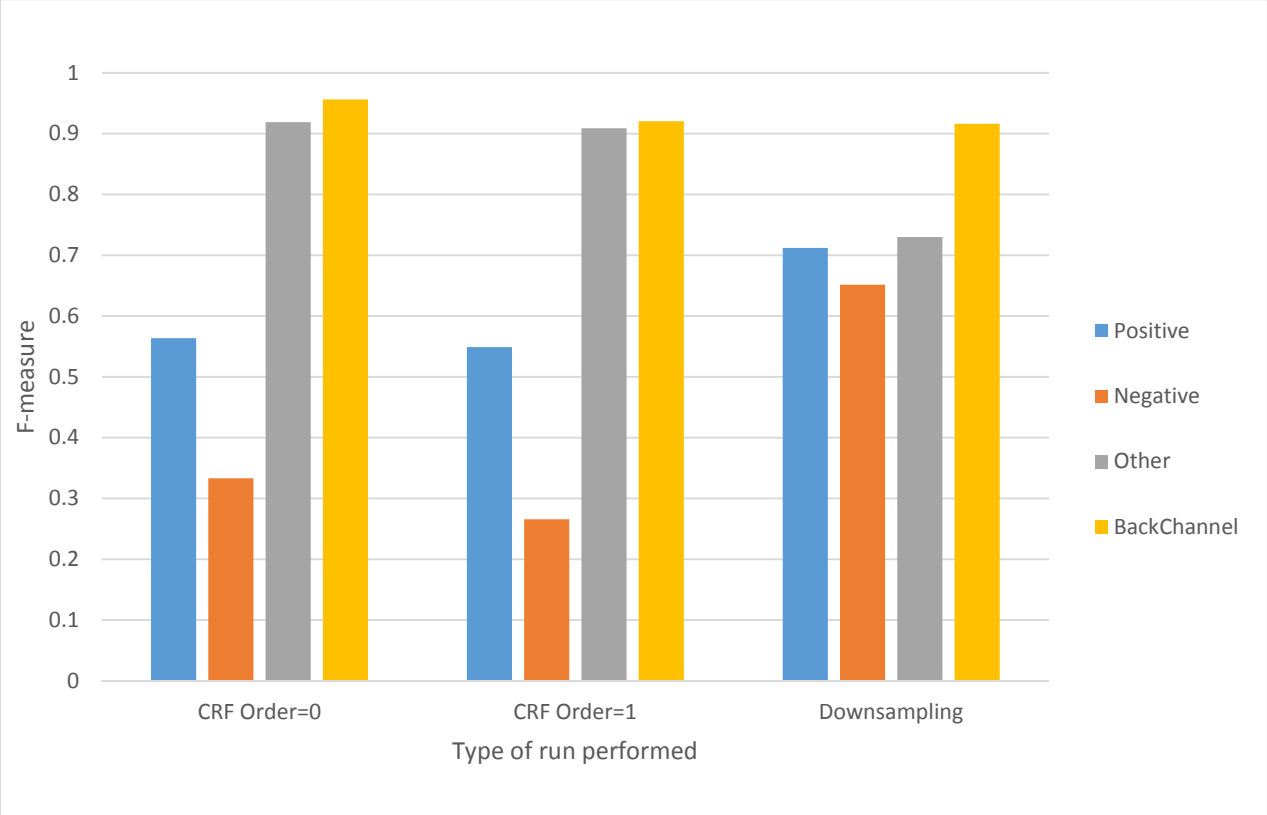


*Chart 2: Behavior of the F-measure for all classes and for different types of runs. The features used for each run are: unigrams, bigrams and baseline conversational (as defined in the "Methodology" section).*

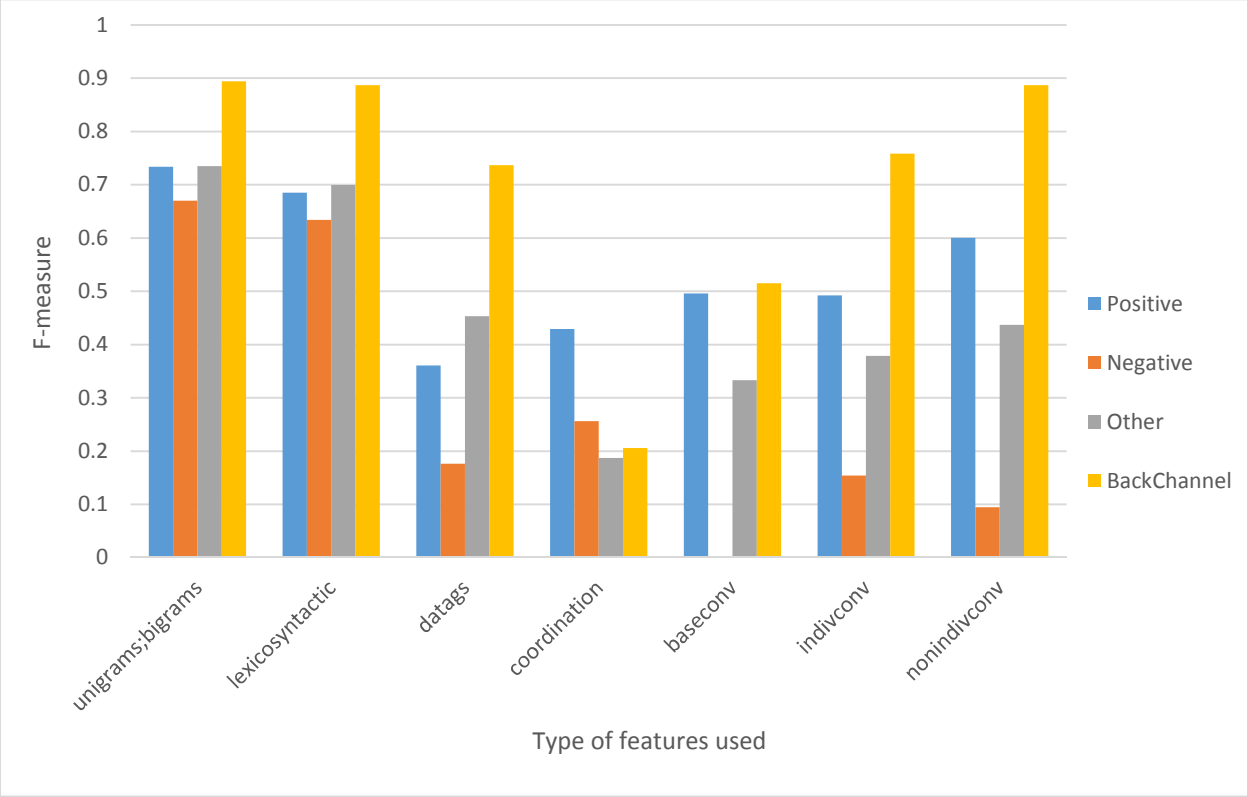



*Chart 3: F-measure for each class when using different feature sets. The features shown on the horizontal axis from left to right correspond to the following feature sets that have been discussed in the "Methodology" section: 1. Unigrams, bigrams; 2. Lexicosyntactic features; 3. Features based on dialog act annotations; 4. Coordination features; 5. Baseline conversational features; 6. Individual conversational feature; 7. Non-individual conversational features.*

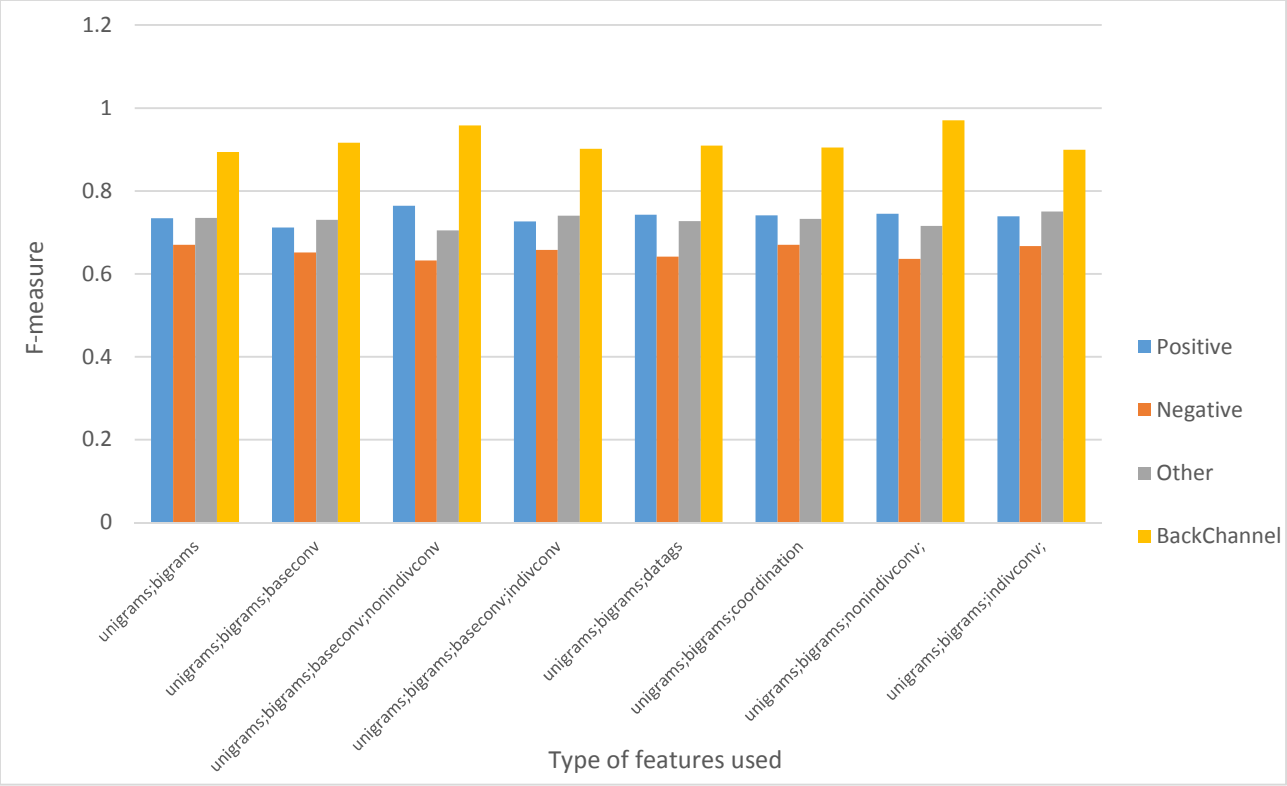



*Chart 4: F-measure for each class when the combination of unigrams/bigrams and other features is used.*

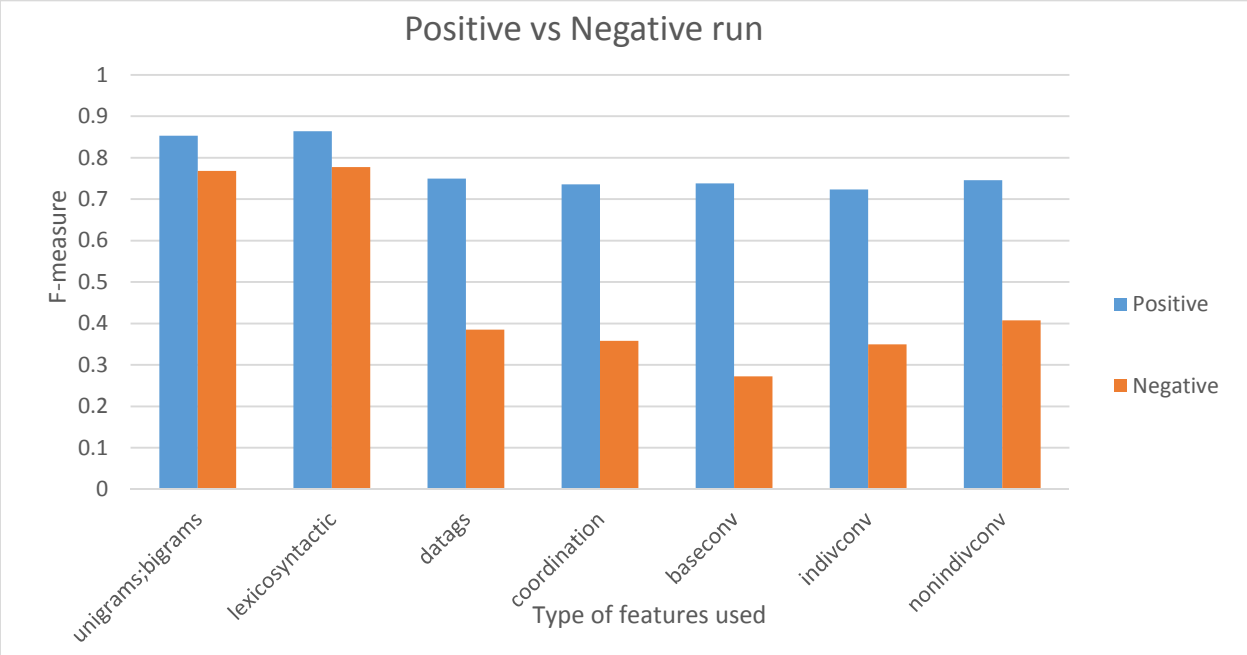



*Chart 5: F-measure for each class when using different feature sets. For these runs only Pos and Neg instances were used. The features shown on the horizontal axis from left to right correspond to the following feature sets that have been discussed in the "Methodology" section: 1. Unigrams/bigrams, 2. Lexicosyntactic features, 3. Features based on dialog act existing annotations, 4. Coordination features, 5. Baseline conversational features, 6. Individual conversational features, 7. Non-individual conversational features.*

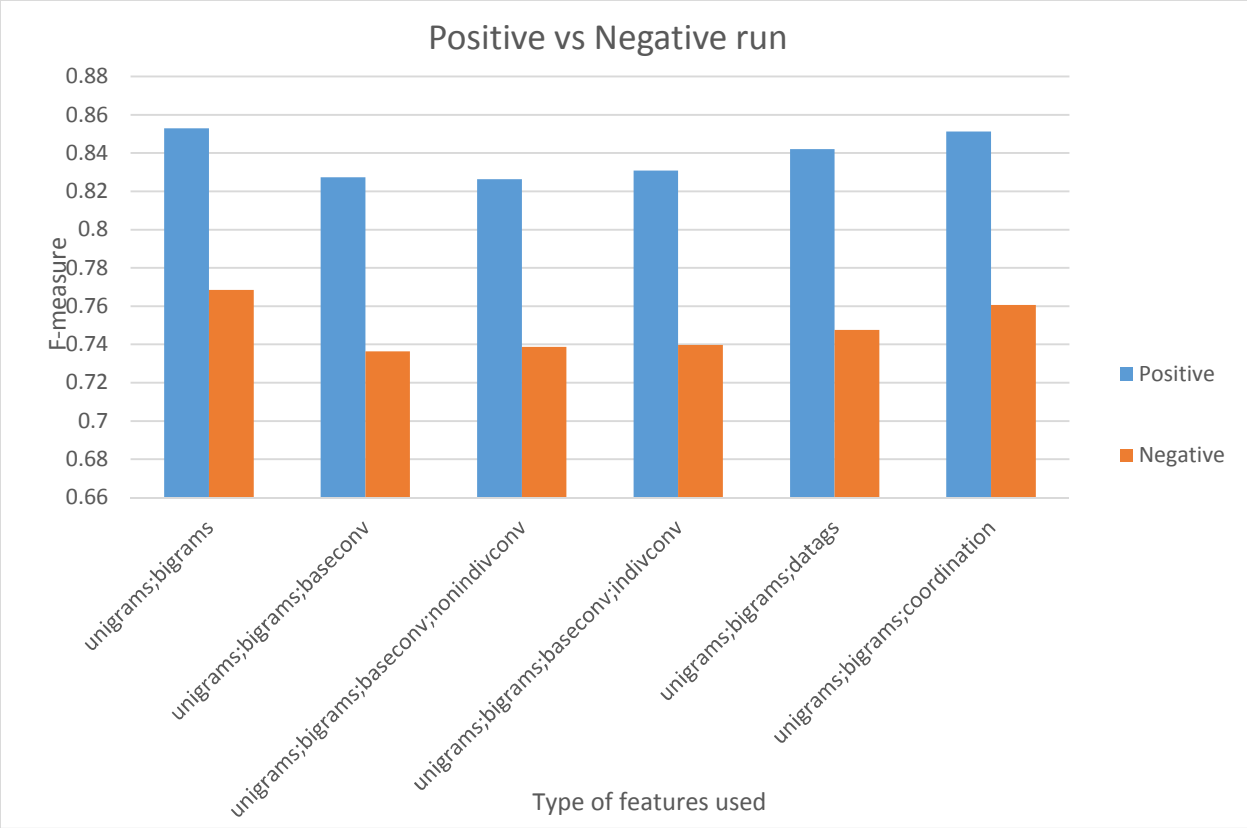



*Chart 6: F-measure for each class when the combination of unigrams/bigrams and other features is used. For these runs only Pos and Neg instances were used.*

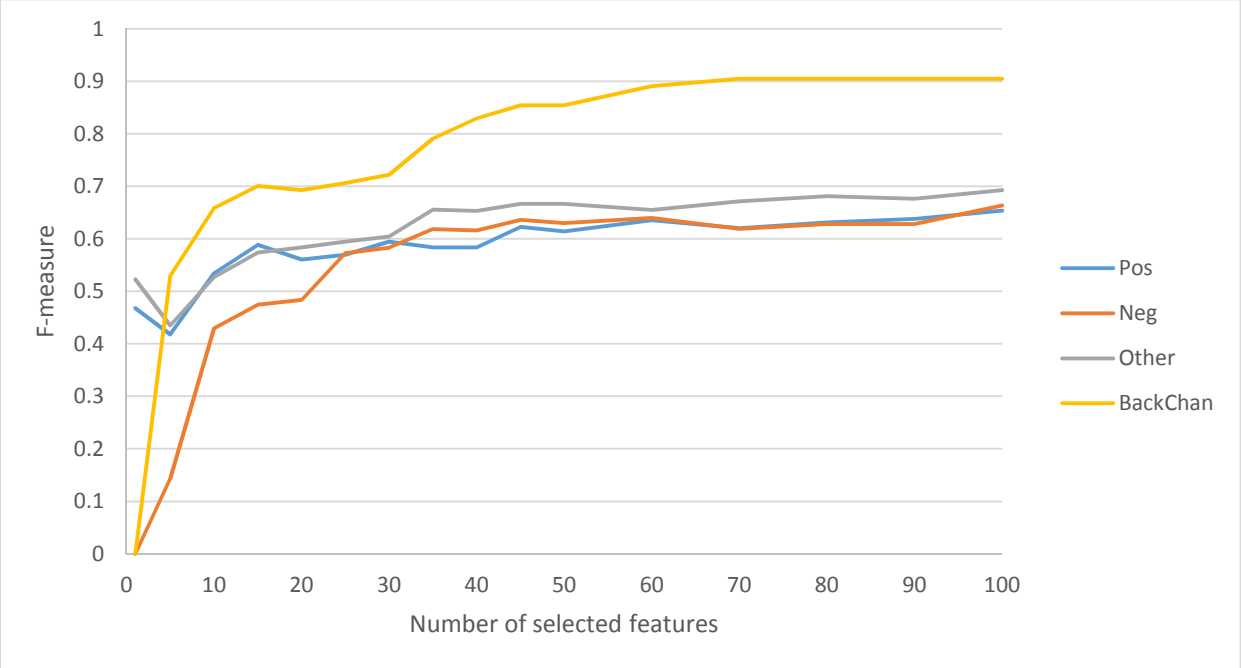



*Chart 7: Changes in F-measure for each class as the number of unigrams/bigrams selected gets different values*

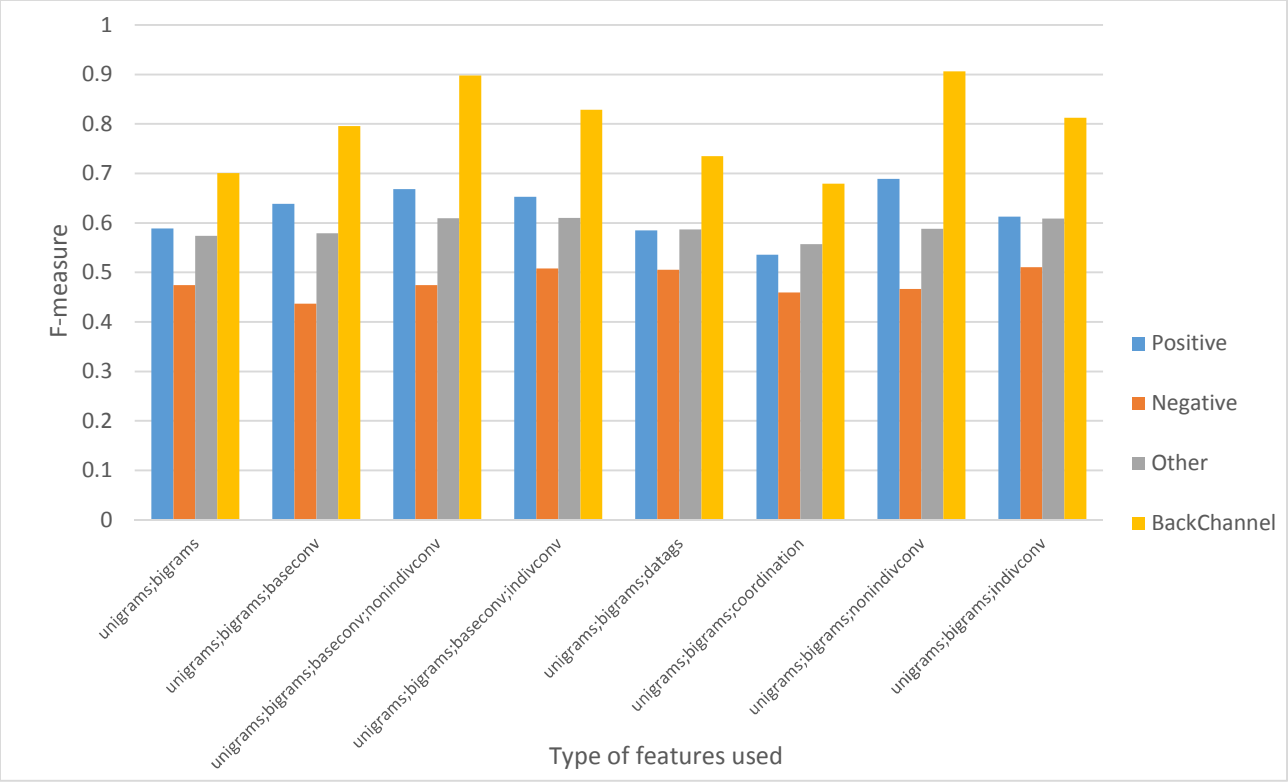



*Chart 8: F-measure for each class when the combination of unigrams/bigrams and other features is used and in the case where only the 15 most frequent unigrams/bigrams are kept in each fold.*

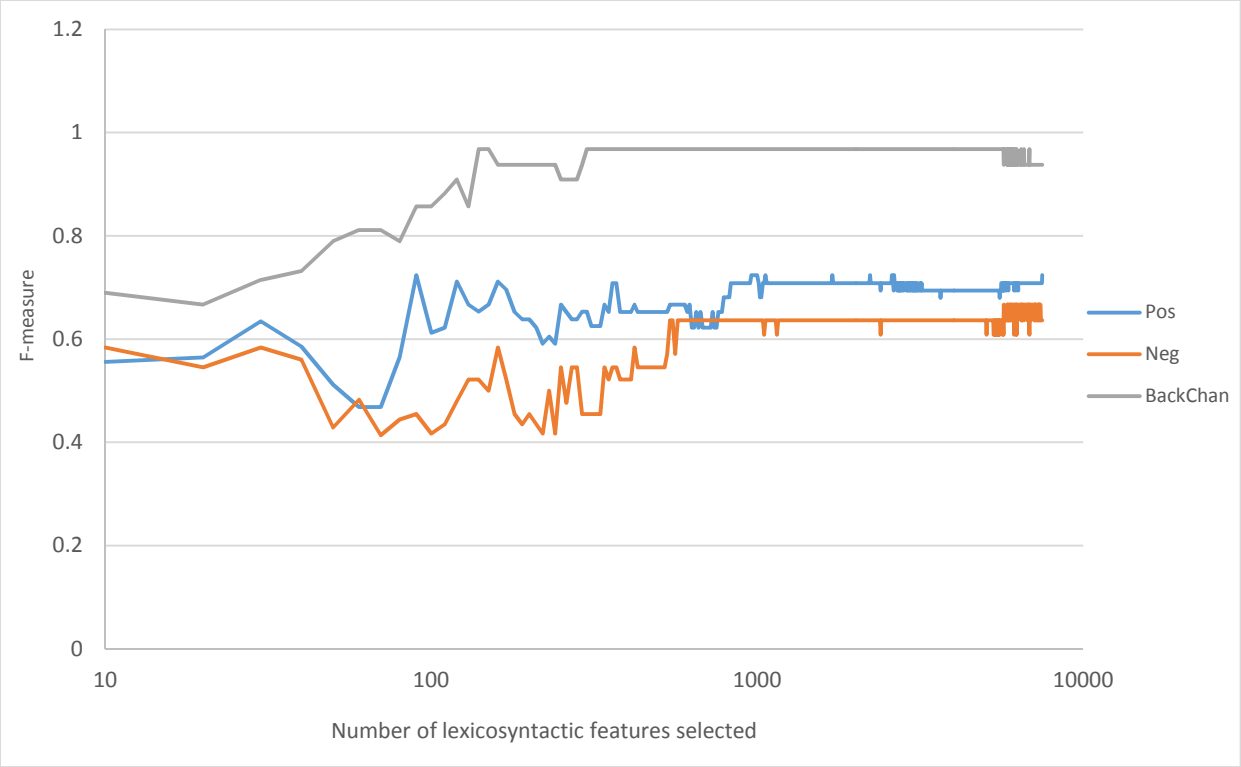



*Chart 9: F-measure per class when only lexicosyntactic features are used*

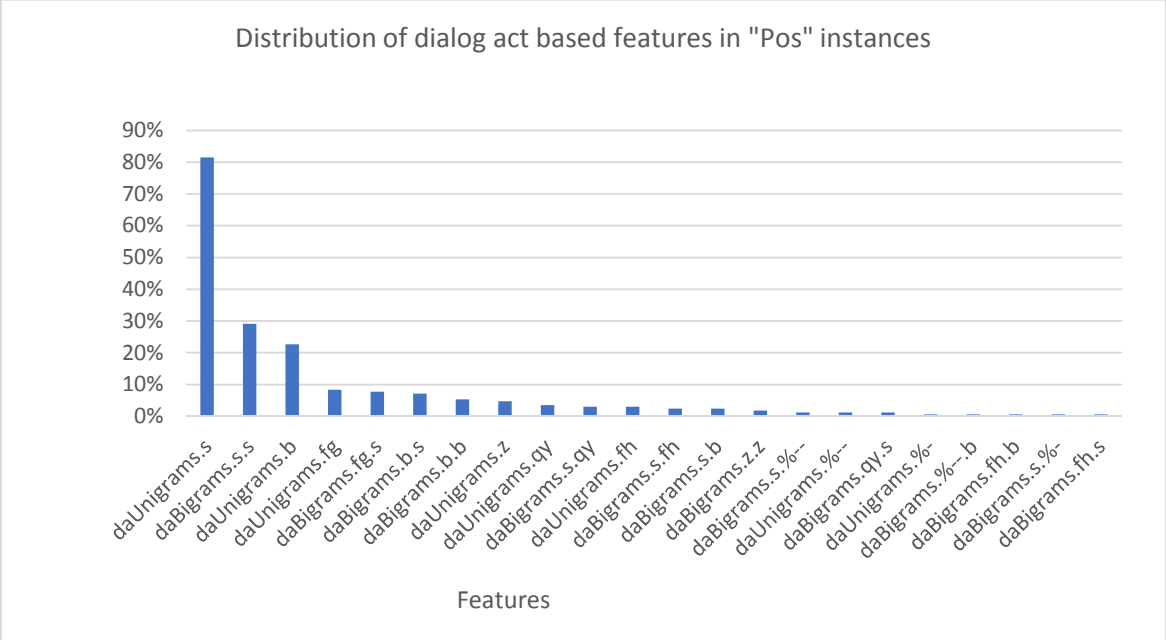



*Chart 10 (a)*

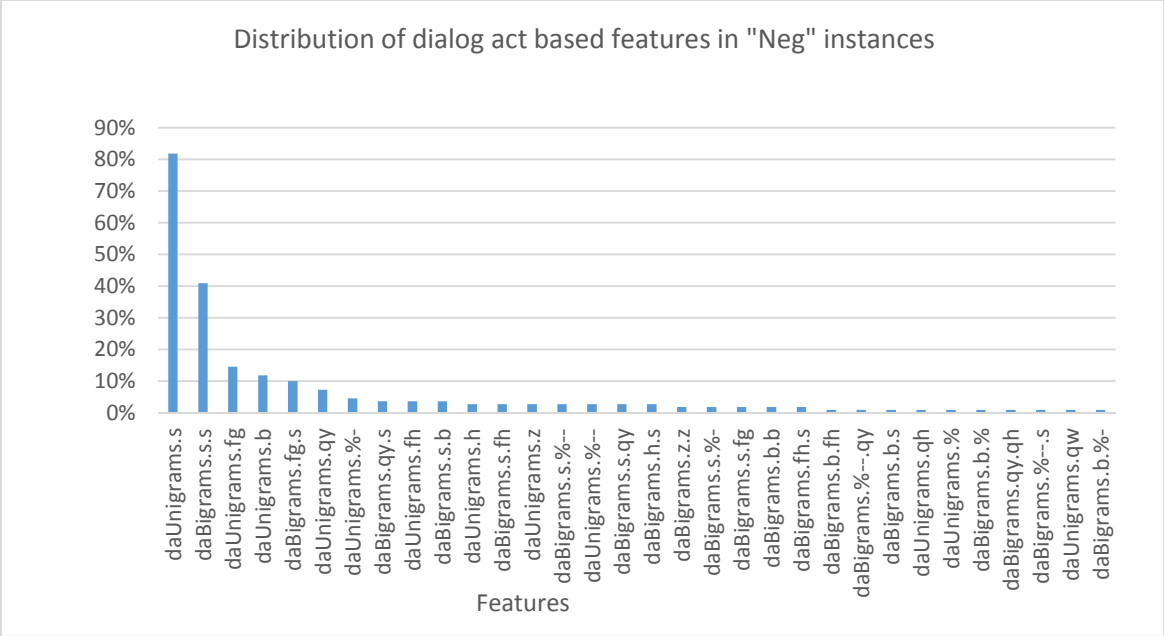



*Chart 10 (b)*

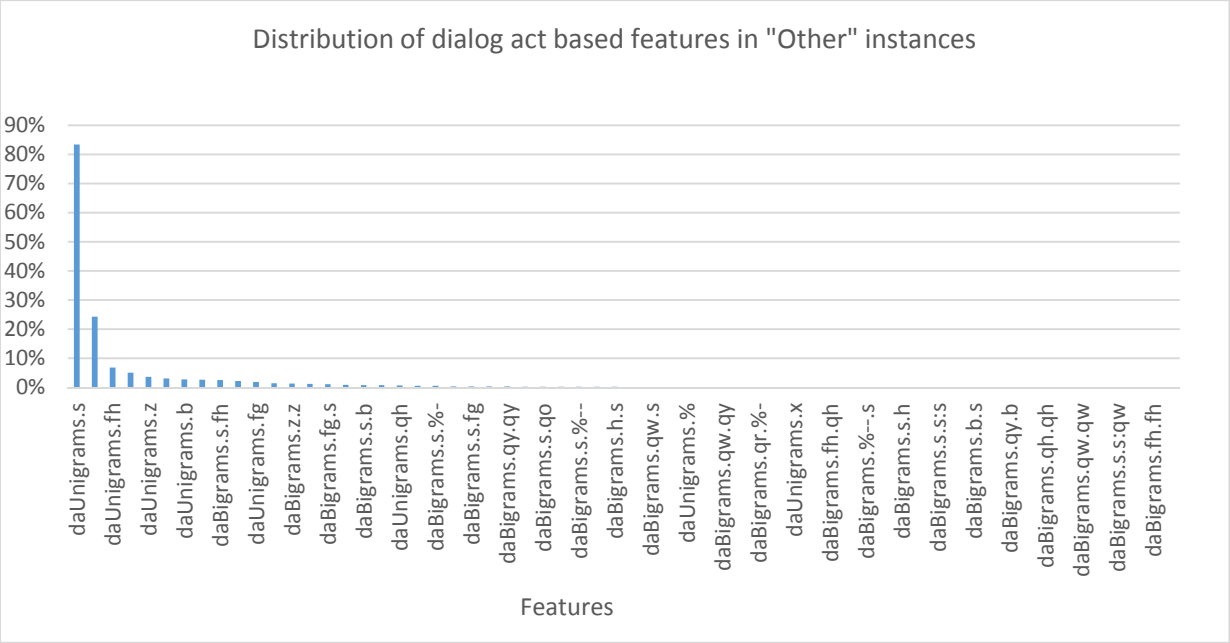



Chart 10 (c)

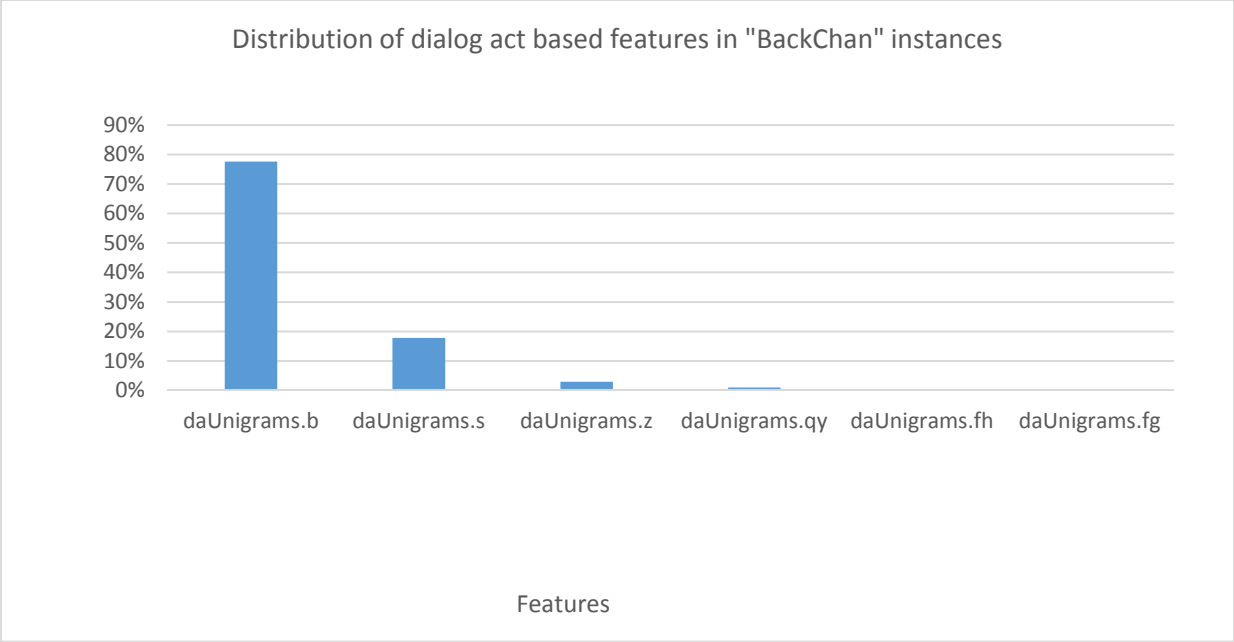



Chart 10 (d)

*Chart 10: Percentage of spurts per class that contain each one of the dialog act based features. The features have been sorted based on the frequency. The horizontal axis shows the dialog act based features which are unigrams and bigrams of the general part of the dialog act parts found in the MRDA corpus. The feature daUnigrams.s refers to the presence of the general part "s" while the feature daBigrams.s.fh refers to two consecutive general parts (the first one is "s" and the second one is "fh")* (*Appendix contains an explanation of what the feature symbols mean)*

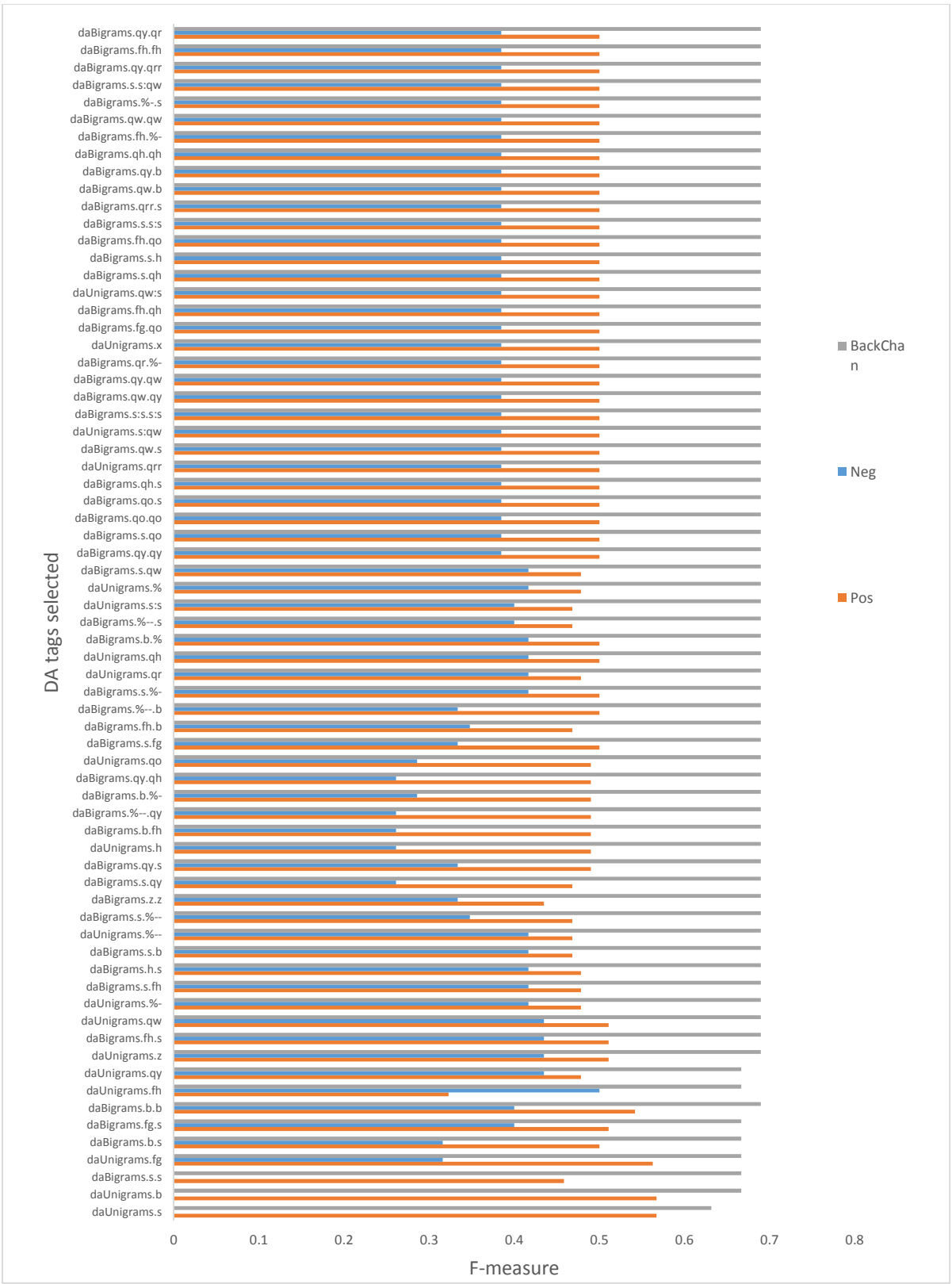
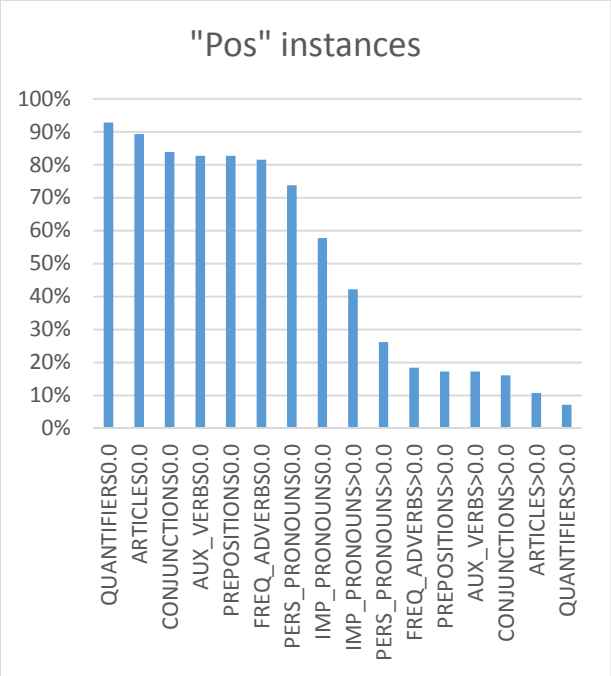
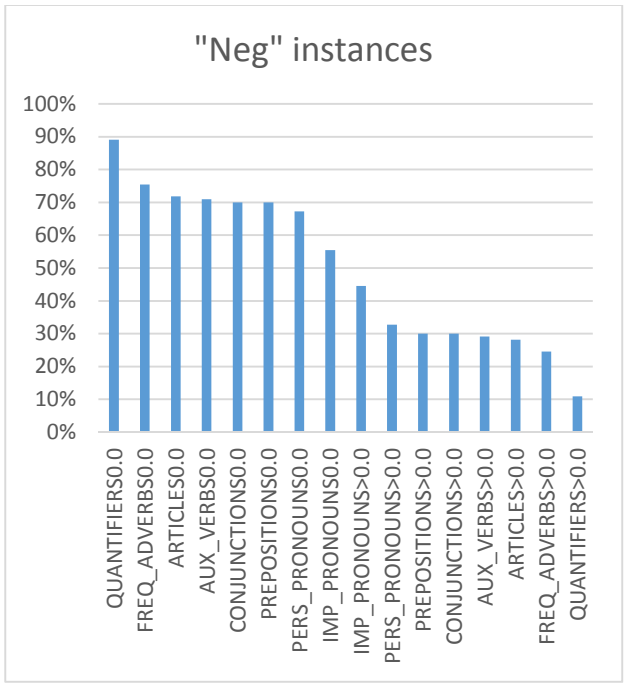


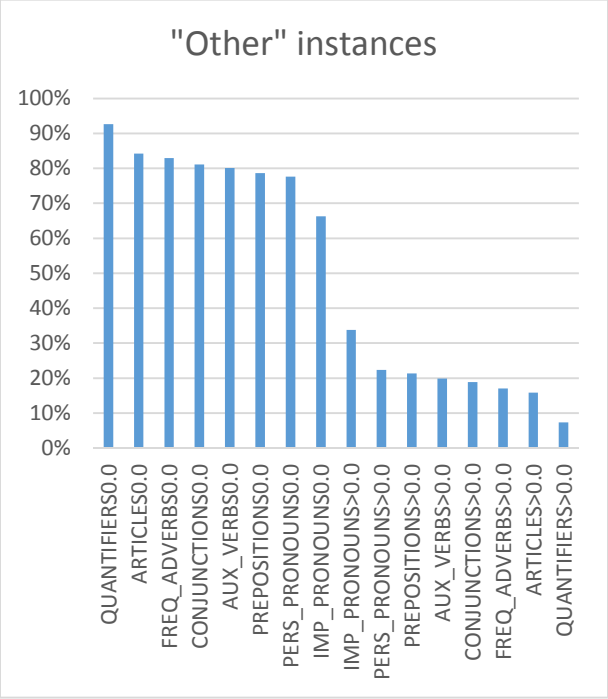

*Chart 11: F-measure per class when only dialog act-based features are used. Each position on the vertical axis implies that for this specific run, the selected features are the ones preceding them on the axis and additionally the label found in this position*


"Pos" instances
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
QUANTIFIERS0.0
ARTICLES0.0
CONJUNCTIONS0.0
AUX_VERBS0.0
PREPOSITIONS0.0
FREQ_ADVERBS0.0
PERS_PRONOUNS0.0
IMP_PRONOUNS0.0
IMP_PRONOUNS>0.0
PERS_PRONOUNS>0.0
FREQ_ADVERBS>0.0
PREPOSITIONS>0.0
AUX_VERBS>0.0
CONJUNCTIONS>0.0
ARTICLES>0.0
QUANTIFIERS>0.0

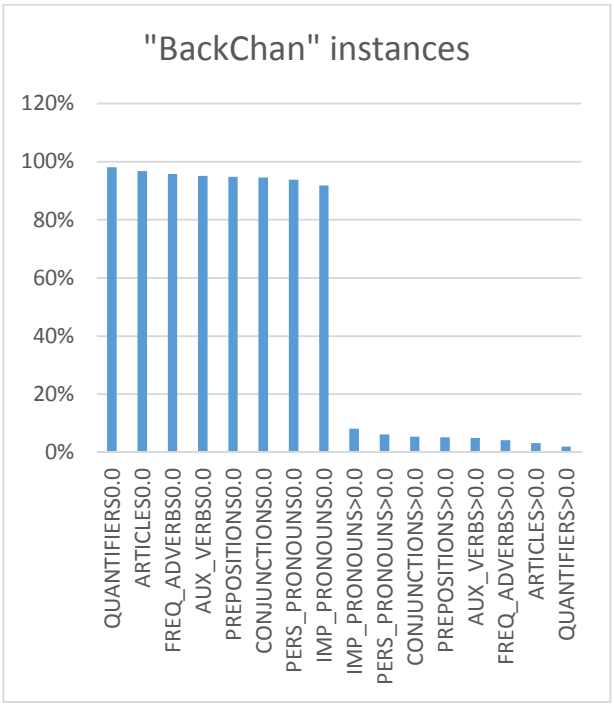
(a)


"Neg" instances
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
QUANTIFIERS0.0
FREQ_ADVERBS0.0
ARTICLES0.0
AUX_VERBS0.0
CONJUNCTIONS0.0
PREPOSITIONS0.0
PERS_PRONOUNS0.0
IMP_PRONOUNS0.0
IMP_PRONOUNS>0.0
PERS_PRONOUNS>0.0
PREPOSITIONS>0.0
CONJUNCTIONS>0.0
AUX_VERBS>0.0
ARTICLES>0.0
FREQ_ADVERBS>0.0
QUANTIFIERS>0.0

(b)


"Other" instances
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
QUANTIFIERS0.0
ARTICLES0.0
FREQ_ADVERBS0.0
CONJUNCTIONS0.0
AUX_VERBS0.0
PREPOSITIONS0.0
PERS_PRONOUNS0.0
IMP_PRONOUNS0.0
IMP_PRONOUNS>0.0
PERS_PRONOUNS>0.0
PREPOSITIONS>0.0
AUX_VERBS>0.0
CONJUNCTIONS>0.0
FREQ_ADVERBS>0.0
ARTICLES>0.0
QUANTIFIERS>0.0

(c)


"BackChan" instances
120%
100%
80%
60%
40%
20%
0%
QUANTIFIERS0.0
ARTICLES0.0
FREQ_ADVERBS0.0
AUX_VERBS0.0
PREPOSITIONS0.0
CONJUNCTIONS0.0
PERS_PRONOUNS0.0
IMP_PRONOUNS0.0
IMP_PRONOUNS>0.0
PERS_PRONOUNS>0.0
CONJUNCTIONS>0.0
PREPOSITIONS>0.0
AUX_VERBS>0.0
FREQ_ADVERBS>0.0
ARTICLES>0.0
QUANTIFIERS>0.0

(d)

*Chart 12: Percentage of spurts per class that have each one of the coordination based properties shown in the horizontal axis. The properties have been sorted based on the frequency value of each one of them. The properties in the horizontal axis should be interpreted as in the following examples,e.g. a spurt that has the property QUANTIFIERS0.0 is a turn that contains words of the LIWC category QUANTIFIERS without seeing words of the same category in the previous turn. Thus the linguistic style coordination score for this category is 0.0. However, a property of type QUANTIFIERS>0 means that the score of coordination is greater than 0.0.*

## 9.5. Equations

$$S_C = \begin{cases} 0 & , if\ current\ turn\ does\ not\ contain\ words\ of\ category\ C \\ & or \\ & if\ it\ contains\ words\ of\ category\ C \\ & but\ the\ preceding\ turn\ doesn\ ot\ contain\ any \\ 1 - \dfrac{number\ of\ turns\ where\ current\ speakers\ uses\ words\ of\ category\ C}{total\ number\ of\ turns\ of\ the\ current\ speaker}, & Otherwise \end{cases}$$

*Equation 1: Coordination score for class C (where C is one of the classes mentioned above e.g. articles, auxiliary verbs, conjunctions etc).*