**TEXT MINING: a Method for Exploring Interactive Alignment in Natural Occurring Conversations – a Case Study of Sentiment Analysis in Shared Reading Groups**

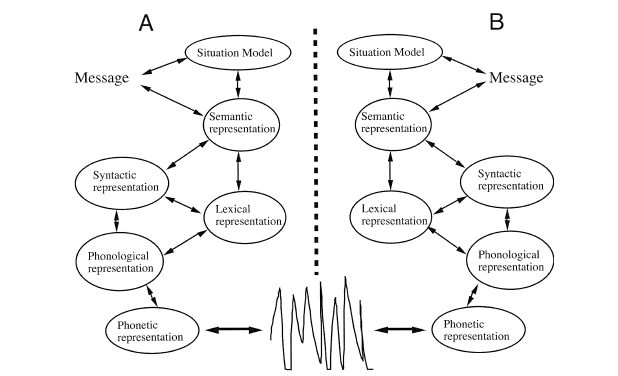
**David Esben Leer Bysted,** *Aarhus University***, Jonas Fenger,** *Aarhus University,*

**Geoff Grevers,** *University of Leeds***, Li Zehui,** *The University of Nottingham Ningbo China***, Julie Mørch Zederkof,** *Aarhus University*

# **1. INTRODUCTION**

Traditional mechanistic accounts of human communication have treated language processing as autonomous and modular. However, investigations behind such models have exclusively been based on studies of monologue and inspiration from telecommunication – not on the basis of dialogue and natural occurring conversations, which are the most basic and natural form of language use (Pickering & Garrod 2004).

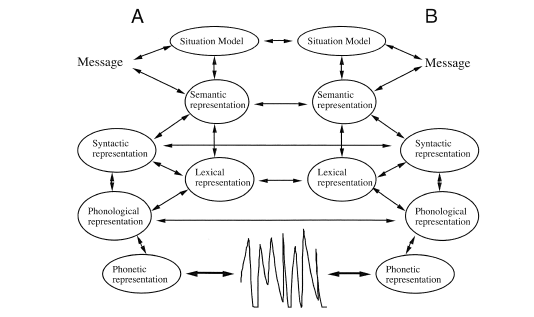
Of the former mechanistic accounts of communication, the Shannon-Weaver model of communication (1949) was the most influential. Their goal was to discover how communication messages could be converted into electronic signals most efficiently, and how those signals could be transmitted with a minimum of error. In studying this, Shannon and Weaver developed a mechanical and mathematical model of communication, known as the *Shannon and Weaver model of communication* (1949).



***Figure 1****: Modification of The Shannon-Weaver model of communication. A and B represent two interlocutors in a dialogue.*

In the above mechanistic model of language processing A and B represent two speakers in a dialogue. Speaker A has a *situation model e.i.* a thing in mind he/she wants to communicate to speaker B*.* Speaker A will individually transform this situation model into meaning (semantics), sentences (syntax), words (lexicon) and sounds (phonology) and transfer all these representations to speaker B in the form of an utterance. Speaker B will now process these sounds, put them into words, sentences, meaning and hopefully into a situation model similar to what speaker A intended. This process is entirely autonomous and modular - the only interactive element is the transition of sound between speakers.

During the last decades methods for analyzing language processing in dialogue have emerged via experimental studies of communication in cooperative problem solving games (Pickering & Garrod 2004). This has revolutionized the general understanding of language models, since these studies suggest that language processing are not autonomous, but instead interactive and mutual. In this framework the core argument is that people prime each other and thereby align language processing interactively at all language levels; phonological, syntactic, lexical, semantic and align situation models.



***Figure 2:*** *The Interactive Alignment model of communication. A and B represent two interlocutors in a dialogue.*

At the highest level, the goal of interlocutors is to align their mental representations, *situation models*. This emerges from alignment at lower levels, both concerned with linguistic decisions (e.g., choice of words) and nonlinguistic processes (e.g., alignment of posture or speech rate). Because of the high-level goal, the interlocutors are particularly concerned with close coupling at these lower levels to reach effective communication.

Interactive alignment has been a hot topic of research throughout the last decades but has only been studied in experimental settings of *cooperative problem solving games* – not in natural occurring conversations. By analysing the *frequency of unique word tokens* and *sentiment* (Pang & Lee 2008)in reading group conversations, we want to explore alignment of lexicon and situation models. From this we expect to prove *text mining* as a valid and groundbreaking method for exploring well-established research results of language use in its *natural ecological setting* – everyday conversation.

The exploration of interactive alignment in everyday conversation in reading groups is furthermore interesting since the core goal of such communication is *open-ending discussions* of literature – not a common goal of *task completion*. It is therefore interesting if individuals will still work towards the same communication efficiency in a setting without a core goal for efficiency. We might therefore expect different structures of alignment than documented in existing literature (Fusaroli et al. 2012).

**2. METHODS**

The corpus used in this exploratory study, was borrowed from *The Reader Therapy Group* in Liverpool (Billington 2011). Their project *Get Into Reading* established two weekly reading groups using a method of *Shared Reading*. The groups either consisted of criminals, individuals suffering from dementia, mental illnesses or people with low literacy skills.

The corpus consists of twelve transcribed reading group sessions from two different reading groups, and all conversations were in english. Additional information about the reading groups were not available for our use.

Preliminary preprocessing of the data was done in order to ensure identical formatting in all documents, i.e. elements were divided by new lines and each element was constructed of an identifier (the speaker) and the spoken sentence. Each document was split up into 10 sections, which would allow us to investigate effects of time.We furthermore divided all sections by person id for us to be able to compare alignment between individuals.

Following, we used the tm package for:

* Eliminating white spaces, numbers, and punctuation (*minimizing irrelevant content)*
* Converting text into lowercase *(making sure same words are represented by the same unicode characters)*
* Removing stopwords *(removing words of sparse semantic content, containing minimal relevant value for our analysis)*
* Stemming *(focusing our analysis on semantics - no syntactics)*

## **2.1. ANALYSIS 1: Frequency of unique tokens over time**

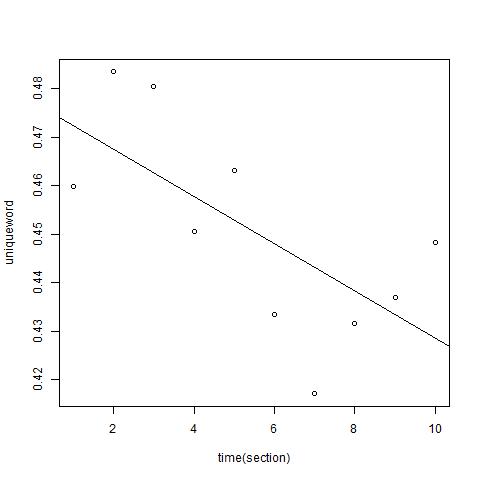
In order to investigate whether the reading group aligned in word-use, we compared unique word token count over time; it provides a method to an insight in the lexical alignment (Fusaroli et al. 2012) of the individuals in each session. A decrease in the frequency of unique tokens over time could indicate that the members of the reading group start to adopt terms used by others, so instead of using different terms for describing similar things, they align their vocabularies and use less unique words. To measure the unique word token frequency over time, we had to provide a method for achieving a reliable time variable. This was done by ‘slicing’ each document into sections and normalizing the data by dividing the total token frequency in each document with the amount of each person's token frequency which allowed us to compare the results on a notionally common scale. We could then compare the unique count of each section with one another, with the prospect that the overall vocabulary would decrease with time/each section. After the data was collected, we could draw a regression line to show the overall decrease.

## **2.2. ANALYSIS 2: Sentiment analysis**

In order to measure whether participants would align on situation models (i.e. sentiment) over time during the sessions, we applied the technique of sentiment analysis (Pang & Lee 2008). Using the AFINN word list (Nielsen 2011) containing 2477 words rated for valence with an integer from -5 to +5, we measured how the language of the individuals changed; whether they had an increased use of positive or negative words. As with the analysis of the lexical alignment, we sliced the documents into equal sized sections to be able to measure the sentiment score over time. Having the hypothesis that individuals might align on situation models throughout the session.

**3. RESULTS**

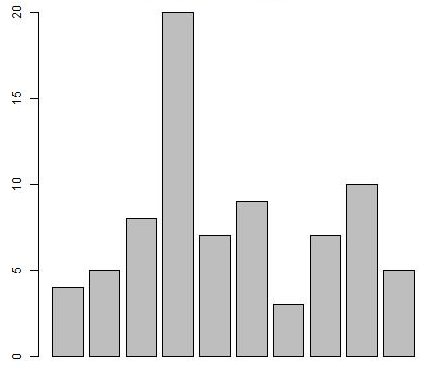
**3.1. FINDINGS: Lexical alignment**

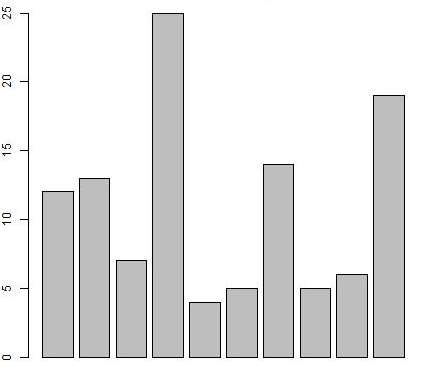
When plotting a regression line (see Figure 3) of the normalized values of unique word tokens according to time, the slope of the regression line indicates that the vocabulary use of the whole group decreases over time, suggesting the reading group participants would settle on a set of terms to use in their discussions i.e. align on lexicon. 

***Figure 3****: A linear regression plot of lexical alignment for one of the reading group sessions. The frequency of unique tokens decreases over time.*

**3.2. FINDINGS: Sentiment alignment**

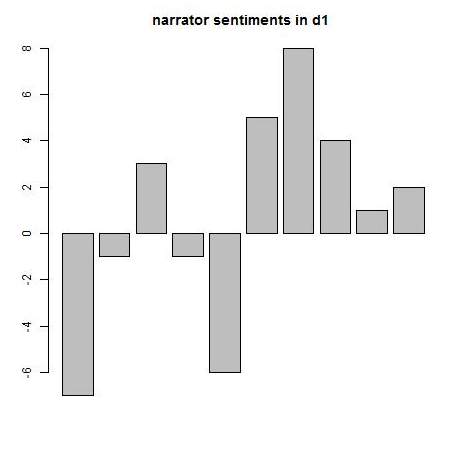
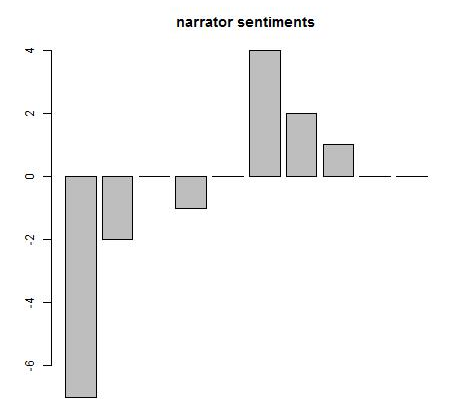
For the sentiment analysis we found two main tendencies. First, our hypothesis that sentiment analysis could be used to explore interactive alignment of situation models between interlocutors was rewarding. Most group members seemed to align throughout most of the reading group sessions. This is indicated by the following plot of sentiment values (see figure 4) for ID 4 and ID 8:





***Figure 4:*** *Sentiments of ID 4 and 8*

Second, groups seemed to align with the sentiment of the instructor. This gives important insights into what an important role the instructor has for the participants well-being and the outcome of shared group reading:



# 

# 

# 

***Figure 5:*** *A bar plot of the sentiment score in a session for the instructor and the rest of the members in Group 2. A similar sentiment score suggests an alignment on the situation model.*

# 

# 

# **REFERENCES**

Billington, J., Dowrick, C., Robinson, J., Hamer, A., Williams, C. (2011). An investigation

into the therapeutic benefits of reading in relation to depression and wellbeing.media/35277/therapeutic\_benefits\_of\_reading\_final\_report\_march\_2011.pdf

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012).

Coming to terms quantifying the benefits of linguistic coordination. *Psychological*

*science*.http://pure.au.dk/portal/files/47690659/Psychological\_Science\_2012\_Fusarol\_0956797612436816.pdf

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in

microblogs. *arXiv preprint arXiv:1103.2903*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends*

*in information retrieval*, *2*(1-2), 1-135.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue.

*Behavioral and brain sciences*, *27*(02), 169-190.

Shannon, C. E., & Weaver, W. (1949). The mathematical theory of information.