# Gossip is more than just story telling
# Topic modeling and quantitative analysis on a spontaneous speech corpus

Boróka Pápay
MTA TK “Lendület” Research Center for Educational and Network Studies (RECENS)
Hungarian Academy of Sciences, Centre for Social Sciences, Budapest, Hungary
papay.boroka@tk.mta.hu

Bálint György Kubik
MTA TK “Lendület” Research Center for Educational and Network Studies (RECENS)
Hungarian Academy of Sciences, Centre for Social Sciences, Budapest, Hungary,
Cleverbridge AG
Cologne, Germany
kubikbalint@gmail.com

Júlia Galántai
MTA TK “Lendület” Research Center for Educational and Network Studies (RECENS)
Hungarian Academy of Sciences, Centre for Social Sciences, Budapest, Hungary
galantai.julia@tk.mta.hu

## Abstract

Gossip is one of the most widespread human activities with multiple functions such as enhancing human cooperation, establishing social order, information sharing, norm enhancing or stress reduction. Gossip has been analyzed mostly by qualitative or survey methods. In this paper, we describe a quantitative approach to identify gossip in a large corpus containing spontaneous talk with LDA topic modeling and quantitative analysis. We aim to identify gossip and its characteristics to analyze its topics, the verbal and non-verbal emotions that were used during gossiping, and other non-textual data such as the number of speakers and the number of persons present during the gossiping events. We also analyze the topics to distinguish gossiping and storytelling by dividing gossip and non-gossip texts in our large spontaneous speech corpora.

Keywords: LDA topic model, sentiment analysis, emotion analysis, spontaneous speech corpus, gossip, reputation, story[1][2]

## 1 Introduction

As two thirds of human conversations are about social topics that can be labeled as gossip, we can state that gossip is the core of social relations and society itself [Dunbar04]. Language caused a significant increase in communication in groups and in information exchange. It also allows us to get information about what happens in a social group, while gossip plays an important role in the sustaining of human cooperation [Dunbar04]. Gossip also often transmits reputational information about individuals, establishing social order, and enhances cooperation [Feinberg14], [Hess06]], [Novak05]. Gossip might also have several purposes for the group in which it occurs, and for individuals who use it.

[1] *Copyright © 2018 for the individual papers by the paper's authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.*

In: A. Jorge, R. Campos, A. Jatowt, S. Nunes (eds.): Proceedings of the Text2StoryIR'18 Workshop, Grenoble, France, 26-March- 2018, published at http://ceur-ws.org

[2] The authors receive funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No 648693).

On a group level, gossip not only enhances group norms, but contributes to interpretation of events, reducing anxiety in stressful situations and managing emotions [Michaelson04], [Mills10].

Gossip is “informal and evaluative talk in an organization, usually among no more than a few individuals, about another member of that organization who is not present” [Kurland11].

In our paper we analyze the distinction between gossiping and storytelling by dividing gossip and non-gossip texts in our large spontaneous speech corpora. To extract the hidden structure of spontaneous speech and to find those thematic topics that people are gossiping about we used topic modeling, specifically Latent Dirichlet Allocation (LDA) which is an unsupervised automated analysis to capture information in large corpora [Blei05]. With topic modeling we can automatically classify and measure issues that occurred during spontaneous speech with those key linguistic and latent semantic features help us to determine the patterns and conversation behaviors during gossiping [Bak14]. When analyzing the hidden structure of spontaneous speech by topic modelling we can correlate the topics with characteristics such as the usage of verbal or nonverbal emotions during gossiping or features such as the number of persons present during conversations.

By using LDA topic modelling in our analysis the gossip topics were clearly separated from all other topics such as storytelling (stories about a third, but external person), cooking, dueling, playing games, etc. We can also assume that the two topics divide also in function and meaning as gossiping usually functions as building one’s own and destroys other individuals’ reputation. Gossip’s so-called storytelling functions serve more to enforce social norms and to maintain social bonding between the sender and the receiver of the story.

The first part gives an insight into our unique dataset. We use our database to test assumptions from the literature about gossip on a spontaneous human speech corpus. The methodology section of the article addresses the tools that were used to test these existing assumptions, such as text preprocessing, topic modelling and quantitative characteristics of the text. In the results chapter, we give qualitative interpretations to the topics in our topic universe as well as attempting to uncover relationships between topic memberships, manually annotated features with strong emphasis on gossip, and prevalent emotions identified using emotion dictionaries.

## 2 Data and database

There has been research on analyzing the presence of gossip from resources like interviews, social media, workplace emails, surveys or anthropological observations [Jones80], [Mitra12] but to extract gossip and to capture its features in everyday human communication from a spontaneous speech corpus is a relatively new approach in gossip research.

For our analysis we used a unique corpus of Hungarian language which consists of approximately 550 hours of spontaneous speech. The documents are transcripts of organic human dialogues, separated by natural silence that are longer than 2 seconds. The high-quality audio recordings were recorded during a Hungarian entertainment programme covering a period of 8 days. We used approximately two thirds of the corpus for our analysis, since the manual transcription of the rest was still ongoing at the time. It is also important to note that the corpus is a work-in-progress. We are in the process of finalizing the transcriptions and conducting steps of quality assurance. The recordings were obtained using personal microphones of eight participants of a gameshow covering the whole interval of their wake times. The contacts of the participants were restricted to a closed environment as they had no or limited possibility to interact with the outside world.

This analysis presents the results of the manual annotation of the HuTongue corpus, which can be a valuable linguistic resource for developing different types of automatic classifiers in the future. Manual annotation provided us the opportunity to tag those parts of the text where the speakers were talking about a person who was a participant or former participant of the gameshow, but was not present. These parts of the text were tagged by the annotators as gossip dialogues. Statements could be formed by the third person’s deeds, personality, and numerous other factors. We also included those texts, where the speaker made a statement about themselves in a relation with the third person. When the speakers were mentioning multiple participants who were not present at the dialogue all mentioned participants were marked individually as a gossip target. Those dialogues were not tagged as gossip where the person whom the speakers were talking about although was not present but was not a participant or former participant of the gameshow (like acquaintances, family relatives, and so on). These discourses were mentioned as storytelling later on in our topic model.

For the analysis of our corpora we used Magyarlanc (translates to “Hungarian chain”) which is a linguistic analyzer tool developed for syntactic analysis of Hungarian language. With the usage of the Magyarlanc toolkit, we were able to conduct POS-tagging of the corpus. It is important to note that this tool can also be used for segmentation, morphological analysis, and dependency parsing of Hungarian texts [Zsibrita13]. These analytic directions are also integral part of planned further analysis of our unique corpus.

To have a deeper insight into gossip's manifestations in spontaneous speech situations, we used annotation marks during the transcription process of our corpora. The annotators used annotation codes to mark the speech about a third person who is not present during the conversation (gossip). During the annotation of gossip the sender and the receiver could be identified by the annotators as well as indicating the names of the participants and the target of gossip. During a dialogue, the annotators could also identify other participants who remained silent during the conversation but their presence could be perceived by the annotator. This tag also provides us the possibility to measure the number of persons present while gossiping. During the transcription process the annotators indicated the exact time interval of speech by using timestamps to sign which participant was talking and for how long. Name tags provides us information about turn-taking and simultaneous speaking situations of the speakers. All tags that we used during the annotation period was thoroughly documented and described for the annotators in a user guide with examples for tag categories. In order to ensure the quality of the corpus, we also used annotation codes that indicates incomprehensible, unidentifiable speech.

For the purpose to detect conversation behaviors when gossiping in everyday informal communication situations we analyzed several types of verbal and non-verbal emotions of the speakers. Annotation marks identified by our annotators to explore the speaker's non-verbal signs of emotions during the conversations were: laughter, crying, sighing, etc.

In our paper we analyzed verbal emotions semantically with emotion and sentiment analysis as well (the dictionaries used were developed by Precognox Company)3. For the emotion analysis we used a six categories dictionary based on Ekman's and Friesen's [Ekman69] theory.

The quality and the compatibility of the corpus is measured in several ways and in multiple dimensions. This was done by automatized means but also with qualitative, random-like monitoring. The work quality of the annotators were measured by giving them the same text files as to compare them by means of matching the transcribed text's accuracy, the annotation tags, name tags, and timestamp usage which are divided into sub-dimensions for more accurate feedback. We compared annotators by comparing their work to each other and by using a reference annotator as well.

These measures are continuously monitored during the corpus construction and documented thoroughly. To measure text similarity we used cosine similarity and Levenshtein distance. In the case of serious quality differences (if an annotator's quality assurance match was under 70%) the text file was re-transcribed and re-annotated. We provided individual feedback to annotators in all of the quality assurance dimension by time-to-time. With these quality assurance tools we were able to monitor 20 hours of the whole corpus.

## 3 Research Directions

In our first research direction we examine what were the participants of the gameshow talking about. We assumed that majority of their speech is about other people. Analyzing spontaneous speech Levin and Arluke [Levin85] concluded that the topic and subject of gossip was mainly concerned about personal habits, manners, appearance, and role performance, and both men and women focused their gossiping conversations mainly on topics such as dating and sex. An important part of our research is that the other people that they talk about can be outside of the closed environment they are in, and can be fellow players. While the reason they talk about other people can be to set examples and norms, gossip about other players can be motivated by reputational motives.

In our next two research directions, we examine the quantitative differences between the dialogue segments assuming that story telling about other people differs from gossiping about fellow players, and also differs from other speeches containing other topics.

Gossip is an evaluative talk and is usually about a third party, who has engaged in a past event and is not present. Gossip is usually among a few individuals [DiFonzo07]. We assume that gossip will be associated with less speakers than non-gossip.

Gossip is also confidential. We assume that during gossip, less people are present in the room, regardless of the number of participants. Gossiping entails confidential topics usually occurring among people close to each other [Shimanoff85]. Close acquaintances or friends presumably speak longer. Also, the level of confidentiality required for gossiping takes longer time to form. We assume that segments that contain gossip are longer than non-gossip segments.

Previous studies on emotions occurring during informal communication showed that emotions are more frequently present than in everyday conversations [Shimanoff85]. When analyzing emotions as conversation behavior in gossiping situations, negative emotions exhibit a tendency to appear more frequently than positive ones.

[3] We would like to thank Precognox for providing us their sentiment and emotion dictionaries

These emotions are expressing most frequently anger, stress or sadness by the speaker. By looking at the subject of gossip and the way of communication mechanisms that speakers use, references show that gossiping while using negative emotions are more likely in an indirect form of speech. [Anderson98]. We assume that during gossip, people express more anger and sadness. We expect more negative emotions in verbal and non-verbal expression of the players.

# 4 Methodology

This chapter gives an overview of the text preprocessing steps and the analytic strategy used. The number of unique terms were reduced using lemmatization, a unique stopword dictionary, and frequency-based filtering. The preprocessed data was then used as an input for topic modeling with Latent Dirichlet Allocation (LDA).

## 4.1 Text preprocessing

The corpus has undergone multiple stages of preprocessing to prepare the transcripts for text analysis. The pipeline described below has been automated to handle the increasingly large corpus.

The analysis of texts written in agglutinative languages like Hungarian requires the lemmatization of the corpus in question due to a potentially large number of words with similar meanings. We have chosen the widely used Magyarlánc software to lemmatize our sizable corpus. This tool enabled us to implement this step in a time-effective and scalable way.

A well-grounded stopword dictionary is key to preparing a corpus for text analysis. The Hungarian stopword list of the Snowball project [Porter01] was chosen as the basis of our extended dictionary. Using the popular Magyarlánc tool for morphological analysis and part-of-speech tagging of Hungarian corpora [Zsibrita13], multiple additions were made to the base list. We excluded all the words from our corpus that were not categorized into a known morphological category by Magyarlánc (represented by "X" in the program's output). Adverbs, apart from verbal adverbs, were also discarded along with adpositions, auxiliary verbs, interjections, particles, determiners, coordinating and subordinating conjunctions. A manual check by researchers ensured that foreign words used as a normal, everyday part of the language and slang words incorrectly categorized by Magyarlánc were not unnecessarily discarded. We also added other nonsensical words to our stop word list identified during the manual checks. In the end, our stopword dictionary contained more than 2000 lemmas.

## 4.2 LDA

As the next step, the document-term matrix of unigram counts was obtained. Even though text preprocessing has left a relatively low number of unique lemmas, terms appearing in less than 5 documents and words present in more than 60% of texts were also removed to discard overly rare and overly frequent unigrams. Digits and punctuation characters were also excluded. The final document-term matrix had 12.961 documents and 8.530 terms.

We have made the decision to use the method of latent dirichlet allocation (LDA) to uncover the underlying topics in our corpus. LDA models the term-topic and topic-document probabilities in a generative way with a Dirichlet distribution as a prior, estimating non-exclusive topic memberships for each document in the corpus. Gensim version 3.2.0, a topic modeling library for Python 3, was used to construct the document-term matrix and for LDA modeling [Rehurek10]. We randomly split our corpus into train, test, and validation set with 50%, 25%, 25% of data, respectively. Our models were configured to use an asymmetric prior learned from the data (alpha parameter set to "auto") and to take 40 passes through the training data. Other parameters were set to their default values. We decided to use 50 topics, a number providing coherent topics and still enabling qualitative assessment by researchers. The average semantic coherence metric as defined by Mimno et al [Mimno05] was -3.31.

During the process of model building and choosing the number of topics, we relied on metrics such as logarithmic perplexity (measured on the test and validation set). Jaccard distances and Kullback-Leibler differences between consecutive training steps, as well as coherence metric by Mimno et al. [Mimno05]. The Gensim Python library provides great functionality to monitor the process of LDA training.

## 5 Results

In this section, we detail our analytic results. After estimating our topic model, we gave a qualitative interpretation to each one of the 50 topics based on the term weights as well as attempting to identify possible connections between different properties of the individual documents and the corresponding topic ratios.

Our results include 50 topics that are present in our speech segments. For our first research direction, we were able to categorize these topics by their main theme. 24 of these topics included speech about everyday life, or so called 'internal issues' like kitchen and food, clothes, body care and so on (Figure 2.1). 10 of the topics seem to be about the entertainment show of which the speakers were part. Discussions about the selection process, duels, and other organized games can be discovered. In other 12 topics, they mostly told stories about other people, who are not part of the participants (Figure 2.2). Two topics were distinctively about each other, called later 'gossip topics', two of the most coherent ones. The other topics were mixed or hard to categorize. Topics were categorized as "gossip topics", if the number of gossip annotation tags provided by human transcribers were significantly high.

For demonstrative purposes, we selected the two topics with gossip (Figure 1), one topic about everyday issues, and one 'story topic' about people from the 'outside' (Figure 2). The most important terms in the two gossip topics are represented by the following word clouds.

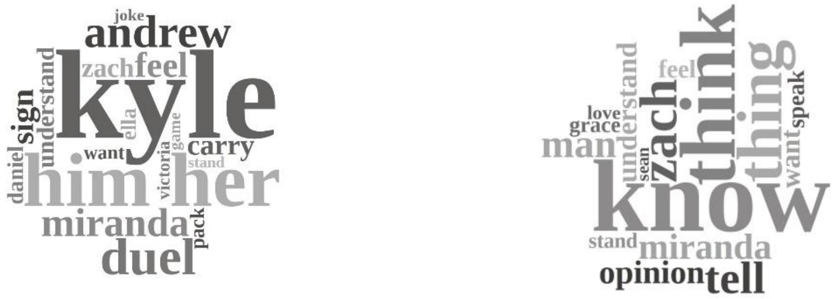

Figure 1: Wordclouds of the most important terms in the two topics with gossip. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

As we can see, topics with gossip contain the names of some of the participants with a high weight. These words most probably describe the actions of these individuals and feelings or actions associated with them.

Some of the most important terms in the two non-gossip topics are the following:

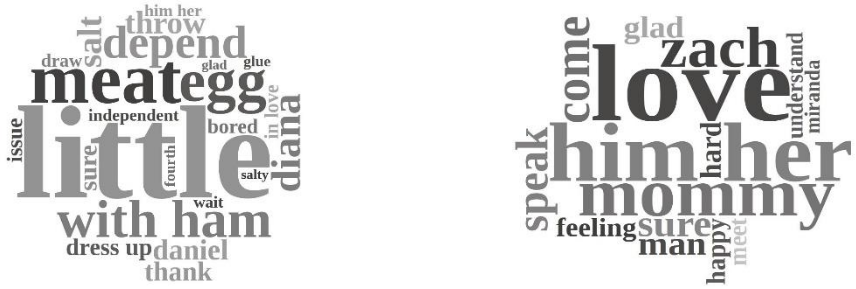

Figure 2: Wordclouds of the most important terms in the two miscellaneous topics. Words were translated from Hungarian to English. Word sizes are proportional to LDA weights. (Source: own visualization)

Words associated with food are well represented in the first topic, while the second is filled with terms referencing outside parties (such as family members or celebrities) as well as their actions and associated feelings, but the second topic was not labeled as a “gossip topic”, since the number of gossip annotation tags were not sufficient.

For our second and third research direction, we wanted to see how segments that contain gossip differ quantitatively from other, non-gossip segments. We analyzed, how the appearance of certain topics correlated with other characteristics and variables from our segments (Table 1).

Topics containing gossip about each other are distinct from other non-gossip topics, including those that are about people from the outside (as family, friends, acquaintances, celebrities).

In contrary to our assumption, texts that contain gossip are not necessarily longer than their non-gossip counterparts. The ratio of gossip statements, coded by annotators is significantly higher in our gossip topics. As we assumed, in one of the gossip topics, less people were present during the conversation, and in both of them the number of individual speakers is significantly lower. Gossip topics usually contain names, personal pronouns, simple verbs as ‘say’, ‘go’ or ‘think’ and verbs related to expressions of own emotions such as ‘feel’ and ‘understand’.

In general, in conversations containing gossip, the participants underused the words from our emotion dictionaries. As expected, they used significantly more anger in both gossip topics, but less joy related words were present in them. Words that express sadness were more likely to be present at the internal issues category then in the gossip categories.

Overall, during gossip conversations, the participants expressed less words that have positive connotation, but more with negative connotation. Interestingly, story topics in general contain less anger, harsher nonverbal emotions and, in some cases, much more positive dictionary. The participants underused the elements of the nonverbal communication during the conversations that contain gossip. They especially underused forms of nonverbal communication as laughter or crying.

## 6 Conclusion

In this paper, we provide some insight into our analysis of a unique, large, and annotated corpus of spontaneous speeches collected during a Hungarian entertainment programme, where participants were placed in a monitored, closed environment for a relatively long period of time. Our analytic strategy was aimed to provide an overview of the dominant topics discussed by the players and to identify relationships between important characteristics (such as topic membership, emotion, or number of participants) and the prevalence of gossip, a widespread activity in human groups associated with many functions.

The manually transcribed and annotated corpus was heavily preprocessed to fit our analytic needs, which entailed, amongst others, the development of a specialized stopword dictionary. The preprocessing step was followed by estimating topic models with the Latent Dirichlet Allocation (LDA) method. The identified topic universe underwent qualitative interpretation by researchers and quantitative relationships were also calculated between features of interest (such as annotation codes indicating gossip, the number of people present, etc.) and emotions identified with emotion dictionaries.

We can conclude that gossiping is different from storytelling and other social topics. Gossip is not only about inform people or to set norms in a community, but it might have a personal impact on the individual with unleashing anger or distress. It is also possible, that it differs from non-gossip because it might be used for reputational purposes. In our analysis it is also a surprise that although we can notice a huge dimension of words of anger while gossiping we cannot notice any non-verbal emotion during this type of communication. In the only category where we can find the usage of non-verbal emotions is the ‘story’ category where participants mainly talk about their outside acquaintances.

Table 1: Example of topics and variables by segments, where segments are coherent units of speech or conversation without silence longer than 2 seconds

| Segment characteristics | internal issues | story | gossip1 | gossip2 |
|---|---|---|---|---|
| **length (in rows)** | -0,074 | -0,054 | -0,018 | -0,076 |
| **gossip ratio** | -0,064 | 0,027 | 0,202 | 0,235 |
| **people present at the conversation** | 0,081 | non sign. | non sign. | -0,041 |
| **people speaking at the conversation** | non sign. | -0,037 | -0,012 | -0,051 |
| **turn taking at conversation** | non sign. | -0,037 | non sign. | -0,047 |
| **ratio of "joyful" words** | -0,016 | -0,036 | -0,058 | -0,091 |
| **ratio of words associated with sadness** | 0,016 | non sign. | non sign. | -0,037 |
| **ratio of words associated with anger** | non sign. | -0,021 | 0,018 | 0,024 |
| **positive_ratio** | -0,023 | -0,045 | -0,067 | -0,092 |
| **negative_ratio** | -0,02 | -0,051 | -0,058 | -0,072 |
| **ratio of non-verbal annotation tags** | -0,026 | 0,014 | -0,046 | -0,105 |
| **the ratio of laughter annotation tags** | -0,042 | non sign. | -0,073 | -0,123 |
| **the ratio of crying annotation tags** | 0,033 | 0,078 | non sign. | -0,048 |

# References

[Anderson98] K. J. Anderson., C. Leaper. Emotion Talk Between Same-and Mixed-Gender Friends. Form and Function. *Journal of Language and Social Psychology. 17(4), pp. 419-448.* 1998.

[Bak14] J. Bak., C.Y. Lin., A. Oh. Self-disclosure topic model for classifying and analyzing Twitter conversations. In Pang, B. W. Daelemans, (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing Doha, Qatar: Association for Computational Linguistics. pp. 1986-1996.* 2014.

[Blei05] D. M. Blei., A. Y. J. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research. pp. 3:993–1022.* 2003.

[Bunt05] G. van de Bunt., R. P. M. Wittek., M. C. de Klepper. The Evolution of Intra-Organizational Trust Networks. The Case of a German Paper Factory: An Empirical Test of Six Trust Mechanisms. *International Sociology, 20(3): pp. 339-369.* 2005.

[Crain12] S. P. Crain., K. Zhou., S. Yang., H. Zha. Dimensionality reduction and topic modeling: from latent semantic indexing to latent dirichlet allocation and beyond. *In: Aggarwal, C.C., Zhai, C. (eds.): Mining Text Data. Springer-Verlag, New York. pp. 129-161.* 2012.

[DiFonzo07] N. DiFonzo., P. Bordia. Rumor, gossip and urban legends. *Diogenes, 54(1), 19-35.* 2007.

[Dunbar04] R. Dunbar. Gossip in evolutionary perspective. *Review of general psychology, 8(2), 100.* 2004.

[Feinberg14] M. Feinberg., R. Willer., M. Schultz. Gossip and ostracism promote cooperation in groups. *Psychological science, 25(3), pp. 656-664.* 2014.

[Hess06] N. H. Hess., E. H. Hagen. Psychological adaptations for assessing gossip veracity. *Review of general psychology, 8(2), 78.* 2006.

[Ekman69] P. Ekman., W.V. Friesen. The repertoire or nonverbal behavior: categories, origins, usage, and coding. *Semiotica, 1, pp. 49–98.* 1969.

[Jones80] D. Jones. Gossip: Notes on women's oral culture. *Women's Studies International Quarterly. Volume 3, Issues 2–3, Pages 193-198.* 1980.

[Kurland00] N. B. Kurland., L. H. Pelled. Passing the word: Toward a model of gossip and power in the workplace. *Academy of Management Review, 25(2), pp. 428-438.* 2000.

[Levin85] J. Levin., A. Arluke. An exploratory analysis of sex differences in gossip. *Sex Roles. 12, (3-4).* 1985.

[Lucas15] C. Lucas., R.A. Nielsen., M.E. Roberts., B. M. Stewart., A. Storer., D. Tingley. Computer-assisted text analysis for comparative politics. *In: Political Analysis 23(2), pp. 254-277.* 2015.

[Mitra12] T. Mitra., E. Gilbert. "Have you heard?: How gossip flows through workplace email", *Proc. AAAI Int. Conf. Weblogs Soc. Media, pp. 242–249.* 2012.

[Michelson04] G. Michelson., M. V. Suchitra. Do loose lips sink ships? The meaning, antecedents and consequences of rumour and gossip in organisations. Corporate Communications: *An International Journal, 9(3), pp. 189-201.* 2004.

[Mills10] C. Mills. Experiencing gossip: The foundations for a theory of embedded organizational gossip. *Group & organization management.* 2010.

[Mimno11] D. Mimno., H.M. Wallach., E. Talley., M. Leenders. A. McCallum. Optimizing semantic coherence in topic models. *In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics. pp. 262-272.* 2011.

[Nowak01] M. A. Nowak., K. Sigmund. Evolution of indirect reciprocity. *Nature, 437(7063), pp. 1291-1298.* 2005.

[Porter10] M. F. Porter. Snowball. A language for stemming algorithms.
*In: http://snowball.tartarus.org/texts/introduction.html.* 2001.

[Rehurek10] R. Rehurek., P. Sojka. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45-50.* 2010.

[Shimanoff85] S. B. Shimanoff. Expressing Emotions in Words: Verbal Patterns of Interaction. *Journal of Communication, 35, pp. 16–31.* 1985.

[Tholander03] M. Tholander. Pupils' gossip as remedial action. *Discourse studies, 5(1), pp. 101-128.* 2003.

[Zsibrita13] J. Zsibrita., V. Vincze., R. Farkas. Magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. *Szeged.* 2013.