# Towards quantifying parliamentary discourse

Longevity of speech patterns in the Czech Chamber of Deputies

Jan Kostkan
au582299@post.au.dk
Aarhus University

## Abstract

How does one measure the influence of a speech in a parliamentary debate? Barron, Huang, Spang, & DeDeo (2018) propose a novel method of quantifying parliamentary discourse during the French Revolution using topic modelling and information theory. In this paper, their method is reviewed and replicated in context of a contemporary institution: the lower house of the Czech Parliament. Furthermore, the analysis is expanded by introducing a considerably longer time frame of measuring a speech's impact. Novel and surprising speeches were found to have more influence over the direction of the debate, when considered in the context of a few hours. However, when impact over multiple weeks is measured, the less risky speeches prevailed. This imposes some limitations on interpretation of results, as proposed in the original paper.

*Keywords*: socio-linguistics, parliamentary discourse, digital humanities

# Introduction

Modelling the evolution of parliamentary debates is an exciting and unexplored way of scaling some qualitative questions to hundreds of thousands of social interactions. The pioneers of the methodology used in this analysis (Barron, Huang, Spang, & DeDeo, 2018) report that debates in the French National Constituent Assembly during the times of French Revolution were accompanied by a constant struggle for direction of the debate between the political left and the conservatives. While speakers on the left kept on bringing new and surprising points, conservatives were trying to keep the discussion in this young parliament within boundaries. In the end, when speeches were modelled in the context of a single day, surprising speeches were found to influence the direction of the discussion more.

Even though executing kings went out of fashion, the conflict between innovation and conservation likely remains alive and well. In this analysis, the method proposed by Barron et al. (2018) is conceptually reviewed and replicated in a new setting: contemporary debates in the lower house of the Czech Parliament: Chamber of Deputies. When relating speeches to others within a single day, the same effect was found: the payoff in influence significantly increases with surprisingness of a speech. However, a novel addition of this paper is modelling speeches in the context of about four months of debate, at which this effect is reversed. Over longer periods of time, it is conservative speeches that survive longer, as the payoff in influence significantly decreases with increasing surprisingness. As this is a replication study, main focus will be placed reviewing the methods and their conceptual backing.

## General background

Parliaments face the task of handling an overwhelming influx of information and ideas (Walgrave & Dejaeghere, 2017). With this overload, it becomes increasingly important how information is accessed and communicated. One way of exercising control over the flow of information is to interact with other representatives in a parliamentary debate and influence what is being discussed. Such power may come with a formal function in the parliament (e.g.

membership in a relevant committee), but it is reasonable to assume that the outcome of some debates depends upon relationships that are not explicitly disclosed to the public (Reh, Héritier, Bressanelli, & Koop, 2013).

Although we cannot model said relationships as such, based on available information, we can model a speaker's influence over the debate, using methods proposed by Barron et al. (2018). Influence of a speech will likely be reflected by it's reception . In this paper, reception is measured by shared and competing usage of language. That is, speech patterns used in an influential speech will be mimicked in future speeches. In the following paragraphs, I will be reviewing some research backing this claim.

The coexistence and competition of speech patterns within a single language, also known as heteroglossia has been the topic of interest since the 19th century (Bailey, 2007). Heteroglossia essentially refers to how we choose different words in different social contexts. Even when referring to the same phenomenon, the choice of wording can impact how a speaker's perspective will be interpreted. For example, although there is a fine line between the words "refugee" and "migrant", they were reported to be associated with different positions in British media; "migrant" being used in more negative contexts (Goodman, Sirriyeh, & McMahon, 2017).

Furthermore one can align, or distance oneself from a perspective by mimicking certain speech patterns. As Tagg (2016) remarks, one's voice co-constructs the voices of others. Research within Communication Accommodation Theory draws a relation between the extent to which speech patterns are mimicked and differences in power between speakers. Speakers with a lower social status have been found to mimic the speech patterns of higher status speakers (Danescu-Niculescu-Mizil, Lee, Pang, & Kleinberg, 2012; Muir, Joinson, Cotterill, & Dewdney, 2016). This conclusion is shared by the first referenced study, which used an ecological dataset and the second one, which is an experimental study.

In this paper, following Barron et al. (2018), this is operationalized by quantifying the composition of speech patterns in a parliamentary speech and measuring composition

differences between them. This process is described in detail in the methods section, but resulting measures are introduced here:

- ➔ *Novelty*, which is how surprising a speech was compared to past speeches. Speeches, in which new speech patterns are introduced will be considered highly novel. Following the theoretical backing of heteroglossia, I am hoping to capture the emergence of new ideas, or new perspectives on existing topics using this measure.
- ➔ *Transience*, which is how surprising a speech was compared to future speeches, or in other words it measures speech pattern longevity. If speech patterns used in speech X are highly mimicked in future speeches, speech X acquires a low transience score.
- ➔ *Resonance*, which is the difference between novelty and transience. This measure is used to explore whose ideas influence the direction of a parliamentary debate. A resonant speech introduces a new perspective (high novelty), that is mimicked in subsequent speeches (low transience). Novelty and transience have to be compared, because transience itself does not uncover who introduced ideas and perspectives that were highly mimicked.

As was hinted, differences between speeches are modelled using a formal Bayesian definition of surprise (Itti & Baldi, 2009). This is an elegant measure of difference, because it allows one to estimate a social context, in which speeches are delivered. In other words, by modelling the present state of heteroglossia, we can quantify how surprising a certain speech was in its context.

Following Itti & Baldi's (2009) definition, the present social context is expressed as a probability distribution capturing the composition of speech patterns used in a certain time frame. This distribution aims to capture the prior expectations representatives have about following speeches, as it is the summary of speech patterns currently in use. These speech patterns carry certain perspectives and ideas that make up a social context of the debate. And what deviates from this context is considered surprising.

To give a more tangible explanation: say there is a debate going on in the Chamber of Deputies about agricultural subsidies. Several past speeches included speech patterns relevant

for this topic, for example “tractor”, “self-sufficiency” and “rye”. However, the next speaker decides it is time to address the refugee crisis and utters different speech patterns, such as “Syria”, “migrants” and “illegal”. As this new composition of speech patterns deviates from the composition established in past speeches, the speech about refugee crisis will be considered highly surprising. Even though this may be an unrealistic example, it is undeniable that in the context of a discussion about agricultural subsidies, a speech about refugees is a novel idea, that could potentially alter the course of further discussion.

Said compositions of speech patterns are expressed as probability distributions, because this allows us to measure surprise on a spectrum and not simply as a binary value (surprising vs. not surprising). The more a probability distribution of context is different from a probability distribution of speech X, the more surprising speech X is. However, there is one more assumption that we are making, when modeling surprise in this way: There is only one context that is assumed to be shared by everyone. I argue that this is a necessary step to take, because the other option would be to model subjective context, hence a context that is different for each representative. This is hardly doable, because we can only guess what the individual contexts are exactly. Such analysis would thus likely require making much more outrageous assumptions.

# Methods

In summary, composition of speech patterns is modelled using Latent Dirichlet allocation (LDA), producing probability distributions for each speech. Measures of surprise are then acquired using Kullback–Leibler divergence (KLD), which is simply the measure of how different two probability distributions are. Both these steps are described in detail later in this section.

Transcripts of debates in the Chamber of Deputies were acquired from Common Czech and Slovak Digital Parliamentary Library (http://public.psp.cz/eknih/). Used transcripts are available in Czech and are authentic, i.e. colloquial language is not corrected to formal Czech.

Speeches were then tokenized, lemmatized and stop words were removed. Lemmatization serves to standardize words to their dictionary forms, so that lemmas used in different grammatical contexts in the original text could be analyzed together. This is especially important in Czech, when compared to English, because of a somewhat more complex morphology. Grammatical cases such as declension or grammatical gender, which only have a limited, or no presence in English considerably increase the number of features a topic model has to evaluate. Furthermore, it is assumed that a noun addresses the same topic, regardless of declension.

For both these tasks, UDPipe parser was used, coupled with a pre-trained model (czech-pdt-ud-2.3-181115.udpipe). The model is based on manually annotated word relationships from Prague Dependency Treebank (Straka & Straková, 2017). Software, including pre-trained models is available online (http://ufal.mff.cuni.cz/udpipe). Said pre-trained model has a reported word tokenization accuracy of 99.9% and lemmatization accuracy of 97.8%. However, it notably struggles with lemmatization of surnames, which were removed from the analysis. Czech surnames are often common nouns or adjectives, that could be used in vastly different contexts. On the other hand, uncommon surnames are not frequent enough to be reliably lemmatized. In both these cases, including surnames in the analysis would negatively interfere with the topic modelling.

Dictionary used for stop-word removal was stopwords-cs (https://github.com/stopwords-iso/stopwords-cs), which consists of common high-frequency words that do not bare much meaning for purposes of this analysis. Stop-word removal is a common procedure in topic modelling (Nikolenko, Koltcov, & Koltsova, 2017). Some custom stop-words were added to said dictionary. These include numbers, month names, addressing words (e.g. "sir", "madam"), titles and full names of all speakers. This was found to improve topic coherence.

Latent Dirichlet allocation (LDA) was used to produce a per-document model of topic distributions and a per-topic model of word distributions (Blei, Ng, & Jordan, 2003). Following the approach taken by Barron et al. (2018), one hundred topics were estimated (K = 100). A total of 384,430 speeches, uttered by 681 unique speakers were evaluated. The

number of speeches corresponded to number of documents evaluated by the algorithm. The model was restricted to use a maximum of 131,072 unique words (V parameter), however the real number of unique words in the vocabulary was lower than that.

Furthermore, prior weights (also known as hyperparameters) were set to $\alpha = 0.1$ for the per-document model and $\beta = 0.1$ for the per-topic model. These priors bias the model towards a lower number of topics per document and a lower number of words per topic, respectively. Resulting per-document and per-topic distributions will thus be more different form one another, than if higher prior weight values were used (Lu, Mei, & Zhai, 2011).

Proposed prior weights differ in topic modelling literature, however, above-mentioned values are the most commonly used ones (Nikolenko et al., 2017). Parameter values used in the original study (Barron et al., 2018) were unfortunately not reported. Therefore, in order to make results of this analysis more compatible with existing literature, I use common parameter values.

For the main task of comparing speech patterns across speakers, the per-document model was used. For each document (speech), a probability distribution is produced, consisting of probability of belonging to each of the 100 topics that have been allocated by the model.
Furthermore, per-document distributions were compared based on Kullback–Leibler divergence, which is a measure of how different two probability distributions are.
Barron et al. (2018) introduced three measures of speech pattern longevity that are based on Kullback–Leibler divergence (KLD): Novelty measures how different (surprising) a speech was compared to past speeches, based on speech patterns used. Transience measures this difference compared to future speeches and resonance then is the difference between novelty and transience.

Based on Barron et al. (2018), the formulas used for calculation of novelty, transience and resonance are as following:

$$Novelty_w(j) = \frac{1}{w} \sum_{d=1}^{w} KLD\left(s(j) \mid s(j-d)\right) \quad (1)$$

$$Transience_w(j) = \frac{1}{w} \sum_{d=1}^{w} KLD\left(s(j) \mid s(j+d)\right) \quad (2)$$

$$Resonance_w = Novelty_w(j) - Transience_w(j) \quad (3)$$

Where j is the rank of a particular speech and s is the function generating a probability distribution of speech pattern composition. Time scale parameter (w) is the number of past and future speeches that speech j is compared to. In this paper, the analysis was conducted at two different time scales (w). First, short-term novelty-transience-resonance scores were calculated using w = 27; A given speech was thus compared to 27 past speeches and to 27 future speeches in this condition. This is a time scale, at which the largest slope of the novelty-resonance line was reported in Barron et al. (2018).

Next, long-term scores were calculated using w = 5000, which was rounded for sake of simplicity and corresponds to roughly the length of two parliamentary sessions. Used corpus consists of a total of 145 parliamentary sessions, held over 9 years. Sessions usually consist of several meetings on different days, which means that 5000 speeches add up to about 2 months of discussion.

## Linear model

To test the relationships between acquired measure, two multilevel linear models were fitted. The first one used resonance as the outcome variable in the short-term condition and the second model fitted the same relationship, but over the long-term. Measures were z-scaled prior to modelling and random intercepts of party and speaker were included.

# Results

## Short timescale

Relationship between novelty and transience in the short-term (w = 27) follows a pattern similar to that which Barron et al. (2018) reported. As can be seen in figure 1. on the left, there is a considerable amount of speeches whose novelty score roughly corresponds to their transience score. However, there is a considerable spread in both directions, meaning that there are many speeches that scored higher on one metric than on the other. Not all highly novel speeches are also highly transient.

The relationship captured in figure 1. on the right is perhaps more interesting. Here we see that novelty is to some extent proportional to resonance, as represented by the regression line. In other words, if surprising speech patterns are used, they tend to stick around for longer, at least on a short-term scale. Nevertheless, there are also some very novel speeches, that did not yield a high resonance, as demonstrated by a large spread of values under the regression line.

A multilevel linear model was used to test the relationship between resonance and novelty in the short-term. The relationship was significant ($\beta \approx 0.36$, $se \approx 0.001$, $p < 0.001$).

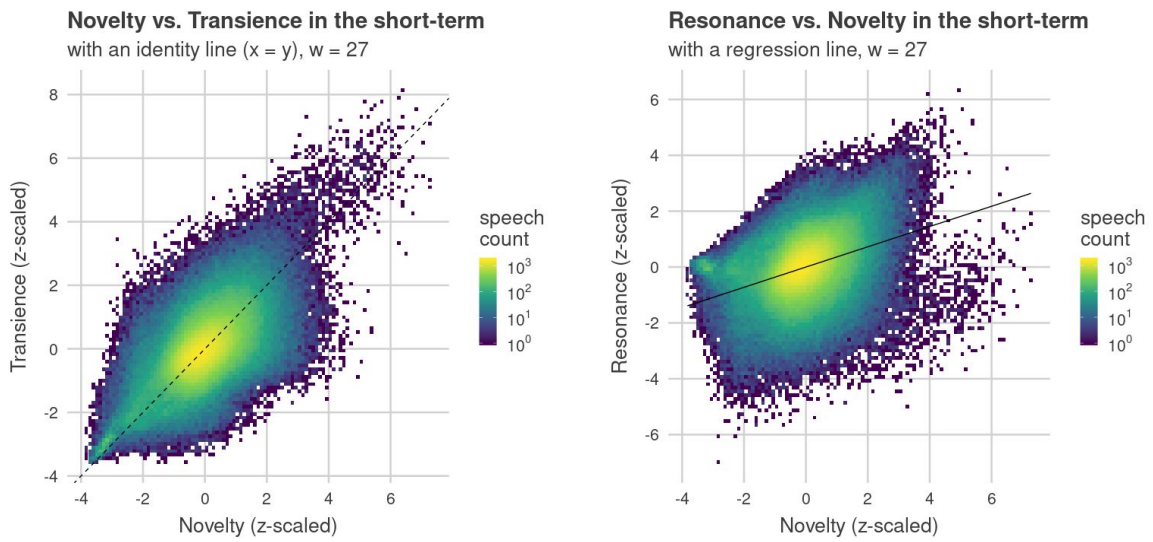


Figure 1.

Short term relationships between measures.
Left: a density plot showing relationship between novelty and transience, plotted with identity line
Right: resonance ~ novelty, including a fitted regression line

## Long timescale

Over the long-term (w = 5000) the relationship between novelty and resonance surprisingly shifts in the other direction: with increasing novelty, resonance drops. The most resonant speeches are ones with a low (around -1), or medium novelty (around 3). However, the overall relationship between these measures is not a clear linear one. We see a huge spread from the regression line and there seems to be a cluster of very highly novel speeches, that at the same time very non-resonant.

Furthermore, at w = 5000 the relationship between novelty and transience becomes even more pronounced. Speeches are densely clustered around the identity line, signifying, that on a long-term scale, transience is highly correlated with novelty.

A multilevel linear model was used to test the relationship between resonance and novelty in the long-term. This relationship was also found significant ($\beta \approx -0.70$, $se \approx 0.001$, $p < 0.001$).

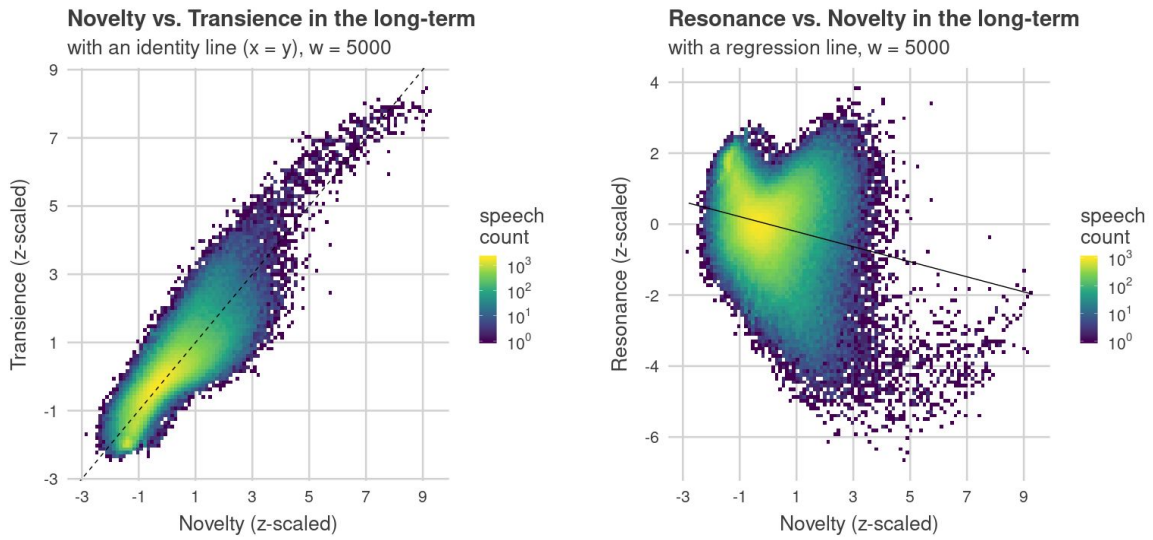


Figure 2.

Long term relationships between measures.
Left: a density plot showing relationship between novelty and transience, plotted with identity line
Right: resonance ~ novelty, including a fitted regression line

# Discussion

In this analysis, speeches presented in Czech Chamber of Deputies between 2010 and 2019 were quantified using topic modelling, and their development across time was modelled using KLD, as a measure of how surprising they were.

First of all, the increasing payoff in resonance with increasing novelty that was found on a short term scale (w = 27), corresponds to one that Barron et al. (2018) reported at the same scale. Same tendency also reported at w = 7 in the original analysis. It seems that the method at hand produces consistent results, even if used to model discourse in a considerably different parliament. The authors interpret this as a manifestation of innovation bias: surprising, or in other way innovative speech is mimicked more than less surprising ones. One explanation of this effect they offer is an overall enthusiasm for everything new and innovative during the early years of the revolution. This may be the case, but it does not explain why does the same trend arise in an established parliament (Barany, 1998).

A more general explanation would be the nature of surprise itself. A stimulus that is surprising, in the formal sense drafted earlier, was reported to attract more sensory attention (Itti & Baldi, 2009). Surprising language patterns were also reported to be recalled better, than ones that do not violate expectations (Hirshman, Whelley, & Palij, 1989). It may be that that when reacting to previous speakers, surprising speeches gain priority. It should also be mentioned that gaining the parliament's attention does not mean gaining its support. This is a distinction that the current method was not designed to dissociate. A highly resonant speech may as well be a highly controversial one. It is likely that representatives opposing a prior speech will use some of the same speech patterns and thus increase the resonance score of the speech they are reacting to. In summary, when interpreting resonance scores produced by current method, it is important to remember that it models the power to influence the direction of a discussion, not a power to determine a discussion's outcome in terms of legislative decisions.

The second finding of this analysis was that the relationship between resonance and novelty is reversed, when modelled in the long-term. That is, resonance payoff decreases with increasing novelty in a longer timeframe. This finding is very surprising, because these results are opposite to those reported in the original study (Barron et al., 2018).

I propose three interpretations of this results, which are largely intertwined. Negative relationship between resonance and novelty at a long-term scale could be:

1. a feature of the method
2. a feature of the institution
3. a feature of speech patterns in general

(1) If a speech is highly novel at a long-term scale, this means that speech patterns used in it were different from those used in 5000 past speeches. That is, speeches scoring high on either novelty or transience are ought to be eccentric, when considered in context of an everyday parliamentary debate. This is not necessarily a methodological problem, as the bar of what is considered highly surprising simply gets higher on longer timescales. Furthermore, there is an advantage of using such a long timeframe: thanks to its sheer extent, a summary of 5000 speeches is perhaps more representative of what a usual day in the parliament is like. It is hard to imagine, that representatives enter each new debate being oblivious to the long-term context of parliamentary discourse.

(2) This pattern may have been produced, because of an institutional punishment of eccentric speeches. A debate in the Czech Chamber of Deputies is a formal interaction, which is explicitly standardized by rules of procedure (Bruteig, 2010). Speeches are often rather long and follow a vague structure, starting with an addressing. Furthermore, MPs are expected to react to other speeches only in their turns, although occasional applause, or comments from the audience are present. This would position the interpretation of this analysis in opposition to what Barron et al. (2018) conclude; Innovation bias turns into a conservation bias. Most highly novel speeches would thus not only be insufficiently impactful to be mimicked, but also systematically neglected.

(3) This effect may also be a natural tendency of speech patterns in any setting, if considered on a long-term scale. If this were the case, it seems that highly novel utterances have their fifteen minutes of fame, after which the discourse stabilizes at an equilibrium. There may be a high risk associated with highly novel speech patterns, that is, only few of them survive over longer timeframes. Although not directly comparable, a high risk of innovation was reported in context of research publications (Foster, Rzhetsky, & Evans, 2015). However, this point is hard to support, because the method used in this analysis is novel itself and has not been applied in other settings yet. We are yet to learn about the relationship between resonance and novelty outside of parliamentary debates and in informal speech.

I find a combination of these three explanations to be the most coherent explanation, the influence of individual factors outlined in previous paragraphs needs to be further investigated. I propose that what we model as high novelty speeches in the long-term analysis are in fact highly eccentric speeches, which do not survive for long in most settings. This is only supported by an institutionalized setting of the Chamber of Deputies that is perhaps even more skeptical of eccentric speeches, than some other settings.

No direct replications of the method used by Barron et al. (2018) have been published to date. The results of this replication point to a few things we should investigate before we fully embrace the method. Namely, the fact that the relationship between resonance and novelty in other settings is unknown. If innovation bias would also be found in non-institutionalized setting, perhaps we would be more careful when drawing conclusions about the mechanics of influence in the French National Constituent Assembly. Furthermore, the effect heavily depends on which time frame was examined. We need to know more about how said relationship differs across timeframes. Barron et al. (2018) do report that novelty significantly predicts resonance up to
w = 100 and mention that on longer time scales, this relationship fades away; It is however not further discussed why that is, or how long it takes for the relationship to turn upside down, as was found in this paper.

## Conclusion

There are high hopes in the field of digital humanities. Although methods are new and untested, they are able to capture some underlying tendencies not visible to the naked eye. In this paper, I have replicated the effect reported in the original study (Barron et al., 2018); That is, on a scale of about half a day of parliamentary discussion, novel and surprising speakers have more influence over the direction of the debate. However, I have proposed that one should be more careful interpreting such effect: first, the historical context by which this effect is explained in the original paper is not present in the corpus used in this analysis. Second, at a scale of about two months, this relationship is reversed and conservative speeches become more influential. All being said, I would be excited to see more publications employing the current method, as acquiring information about the effect in other social settings should provide a more sound base for the interpretation of future results.

# References

Barany, G. (1998). Political Culture in the Lands of the Former Habsburg Empire: Authoritarian and Parliamentary Traditions. *Austrian History Yearbook*, *29*(1), 195–248.

Barron, A. T., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*, *115*(18), 4607–4612.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bruteig, Y. M. (2010). Czech parliamentary discourse. *Discourse Approaches to Politics, Society and Culture (DAPSAC)*, 265.

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, *80*(5), 875–908.

Goodman, S., Sirriyeh, A., & McMahon, S. (2017). The evolving (re) categorisations of refugees throughout the "refugee/migrant crisis." *Journal of Community & Applied Social Psychology*, *27*(2), 105–114.

Hirshman, E., Whelley, M. M., & Palij, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory and Language*, *28*(5), 594–609.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306.

Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, *14*(2), 178–203.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102.