Argumentation is key:
A keyword-based study of arguments in online discourse

Natalie Dykes[◊], Stefan Evert[◊], Joachim Peters[○], Philipp Heinrich[◊]
[◊]Chair of Computational Corpus Linguistics, University of Erlangen-Nuremberg
[○]Chair of German Linguistics, University of Erlangen-Nuremberg

Introduction

Argument mining – the automatic identification and classification of arguments – has attracted growing interest in natural language processing and computational discourse analysis (Cabrio, Tonelli, & Villata, 2013; Janier & Saint-Dizier, 2018). However, authentic texts are challenging due to non-traditional forms of argumentation and implicitness: while they are often persuasive, they usually do not follow the structure of premise and conclusion from traditional formal logics.
We propose a corpus linguistic approach to the comparative cross-genre study of argumentation.

Data and Method

Our data comes from a German web corpus of 14 million tokens in 9746 texts crawled with BootCat (Baroni & Bernardini, 2004). The corpus covers the discourse on multidrug-resistant organisms (MDRO), clinical hygiene and antibiotics-induced diseases. It has been manually annotated with metadata: for each text, the author and the targeted readership is assigned to an actor group (general public, doctors, hospitals, media…) and its topicality is determined (relating directly to MDRO or to a similar, but broader topic). Extensive manual cleaning ensured that all texts were relevant and resolved encoding errors. The corpus was uploaded to CQPweb (Hardie, 2012), allowing us to define and analyse sub-corpora with specific metadata configurations. We use state-of-the-art tools for tagging and lemmatisation (Proisl & Uhrig, 2016; Schmid, 1994; Schmid, Fitschen, & Heid, 2004).

The sub-corpora selected in the present study are:

1. Mass media articles (1.1k texts; 1.3M tokens)
2. Online sources relating to alternative medicine, promoting methods different from conventional medical practices (432 texts; 926k tokens)
3. National, international and regional institutions disseminating information to the public (417 texts, 575k tokens)

These datasets are compared using two kinds of keyword analysis. Firstly, keywords are generated through pairwise comparisons. A second keyword set is obtained by comparing each sub-corpus to a general reference corpus consisting of 3 years of the widespread newspaper *Frankfurter Allgemeine Zeitung* (341k texts, 177.9M tokens). In both cases, we use a conservative version of the effect-size based log ratio (Hardie, 2014) as a keyness measure, taking the lower end of a Bonferroni-adjusted 99% confidence interval (*LRC* – Evert, Dykes, & Peters, 2018).

Results

In this section, we present results for the comparison between sub-corpora 1 and 2. In order to gain an overview of the differences between the two datasets, their keywords were visualised in a semantic map, as shown in figure 1. Text size represents the overall keyness, i.e. the LRC value of a given item compared to the large newspaper corpus. The colour indicates a word's keyness in the direct comparison between the two sub-corpora: redder shades for keywords that are more strongly associated with alternative medicine texts, and greener shades for keywords more strongly associated with the mass media articles. The overall layout is based on semantic similarity (FastText embeddings and multidimensional scaling).

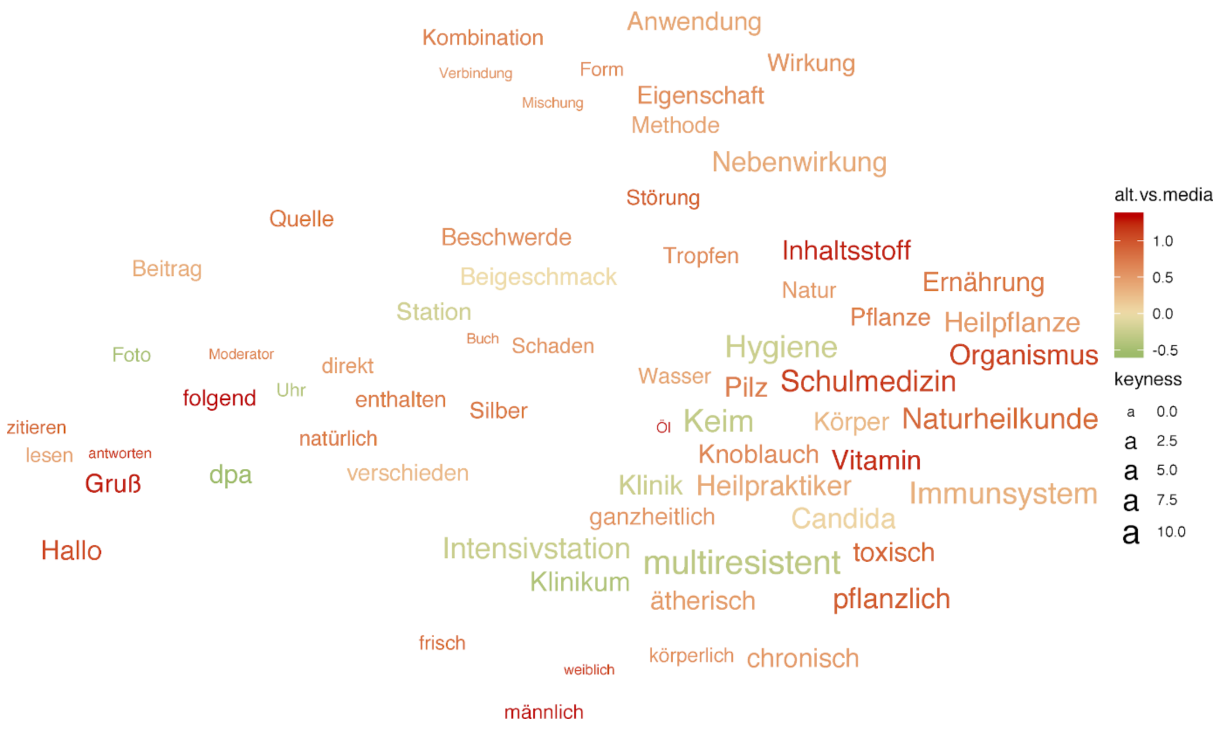

Figure 1: Visual representation of keywords

Keywords clustering towards the left of the plot are mostly false positives with respect to our interest in argumentation patterns – they include greetings (*Hallo* 'hello') and artefacts from boilerplates on the websites (*Beitrag* 'post', *zitieren* 'cite'). The cluster at the top suggests that words relating to the application of medical products are slightly more frequent in alternative medicine, but prevalent in both datasets – as is to be expected from a thematic point of view (*Wirkung* 'effect', *Mischung* 'mixture', *Anwendung* 'application').

Keywords dominant in the media articles primarily relate to hospitals and multidrug resistance; suggesting a focus on the circumstances of contracting clinical infections (*Hygiene* 'hygiene', *Intensivstation* 'intensive care unit').

The visualisation provides a general overview of both sub-corpora. Subsequently, the keywords were annotated for discourse patterns. While we differentiated between four different argumentation patterns and 17 subtypes, the overview focuses on the most frequent argumentation schemes found in the data:

- *argument from effect to cause* (Walton et al., 2008, p. 172):

  [Major premise] Generally, if A occurs, then B will (might) occur.
  [Minor premise] In this case, B did in fact occur.
  [Conclusion] Therefore, in this case, A also presumably occurred.

- *argument from positive consequences* (Walton et al., 2008, p. 101)

  [Major premise] If A is brought about, then good consequences will occur.
  [Conclusion] Therefore, A should be brought about.

Comparison	Actors	Effect – cause	Positive consequences	Other arguments	False positives
Media / alternative	29%	12%	3%	12%	44%
Alternative / media	7%	7%	18%	13%	55%
Media / FAZ	38%	9%	6%	14%	33%
Alternative / FAZ	21%	9%	15%	15%	40%

Table 1: Overview of arguments and actors across keyword comparisons

1) Media articles vs. alternative medicine – comparative keyword analysis

About one third of the mass media keywords refer to discourse actors; the largest group being hospitals and their representatives. While these keywords do not directly reflect an argumentative pattern, they are interesting regarding discursive evaluations, which can be seen as the foundation for argumentation.

Hospitals are presented with a strongly negative bias, notably by referencing vulnerable patient groups such as infants or the elderly. This can be interpreted as the construction of a contrast between hospitals' responsibility to those who need special care and the presence of bacteria to which these persons are particularly susceptible.

Media keywords directly referring to argumentation are mostly variants of the two patterns described above.

In this sub-corpus, the argumentation scheme *from effect to cause* refers to potential reasons for the spread of disease, including agricultural practices, working conditions in hospitals or the economic interests of clinics (*Landwirt* 'farmer', *billig* 'cheap', *Hygienemangel* 'hygiene deficiency').

The *argument from positive consequences* takes the opposite perspective, proposing solutions to the spread of clinical infections (*Screening*, *desinfizieren* 'disinfect').

The keywords from the alternative medicine sub-corpus differ not only on the obvious lexical level, but also suggest very different communication strategies regarding their coverage of discursive categories. Actors prominent in mass media – medical staff, patients or hospital representatives – are not present in the keywords at all. Instead, actor categories refer to spiritual institutions like churches, which are not established participants of the mass media discourse on MDRO. From an argumentative perspective, the following tendencies emerge:

- The *argument from effect to cause* is used in a different context than in mass media. While newspapers primarily list causes like hygienic and economic conditions in hospitals and livestock farming, alternative medicine questions the foundation of traditional medical practice (*Nebenwirkung* 'side effect', *impfen* 'vaccinate', *Schulmedizin* 'traditional medicine'; literally 'school medicine').
- *Arguments from positive consequences* are strongly represented by keywords. Again, their implications differ from those in mass media: positive effects are promoted by substances that are portrayed as either additions or substitutions to traditional medicine, including colloidal silver (*kolloidal, Silber*), mustard oil (*Senföl*) and horseradish (*Meerrettich*).

2) Media articles vs. alternative medicine – newspaper reference corpus

The second comparison evaluates keywords from the two sub-corpora against *Frankfurter Allgemeine Zeitung*. Our results confirm that the choice of reference corpus yields different discursive perspectives on the same datasets (Fischer-Starcke, 2009).

In both analyses of the mass media corpus, the *argument from effect to cause* is similarly frequent. However, the kinds of causes indicated for problems regarding clinical hygiene and infections differ categorically. The calculation against the alternative medicine corpus highlights livestock farming and hospitals' strive for economic efficiency while the comparison to the general newspaper corpus emphasises causes within the hospital, such as inadequate medical treatment or negligence by staff.

The keywords from the alternative medicine corpus show a more varied range of annotation categories than was the case for the comparison with media data on the same topic. In the realm of non-argumentative words, lemmas referring to bacteria and viruses are more likely to be key (18% of keywords vs. 1.4%. These words are included in the actor category because bacteria are often described as intentionally and strategically harming patients, cf. Nerlich & Koteyko, 2009). The *arguments from positive consequences* in the form of naming 'alternative' solutions to traditional medical approaches form the largest argumentative category in both comparisons, while the lexical overlap is small.

In addition, argumentation types which are not present in the comparison between alternative medicine and media articles become visible in the comparison to a more general reference corpus. An example is a type of *argument from effect to cause* criticising the high prescription rate of antibiotics by medical staff – a frequent motive in the mass media articles.

Discussion

Our results suggest that keywords can fruitfully be used to study argumentation in different parts of a thematically constrained corpus. Contrary to fully automated argumentation mining methods, a corpus approach combines the quantitative perspective with a qualitative one, which facilitates the finding of argumentation patterns in spite of explicitness and thematic specificity.

The comparison between keywords obtained through directly contrasting sub-corpora and comparing each against a general reference corpus indicates the importance of perspective. The former calculation yields items highlighting specific differences between the sub-corpora – also leading to a higher false positive rate. The keywords from the respective comparisons against a much larger, thematically more open corpus show a tendency of convergence towards the shared topic.

References

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *International Conference on Language Resources and Evaluation: Vol. 4,* Proceedings *of the IVth International Conference on Language Resources and Evaluation (LREC)* (pp. 1313–1316). Paris: ELRA.

Cabrio, E., Tonelli, S., & Villata, S. (2013). From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems* (pp. 1–17). Berlin: Springer.

Evert, S., Dykes, N., & Peters, J. (2018). A quantitative evaluation of keyword measures for corpus-based discourse analysis*. Corpora and Discourse International Conference*, Lancaster, UK.

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice – A corpus-stylistic analysis. *International Journal of Corpus Linguistics*, *14*, 492–523.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*, 380–409.

Hardie, A. (2014). *A single statistical technique for keywords, lockwords, and* collocations. Internal CASS working paper no. 1, version 1.5, April 2014.

Janier, M., & Saint-Dizier, P. (2018). Evaluating the strength of arguments on the basis of a linguistic analysis: A aynthesis. In *18th workshop on Computational Models of Natural Argument.*

Nerlich, B., & Koteyko, N. (2009). MRSA – Portrait of a superbug: A media drama in three acts. In A. Musolff & J. Zinken (Eds.), *Metaphor and Discourse* (pp. 153–172). London: Palgrave Macmillan.

Proisl, T., & Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop.*

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (NeMLaP), (pp. 44–49).

Schmid, H., Fitschen, A., & Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *International Conference on Language Resources and Evaluation: Vol. 4, Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC).* Paris: ELRA.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press.