Exploring and Visualizing Variation in Language Resources

Peter Fankhauser*, Jörg Knappen†, Elke Teich†

* Institut für Deutsche Sprache (IDS)
R 5, 6-13, 68161 Mannheim, Germany
fankhauser@ids-mannheim.de
† Universität des Saarlandes
Universität Campus A2.2, 66123 Saarbrücken, Germany
j.knappen, e.teich@mx.uni-saarland.de

Abstract

Language resources are often compiled for the purpose of variational analysis, such as studying differences between genres, registers, and disciplines, regional and diachronic variation, influence of gender, cultural context, etc. Often the sheer number of potentially interesting contrastive pairs can get overwhelming due to the combinatorial explosion of possible combinations. In this paper, we present an approach that combines well understood techniques for visualization *heatmaps* and *word clouds* with intuitive paradigms for exploration *drill down* and *side by side comparison* to facilitate the analysis of language variation in such highly combinatorial situations. Heatmaps assist in analyzing the overall pattern of variation in a corpus, and word clouds allow for inspecting variation at the level of words.

Keywords: Language Variation, Corpus Comparison, Visualization

1. Introduction

Language resources are often compiled for the purpose of variational analysis, such as studying differences between genres, registers, and disciplines, regional and diachronic variation, influence of gender, cultural context, etc. Often the sheer number of potentially interesting contrastive pairs can get overwhelming due to the combinatorial explosion of possible combinations. For example, the LOB/Brown family of corpora (Hinrichs et al., 2010) compiled for synchronic and diachronic analysis of British and American English comprises 2 x 2 x 15 subcorpora – 2 for British vs. American English, 2 for the two time slots 60s and 90s, and 15 registers. This in principle allows for 3600 (asymmetric) contrastive pairs. Even when focusing on only one contrastive aspect – region or time – and taking symmetry into account, this leaves 2 x 152 synchronic pairs among registers, such as *A* (Press Reportage) vs. *K* (General Fiction) in British English in the 60s, 2 x 15 diachronic pairs, such as *A* in the 60s vs. *A* in the 90s in British English and another 2 x 15 pairs comparing individual registers between British English and American English, which still adds up to 510 potentially interesting pairs of contrast.

In this paper, we present an approach that combines well understood techniques for visualization *heatmaps* and *word clouds* with intuitive paradigms for exploration *drill down* and *side by side comparison* to facilitate the analysis of language variation in such highly combinatorial situations.

2. Approach

2.1. User Interface

Figure 1 provides an overview of the user interface. At the top, there are three heatmaps. The left heatmap visualizes the overall distance for all 4x4 pairs of British English (60s and 90s) with American English. This heatmap serves for drilling down to particular pairs for closer inspection. The two drill downs 1 (for *GB* 61 vs. *US* 61) and 2 (for *US* 61 vs. *GB* 61) are displayed in the middle and right heatmaps, which visualize the distances between the individual registers ($A - R$) for British English in the 60s vs. American English in the 60s, and vice versa. Distance colors range from greenish to reddish; the color keys to the left and to the right provide more detail.

The left heatmap illustrates that the diachronic difference within a regional variety is generally smaller than the synchronic difference between British and American English, with the largest difference between *US91* and *GB61*. The two detailed heatmaps clearly show the general divide of informational production (*H* and *J*, and to a lesser extent Press: ($A - C$) vs. involved production (the fiction registers *K* through *P*). This general divide holds also for other combinations of region and time, i.e. the overall pattern of the register heatmaps is similar for all 4x4 combinations of region and time.

Each heatmap also weights words for the currently selected pair of subcorpora, visualized by word clouds. The size of a word corresponds to its contribution to the distance, its color corresponds to its relative frequency in the selected (sub)corpus, ranging from blueish to reddish, as the color keys indicate. Both, size and color are scaled logarithmically. Word clouds for the main diagonal show the word weights for the selected subcorpus in comparison to the rest of the corpus (not shown in Figure 1), otherwise they show the word weights for the selected pair of corpora. The word cloud to the left shows words generally typical for British English as opposed to American English in the 60s, the other two word clouds zoom in on this comparison on the specific contrast of *H* (Miscellaneous) for British English (60s) vs. American English (60s), and vice versa. As can be seen words typical for British English comprise spelling variants (*colour, labour, centre, defence, towards*),

Figure 1: Contrast between British and American English, 60s, H (Miscellaneous).

topical words (*britain, british, london*), but also indications for grammatical preferences (*have, been, should, be*). The word clouds for *H* show similar kinds of differences with a notable focus by topical words on civil matters in British English vs. military matters in American English.

The colors panel to the top right in Figure 1 currently allows to choose between two color schemes, *rbow* for visualizing relative difference by a divergent color map, *ryb* as an alternative color scheme for red-green blind people. The p-value panel to the middle right allows to filter words in the word clouds by different levels of significance, by default 0.05 (95 % confidence. Note that "significance" levels 0.25 and 0.5 are highly unusual and practically disregard significance, nevertheless, in some situations they can give interesting insights.

The two drill down word clouds are selected not only on the basis of the selection in the drill down heatmaps, but also on the basis of the selected overall contrastive pair in the left heatmap. This coordinated selection allows for easily exploring a particular register variation between British and American English across time.

For example, Figure 2 gives the typical words for British vs. American English in the 90s – again a mixture of spelling variants, topical words, and grammatical preferences, e.g., the overrepresentation of *the, they, which* in British English compared to American English. Conversely, the corresponding diachronic contrast between the 60s and the 90s within British English in Figure 3 is mainly about topical words, including years around the 60s and the 90s respectively.

Finally, every selection setup gets a unique URL by means of a so called fragment identifier for further reference.

Figure 2: Typical words for British vs. American English, 90s, *H*.

2.2. Corpus Representation and Distance Measure

The individual corpora are represented by means of unigram language models smoothed with Jelinek-Mercer smoothing:

$$p(w) = (1 - \lambda)\hat{p}(w) + \lambda\hat{c}(w)$$

where $\hat{p}(w)$ is the observed probability of the word of the subcorpus (its relative frequency as the maximum likelihood estimate), $\hat{c}(w)$ is the observed probability of the word in the entire corpus, and $\lambda = 0.05$. For a discussion of more smoothing methods for unigram language models see, for example, (Zhai and Lafferty, 2004).

On this basis, the distance between corpora (P and Q) is measured by relative entropy D, also known as Kullback-Leibler Divergence:

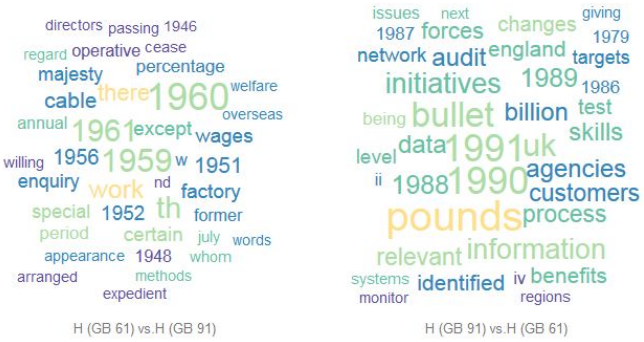

Figure 3: Typical words for 60s vs. 90s, *H*, British English.

$$D(P \parallel Q) = \sum_w p(w) log_2 \frac{p(w)}{q(w)}$$

Here $p(w)$ is the probability of a word w in P, and $q(w)$ is the probability of w in Q. Relative entropy thus measures the average amount of additional bits per word needed to encode words distributed according to P by using an encoding optimized for Q. Note that this measure is asymmetric, i.e., and has its minimum at 0 for $P = Q$ (MacKay 2002).

The individual word weights are calculated by the pointwise Kullback-Leibler Divergence (Tomokiyo and Hurst, 2003).

$$D_w(P \parallel Q) = p(w) log_2 \frac{p(w)}{q(w)}$$

The p-value for assessing the statistical significance of an observed difference in overall word frequencies is calculated based on an unpaired Welch t-test on the observed word probabilities in the *individual* documents of a corpus. This is particular useful when a subcorpus contains only a few documents or a word only occurs in few documents and thus cannot be regarded as representative for a subcorpus.

A more detailed evaluation of these measures and comparison with other measures for comparing corpora, e.g. (Kilgarriff, 2001), is beyond the scope of this paper and will appear in another venue.

2.3. Implementation

The underlying system is implemented in javascript and HTML 5, and thus compatible with most modern web browsers. Word clouds are realized based on Jason Davies' implementation[1], and the heatmaps are realized based on Michael Bostock's library for Data Driven Documents (D3)[2].

The distance matrices, word weights and p-values together with some metadata for tooltips and headings are represented in json (javascript object notation). They are currently precomputed by means of a simple processing pipeline implemented in perl, which requires about half an hour to generate all 3600 contrast pairs for the Brown/LOB family of corpora (each of the 4 corpora has about 1 million tokens) on a moderately equipped laptop.

We plan to make both, the pipeline for computing distance matrices and word weights, and the user interface in javascript available as open source.

3. Related Work

While we are not aware of any visualization of language variation targeting specifically the explorative analysis of variation among many possible pairs of contrast, there do exist a number of approaches with similar goals. Here we can only give an exemplary selection; for a comprehensive overview see, for example, TAPoR 2.0 (Text Analysis Portal for Research)[3].

The MONK workbench (Unsworth and Mueller, 2009) allows to compare pairs of corpora using Dunning's log-likelihood ratio (Dunning, 1993) for word weighting. Apart from the different distance measure (relative entropy as opposed to log-likelihood ratio), the main difference of our approach is that we combine the macro perspective of overall distance with the micro perspective of individual word weights to allow for an explorative analysis of variation. The Voyant Tools (Sinclair et al., 2012) provide a plethora of explorative visualizations for text, including word clouds, co occurrences, and word trends based on frequencies. The focus of these tools, however, lies on summarizing and visualizing one text or corpus, rather than on exploring variation among corpora.

At a very general level, the presented system has taken inspiration from Hans Rosling's forward thinking Gapminder project[4], which showcases explorative visualization of multivariate data in the field of economics. In the longer term, a Gapminder perspective on variation in language resources is certainly worthwhile to explore.

4. Summary and Future Work

In this paper we have introduced an approach for exploring and visualizing variation in language resources which in particular takes into account the combinatorial explosion of contextual dimensions. By combining heatmaps and word clouds it follows the four mantras of scientific data visualization (Shneiderman, 1996; Keim et al., 2006): (i) analyze the data first (ii) show the most important features (iii) zoom, filter, and analyze further (iv) show details on request.

In (Fankhauser et al., 2014) we describe the integration of word clouds with the IMS Open Corpus Workbench[5] (Evert and Hardie, 2011; Hardie, 2012), which generates queries to the REST-API of CQP Web. This allows to easily inspect particular words in their context, and deploy CQP Web's visualization tools for further analysis.

Future work will be devoted to technical as well as methodolical issues: Apart from the integration with a concordance search engine such as CQP Web, we also want to support

[1] http://www.jasondavies.com/wordcloud/about/

[2] http://d3js.org/

[3] http://tapor.ca/

[4] http://www.gapminder.org/

[5] The IMS Open Corpus Workbench. http://cwb.sourceforge.net/

importing external corpora and exporting distance matrices and word weights for analysis with other tools. Moreover, we want to experiment with other visualization mechanisms for the distance matrices, such as dendrograms and scatter plots.

On the methodological side the main challenge lies in supporting a broader variety of feature sets beyond simple unigram language models. This includes latent language models such as topic models (Blei et al., 2003) and hidden markov models (Goldwater and Griffiths, 2007), but also enriched representations such as part-of-speech tagging, and other extensions of unigram models. Such richer feature sets allow to focus the analysis by means of feature selection, but also bear new challenges in measuring and visualizing the contribution of features to a contrast at hand.

5. References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, March.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, UK. University of Birmingham.

Peter Fankhauser, Hannah Kermes, and Elke Teich. 2014. Combining macro- and microanalysis for exploring the construal of scientific disciplinarity. In *Proceedings of Digital Humanities 2014, Lausanne, Switzerland, to appear*.

Sharon Goldwater and Thomas L. Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In John A. Carroll, Antal van den Bosch, and Annie Zaenen, editors, *ACL*. The Association for Computational Linguistics.

Andrew Hardie. 2012. CQPweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3).

Lars Hinrichs, Nicholas Smith, and Birgit Waibel. 2010. Manual of information for the part-of-speech-tagged, post-edited brown corpora. *ICAME Journal*, (34):189–231, April.

Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, and Hartmut Ziegler. 2006. Challenges in visual data analysis. In *In Proceedings of the Tenth International Conference on Information Visualization*, pages 9–16.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA. IEEE Computer Society.

Stéfan Sinclair, Geoffrey Rockwell, and the Voyant Tools Team. 2012. Voyant tools (web application). Technical report, VoyantTools.org.

Takashi Tomokiyo and Matthew Hurst. 2003. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, (MWE '03)*, volume 18, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Unsworth and Martin Mueller. 2009. The MONK project final report. Technical report, The MONK Project, September.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April.