

Katharsis – A Tool for Computational Drametrics

Thomas Schmidt (thomas.schmidt@ur.de), University of Regensburg, Germany

Manuel Burghardt (burghardt@informatik.uni-leipzig.de), University of Leipzig, Germany

Katrin Dennerlein (katrin.dennerlein@uniwuerzburg.de), University of Würzburg, Germany

Christian Wolff (christian.wolff@ur.de), University of Regensburg, Germany

Please cite as:

Schmidt, T., Burghardt, M., Dennerlein, K. & Wolff, C. (2019). Katharsis - A Tool for Computational Drametrics. In: *Book of Abstracts, Digital Humanities Conference 2019 (DH 2019)*. Utrecht, Netherlands.

1. Introduction

With his idea of 'Distant Reading', Moretti (2000) introduced an important *leitmotif* in the Digital Humanities that has led to an ongoing discussion about quantitative methods in literary and cultural studies (Clement et al., 2008; Crane, 2006). We believe that the literary genre of drama is particularly well suited for quantitative analyses and hence adapt the concept of "Drametrics" (as proposed by Romanska, 2015) as a term for the distant reading of dramatic texts. In addition to the actual dialogs, dramatic texts contain other structural elements that can be easily quantified, such as the characters of the play as well as an explicit act and scene structure. Keeping these features in mind, it is hardly surprising that we find a number of recent studies dedicated to the quantitative analysis of drama (e.g. Ilseman, 2013; Wilhelm et al., 2013; Nalisnick and Baird, 2013; Trilcke et al., 2015; Dennerlein, 2015; Xanthos et al., 2016; Willand and Reiter, 2017; Krautter, 2018). At the same time, there have been quantitative approaches to the analysis of drama that date far back into the pre-digital age. As an example for early approaches to quantitative analyses of drama, we would like to refer to the ideas of Marcus' (1973) mathematical poetics, which also contains interesting approaches for quantitative drama analysis.

2. Solomon Marcus' Mathematical Poetics

Marcus suggests the scenic presence of characters as a basic computable measure of a play, which, for each dramatic text, can be visualized by means of a *configuration matrix* (Marcus, 1973). The matrix (cf. figure 1) contains one row for each character of the play, and one column for each scene. Whenever a character appears on stage, the value 1 is entered into the corresponding cell; if a character is not present in a scene 0 is entered as a value.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
B	0	1	0	1	1	0	1	1	1	1	0	1	1	1	1
C	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0
D	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1
E	0	1	0	0	0	0	1	1	0	1	1	0	0	0	1
F	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0
G	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1

Figure 1: An example configuration matrix visualizes the appearances of characters (A-G) throughout the 15 scenes of the play.

Configuration matrices can be used to compute various quantitative aspects of a drama, for instance: the *scenic distance* and *proximity* of characters and even specific relationships between characters (e.g. *dominance*, *alternation*, *independence* or *concomitance*) as well as the overall *configuration density* of plays (Marcus, 1973). The *configuration density* is calculated by dividing the number of cells holding a 1 by the total number of cells. In other words, the configuration density indicates how many of the potential character appearances have actually been realized. It can be understood as a measure of a play's 'population density'. When every character appears on the stage in every scene, the play has a theoretical maximum configuration density value of 1.

During the 1970s and early 1980s, several studies applied Marcus' mathematical approach for the analysis of texts, always dealing with very few samples of text (cf. Marcus, 1974; Marcus, 1977; Marcus, 1984). In these studies, configuration matrices proved to be useful in text analysis, as they fasten and simplify the overview of a character's first or last appearance, co-presence or avoidance with other characters. Some years later, Ilseman (1998) took on the ideas of Solomon Marcus to explore Shakespeare's plays in a quantitative way. Ilseman (1998) used the frequency and lengths of characters' speeches as further parameters and found that the configuration density is an important aspect of genre-distinct quantitative patterns for comedies, romances, tragedies and history plays. In 2005 and 2008, Ilseman used the frequencies and distributions of speech lengths to discuss authorship attribution in Shakespeare's plays.

3. The *Katharsis* Tool

In order to be able to automatically analyze quantitative aspects of dramatic texts according to Marcus' character configurations and Ilsemann's analysis of speech lengths and frequencies, we have created *Katharsis*, a tool for computational *drametrics*. The *Katharsis* tool comprises a *parsing component* that extracts and calculates various quantitative parameters as suggested by Marcus (1973) and an analysis component that searches for dramatic texts of a certain author, genre, timeframe, etc. Currently, a test corpus of approx. 100 German drama texts from the *TextGrid Repository* [1](#) is available for analysis. The texts are available as TEI-XML, allowing for the extraction of metadata (title, author, year etc.) and speeches with the corresponding speaker and structural information. Note that the tool can be extended with further plays from other authors and genres if the texts are encoded in TEI-XML. Furthermore, the quantitative metrics are independent of the language. Figure 2 shows the *Katharsis* results for a search for dramatic texts by Friedrich Schiller. Users can download any quantitative information displayed in the screenshot in JSON format for individual analysis.

Title	Author	Genre	Year	Acts	Scenes	Speakers	Configuration density	Number of speeches	Average length of speeches	Media length of speeches	Longest speech	Shortest speech
Die Jungfrau von Orleans	Schiller, Friedrich	Tragedy	1800	5	59	48	0.07	782	27.89	15	580	1
Maria Stuart	Schiller, Friedrich	Tragedy	1799	5	52	23	0.11	950	28.04	14	282	1
Die Piccolomini	Schiller, Friedrich	Tragedy	1797	5	31	27	0.13	722	25.46	13	235	1
Wilhelm Tell	Schiller, Friedrich	Schauspiel	1802	5	15	79	0.11	1016	23.10	12	611	1
Wallensteins Tod	Schiller, Friedrich	Tragedy	1797	5	63	39	0.08	1240	21.44	9	319	1
Don Carlos, Infant von Spanien	Schiller, Friedrich	Tragedy	1782	5	69	34	0.08	1545	22.81	11	300	1
Die Verschwörung des Fiesco zu Genua	Schiller, Friedrich	Tragedy	1782	5	75	44	0.07	1243	20.51	11	273	1
Kabale und Liebe	Schiller, Friedrich	Tragedy	1782	5	37	15	0.16	920	28.80	15	333	1

Figure 2: Summary of quantitative information calculated by *Katharsis* for dramatic texts by Friedrich Schiller.

With the help of *Katharsis* researchers are able to examine a specific drama in more detail. The tool provides an interactive configuration matrix to explore character appearances and speech statistics for each configuration (figure 3).

Maria Stuart	Act 1								Act 2					
	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6	Scene 7	Scene 8	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
KENNEDY	■	■		■										
PAULET	■	■	■									■		
MARIA		■	■	■										
MORTIMER			■		■	■	■					■	■	■
BURLEIGH							■	■			■	■		
DAVISON									■					
KENT									■					

Speaker: BURLEIGH
 Number of speeches: 20
 Average length of speeches: 29.10
 Median length of speeches: 15.5
 Maximum length of speeches: 161
 Minimum length of speeches: 6

Figure 3: *Katharsis* snippet of the interactive configuration matrix for the play *Maria Stuart*, by Friedrich Schiller.

Furthermore, *Katharsis* produces a table and several interactive bar charts to analyze the distribution of speakers and speech statistics on the structural levels (act and scene) and the progression of these metrics throughout the course of a play (for an example see figure 4).

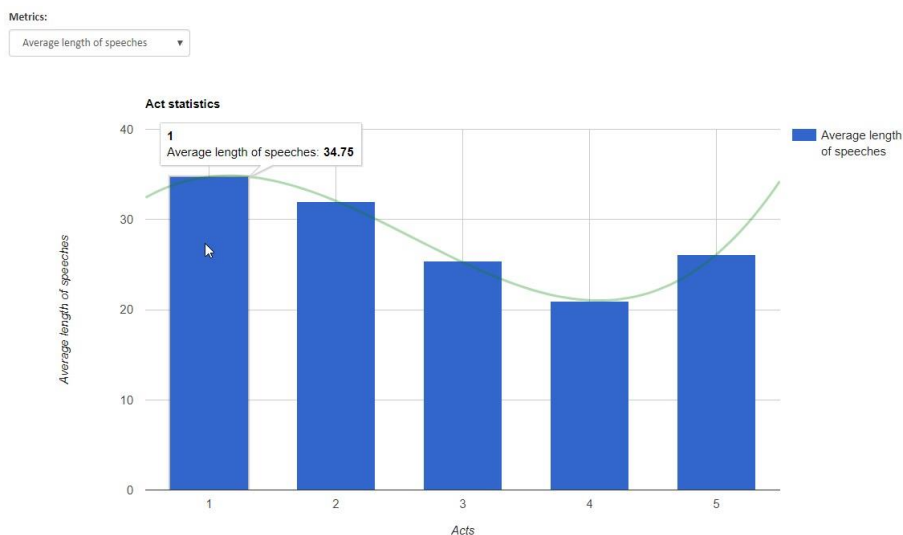


Figure 4: Average length of speeches (measured in number of words) throughout all acts of the play *Maria Stuart* by Friedrich Schiller.

Another segment of the tool shows statistics concerning the comparison of speakers like speech statistics and the distributions of scenic presence. Furthermore, following Marcus' (1973) approach, specific character relations derived from the configuration matrix can be explored. For each character of the play, the tool displays relations to other characters which may be of the type *dominate/dominated*, *alternative*, *independent* or *concomitant*.

The last component concerning the analysis of individual dramatic texts follows Ilsemann's (2005; 2008) idea to examine the distribution of speech lengths in the play. We calculated the speech length by counting the number of words. Users can analyze an interactive histogram and a curve chart. Different speech lengths can be included in the visualization dynamically to narrow down the range of speech lengths for more in-depth analysis (see figure 5 for an example with a comparison).

Finally, *Katharsis* can be used to analyze and compare self-created collections of plays by means of various quantitative aspects. The comparison of different genres and authors is a pre-configured comparison. Figure 5 illustrates a comparison of speech lengths for Goethe and Schiller showing that Goethe's most frequent speech length is seven while Schiller's is rather low with only four words. This might be one reason why the plays of Goethe never were that successful on stage like those of Schiller.

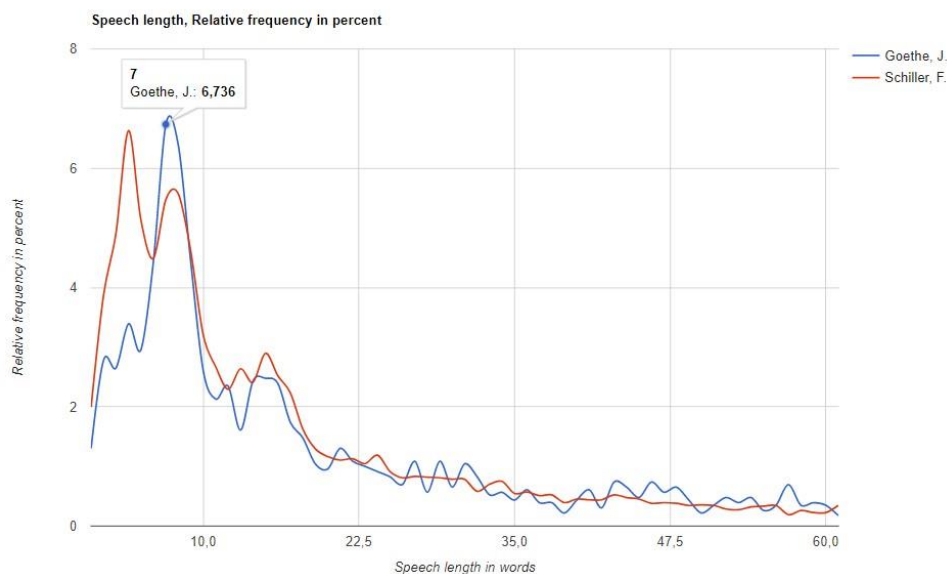


Figure 5: Comparison of the relative distribution of speech lengths for the plays of Goethe and Schiller.

The *Katharsis* tool is available online and can be tested as a live demo in any current web browser:
<http://lauchblatt.github.io/Katharsis/index.html>

4. Case Studies on Quantitative Drama Analysis

In this section, we illustrate the usefulness of *Katharsis* by means of short case studies: An important computable aspect of dramatic texts are the encounters of characters on stage in different configurations. A case study that used *Katharsis* on 13 tragedies, 17 comedies, one tragicomedy and one *Schauspiel* of the German authors Andreas Gryphius, Christian Weise, and Gotthold Ephraim Lessing verified the hypothesis that there is a trend for comedies to have higher configuration densities than tragedies (Dennerlein, 2015). For dramatic German texts from 1600 to 1800 the mean length of speeches in comedies (as compared to tragedies) is lower (see figure 6), whereas the total number of speeches is higher (see figure 7), which means characters in comedies seem to interact in a more dialogic manner.

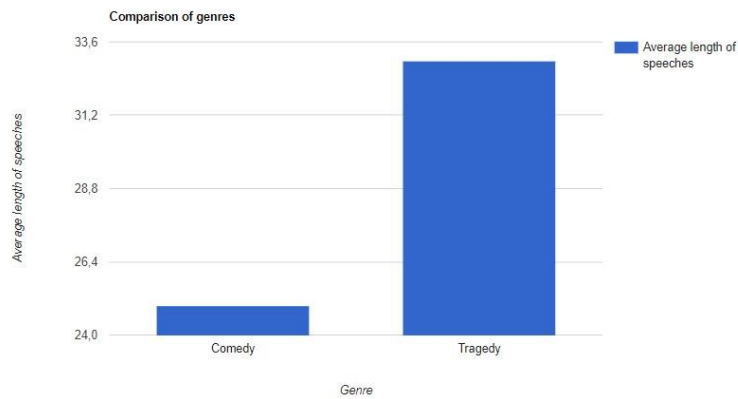


Figure 6: Average length of speeches in comedies and tragedies of the corpus.

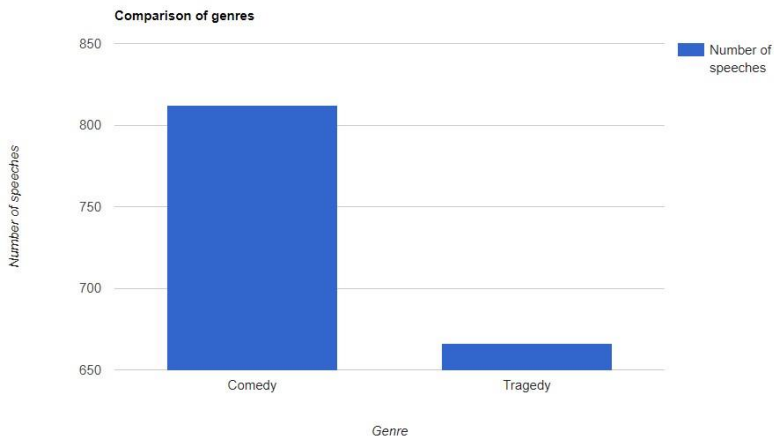


Figure 7: Average number of speeches in comedies and tragedies of the corpus.

This seems plausible with regard to some characteristics of tragedies and comedies already known: Tragedies more often feature monologues because they provide the ideal occasion to reflect on jealousy, hatred, guilt, plans of murder, or suicide. A general lack of communication, or communication difficulties, may be associated with the fact that generally fewer characters share the stage. In comedy, however, protagonists more often encounter each other. Typical comic effects such as confusions between characters or characters exchanging roles as well as speeches delivered at spectators, are staged in the presence of several characters and may result in a rather high configuration density.

5. Future Work: Sentiment Analysis for Drama

To enhance the applicability of *Katharsis* as a tool for computational drametrics, we are currently preparing to include basic sentiment analysis techniques (Liu, 2016) as an addition to mere structural parameters. While sentiment analysis has been particularly popular in the field of computational linguistics, the approach is also gaining popularity in literary studies (Alm and Sproat, 2005; Nalisnick and Baird, 2013; Mohammad, 2011). So far, we have evaluated different sentiment analysis techniques for the context of historic, German language plays (Schmidt and Burghard, 2018a; Schmidt and Burghardt, 2018b; Schmidt, Burghardt and Dennerlein, 2018a; Schmidt, Burghardt and Dennerlein, 2018b). A first *Katharsis* prototype

(Schmidt and Burghardt, 2018b; Schmidt, Burghardt and Dennerlein, 2018b) that implements sentiment analysis for 12 German plays by Gotthold Ephraim Lessing is available online: http://lauchblatt.github.io/Katharsis/sa_selection.html

In the long term, we plan to combine character-to-character sentiment analysis (cf. Nalisnick and Baird, 2013) with the existing configuration matrices, thus not only transferring Marcus' approach of mathematical drama analysis to a digital tool, but rather enhancing it by using additional parameters such as character sentiment.

Appendix A

Bibliography

1. **Alm, C. O. and Sproat, R .** (2005). Emotional sequencing and development in fairy tales. *International Conference on Affective Computing and Intelligent Interaction*. Berlin: Springer, pp. 668-674.
2. **Clement, T., Steger, S., Unsworth, J. and Uszkalo, K.** (2008). *How not to read a million books*. Available online at <http://people.brandeis.edu/~unsworth/hownot2read.html>
3. **Crane, G.** (2006). What do you do with a million books? *D-Lib Magazine* . Available online at <http://www.dlib.org/dlib/march06/crane/03crane.html>
4. **Dennerlein, K.** (2015). Measuring the average population densities of plays. A case study of Andreas Gryphius, Christian Weise and Gotthold Ephraim Lessing. *Semicerchio. Rivista di poesia comparata* LIII: 80-88.
5. **Ilseman, H.** (1998). *Shakespeare disassembled. Eine quantitative Analyse der Dramen Shakespeares*. Frankfurt: Peter Lang Pub Inc.

6. **Ilsemann, H.** (2005). Some statistical observations on speech lengths in Shakespeare's plays. *Shakespeare Jahrbuch*, 141: 158–68.
7. **Ilsemann, H.** (2008). More statistical observations on speech lengths in Shakespeare's plays. *Literary and Linguistic Computing*, 23(4): 397-407.
8. **Ilsemann, H.** (2013). Quantitativ-statistische Dramenanalyse: Welche Aussagekraft haben Häufigkeitsverteilungen der Replikenlängen? *Forum Computerphilologie*. Available online at <http://computerphilologie.digitalhumanities.de/jg09/Ilsemann.html>
9. **Krautter, B.** (2018). Quantitative microanalysis? Different methods of digital drama analysis in comparison. *Book of Abstracts, DH 2018*. Mexico-City, Mexico, pp. 225-228.
10. **Liu, B.** (2016). *Sentiment Analysis. Mining Opinions, Sentiments and Emotions*. New York: Cambridge University Press.
11. **Marcus, S.** (1973). *Mathematische Poetik*. Frankfurt: Athenäum.
12. **Marcus, S.** (ed.) (1974). Poetics and Mathematics. *Special issue of the journal POETICS*, 10.
13. **Marcus, S.** (ed.) (1977). The formal study of drama. *Special issue of the journal POETICS*, 6, 3/4.
14. **Marcus, S.** (ed.) (1984). The formal Study of Drama. II. *Special issue of the journal POETICS*, 13, 1/2.
15. **Mohammad, S.** (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Association for Computational Linguistics*. Portland, Oregon: Association for Computational Linguistics, pp. 105–114.
16. **Moretti, F.** (2000). Conjectures on World Literature. *New Left Review* 1: 54–68.
17. **Nalisnick, E. T. and Baird H. S.** (2013). Character-to-Character Sentiment Analysis in Shakespeare's Plays. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 479–483.
18. **Romanska, M.** (2015). Drametrics: what dramaturgs should learn from mathematicians. In Romanska, M. (ed.), *The Routledge Companion to Dramaturgy*. Routledge, pp. 472-481.
19. **Schmidt, T. and Burghardt, M.** (2018a). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 139-149.
20. **Schmidt, T. and Burghardt, M.** (2018b). Toward a Tool for Sentiment Analysis for German Historic Plays. In Piotrowski, M. (ed.), *COMHUM 2018: Book of Abstracts for the Workshop on Computational Methods in the Humanities 2018*. Lausanne, Switzerland: Laboratoire laussannois d'informatique et statistique textuelle, pp. 46-48.
21. **Schmidt, T., Burghardt, M. and Dennerlein, K.** (2018a). Sentiment Annotation of Historic German Plays: An Empirical Study on Annotation Behavior. Sandra Kübler, Heike Zinsmeister (eds.), *Proceedings of the Workshop on Annotation in Digital Humanities (annDH 2018)* Sofia, Bulgaria, pp. 47-52.
22. **Schmidt, T., Burghardt, M. and Dennerlein, K.** (2018b). „Kann man denn auch nicht lachend sehr ernsthaft sein?“ – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen. In *Book of Abstracts, DHd 2018*, Cologne, Germany.
23. **Trilcke, P., Fischer, F. and Kampkaspar, D.** (2015). Digital Network Analysis of Dramatic Texts. *Book of Abstracts, DH 2015*. Sidney, Australia
24. **Wilhelm, T., Burghardt, M., and Wolff, C.** (2013). "To See or Not to See" - An Interactive Tool for the Visualization and Analysis of Shakespeare Plays. In R. Franken-Wendelstorf, E. Lindinger, and J. Sieck (Eds.), *Kultur und Informatik: Visual Worlds & Interactive Spaces*. Glückstadt: Verlag Werner Hülsbusch, pp. 175–185.
25. **Willand, M. and Reiter, N.** (2017). Geschlecht und Gattung. Digitale Analysen von Kleists ›Familie Schrockenstein‹. In *Kleist-Jahrbuch 2018*. Stuttgart: J.B. Metzler, pp.177-195.
26. **Xanthos, A., Pante, I., Rochat, Y and Grandjean, M.** (2016). Visualising the dynamics of character networks. *Book of Abstracts, DH 2016*. Kraków, Poland, pp. 417-419.

Notes

1.

Available online via <https://textgridrep.org/> (Note: all URLs mentioned in this article were last checked April 29, 2019)