Using Methods of Computational Linguistics for Resolving the "Homeric Question"

Christoph Beierle
christoph.beierle@fernuni-hagen.de
University of Hagen, Germany

Norbert Blößner
n.bloessner@fu-berlin.de
Freie Universität Berlin, Germany

Sebastian Kruse
sebi.sk@gmail.com
University of Hagen, Germany

Abstract

The 'Homeric Question' today is no longer a question about Homer as a person, but a question of the genesis and history of the early Greek epic texts. It is still unresolved. Three mutually incompatible Homer theories (Analysis, Neoanalysis and Oral Poetry Theory) compete, which are based only on manual selections of the text, and this fact seems to be part of the problem. In this paper, we report on the development of a toolbox providing methods of computational linguistics intended to improve our capacity to examine texts in their entireity.

Introduction

The Greek epics, traditionally associated with Homer and Hesiod, play a special role among the early texts of Europe. Not only does European literature begin with these poems, but they also mark the transition from oral tradition to written texts. This intermediate position poses very special problems to every philologist, which are best known under the name 'Homeric Question' (going back to Friedrich August Wolf's Prolegomena ad Homerum, 1795): Have these texts been orally composed and later written down (as Oral Poetry Theory claims)? Or are they poems written by a single great poet experienced in oral tradition (as Neoanalysis maintains)? Or do they combine different passages stemming from different times and poets, compiled later (as Analysis assumes)? The core question is: How can we know?

All three theories present evidence for their findings, but this evidence consists of preselected material – obviously selected according to the principle to present what fits best to the own theory. In order to improve our view, it is useful to look at the complete data instead. That is why as early as the 1970s, the University of Regensburg launched a computer aided project aiming at providing the linguistic data needed for an overview. The project was founded by Ernst Heitsch and Xaver Strasser, and its conception and aims are precisely described in Strasser's dissertation thesis (1984).

Oral texts (as we know from Parry, etc.) are constructed not from single words (Lemmata), but from repeated word connections, which the oralists call 'formulae', but a neutral observer would better call 'iterata' (= lat. 'repetitions'). It has been known that repetitions are a key component of testing Homer theories since the 19th century, but at that time there was no reliable way to collect all of the required data. The Regensburg Project created the first complete directory of Epic repetitions (iterata), based on a lemmatized concordance of all Epic word forms. In the "Regensburger Iteratenverzeichnis" (RIV), which is still unpublished, 'iterata' are defined as semantic and syntactic meaningful phrases that occur at least twice in the corpus. It is e.g. possible a) to find passages that are interlinked by the usage of the same iterata throughout the corpus, and b) to have meaningful information about the usage (e.g. frequency, compactness) of words and phrases.

Therefore, the RIV offers the possibility of collecting and presenting all iterates of a certain type, i.e. of specific frequencies or distributions over the epic texts.

The new possibilities have been used for collecting and researching a complete group of iterata: the so-called 'Singulaere Iterata der Ilias', i.e. those repetitions which remain unparalleled in the Iliad. The results have been published in four dissertations (Ramersdorfer, 1981; Csajkas, 2002; Blößner, 1991; Roth, 1989). This group of repetitions is of special interest, because the *Iliad* is the largest and (according to common opinion) oldest of our epic texts. Therefore, an Oral Theory would expect that it is very small, because why should 'old formulae' be so rare in our oldest and largest text? However, this group contains 3,739 Iterata (out of 18,961 Iterata in sum), and in addition, linguistic and semantic research gives evidence

in many cases that Oral Theory assumptions do not really explain the facts. It looks as if the claims of the Oral Theory are fundamentally based on (wrong) generalizations of some (correct) results. But also the Neoanalyst position is weakened by demonstrations that, in some hundred cases at least, passages of the *Iliad* presuppose the knowledge of 'younger' texts. These results do not only diminish the weight of widely spread theories, but offer concrete data on which new, and more reliable, theories can be built (cf. Blößner, 2006).

Since these examinations, the methods of Computational Linguistics have improved a lot as has the processing power of today's computer hardware. This paper presents an approach to finding further subsets of iterata that continue and extend the idea of improving Epic theories.

Extending the idea of the 'Singulaere Iterata of the Iliad'

The search aims at finding passages in the text which react to each other. With these results, existing theories can be tested and better ones can be built.

Searches of this kind are applications of the scholars' implicit knowledge. So in order to be able to create a computer assisted system this expert knowledge has to be transferred to describable rules and algorithms. The idea of the 'singulaere Iterata', e.g., defined a "conspicuousness" of a phrase that has an unexpected distribution (contrary to the expectation of some theories). This heuristic was proven valid by the results of the four dissertations mentioned above.

Next we will present some ideas that have a well-known foundation within Computational Linguistics but also can be seen as an extension of the 'singulaere Iterata' idea.

The SIoI compares frequencies between two corpora, the Iliad and the complete corpus but the Iliad, where one frequency is very rare (one). So it can be seen as a subset of the **frequency list comparison**, which searches for terms that differ largely between two (or more) corpora. This method uses the frequency class which for a given corpus *K* and a term *t* is defined as

$$HK_t^K = \left\lfloor 0.5 - log_2 \frac{freq_t^K}{freq_{\alpha^K}^K} \right\rfloor$$

where $freq_t^K$ is the frequency of *t* in *K* and α is the most frequent term in *K*. Now one could search for Iterata *I* where

$$HK_I^{notIliad} \gg HK_I^{Iliad}$$

which obviously extends the SIoI.

With the method of the frequency list comparison one could especially search for iterata that are rare in the *Iliad* but frequent outside of it, but it is hard to decide whether they are of significance regarding the search for parallel passages. So we could use the density of the occurrences outside of the *Iliad* as a filter for this class of Iterata. A well-known metric for this question is the **Chi-squared** test which for partitions *R* (e.g. overlapping passages of 300 verses ignoring book boundaries) of a corpus *K* and an iteratum *I* is according to (Rayson et al., 2004) defined as:

$$\chi_I^2 = \sum_R \frac{\left(freq_I^R - exp_I^R\right)^2}{exp_I^R}$$

A high Chi-squared value now suggests that this iteratum is compact in the sense that many of its occurrences are close to each other in respect to the overall corpus.

Another approach to a density filter is the **log-likelihood-ratio** as proposed in (Dunning, 1993) that according to (Moore, 2004) is also valid for rare events (in this context iterata with low frequency). For partitions *R* of a given corpus *K* and an Iteratum *I* it is defined as:

$$\begin{aligned} sig_I^{R,K} &= 2 \cdot \left[\left(freq_I^R \cdot \log \frac{freq_I^R}{exp_I^R} \right) + \left(freq_I^{K/R} \cdot \log \frac{freq_I^{K/R}}{exp_I^{K/R}} \right) \right] \\ exp_I^R &= freq^R \cdot \frac{freq_I^K}{freq^K} \\ exp_I^{K/R} &= freq^{K/R} \cdot \frac{freq_I^K}{freq^K} \end{aligned}$$

Again the expected occurrences of an iteratum in a given passage are compared to the actual frequency. But also the same metric is applied to the rest of the corpus aside from the considered passage. Again a high log-likelihood-ratio value indicates that the iteratum is compact in this passage. Also it is possible to find passages where an iteratum is more frequent than expected.

An issue may arise as this metric is rather complicated from the perspective of a scholar in the humanities and it remains to be well explained; issues like this have been addressed in the "ACID for the Humanities" of the DARIAH project (Büchler, 2013) and also in the conclusions of Bestgen, 2013.

First results and conclusions

The ideas presented above have been fully implemented, and this implementation has been used in first applications. It has been suggested to use the Chi-squared test and log-likelihood-ratio to test the validity of an assumption of the Oral poetry, which says that

the singer could choose freely from a given set of phrases while performing. This should lead to a rather equal distribution of highly frequent terms. Using our implementation, first results raise concerns regarding the validity of this Oral poetry thesis as the analysis shows that there are many high frequency Iterata that are also "compact" in some parts of the works. Further and more philological work has to be done by analyzing those iterata to find out whether there are semantic reasons for this, and to be able to explain the passages that differ strongly. In general, we expect that the further development of our implementation and its application to different theses proposed by the Homer theories will lead to new insights into the problems named the 'Homeric Question'.

Bibliography

Bestgen, Y. (2014): Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing* 29(2), 164–170.

Blößner, N. (1991): *Die singulären Iterata der Ilias. Bücher 16–20*, Stuttgart (Teubner).

Blößner, N. (2006): Relative Chronologie im frühgriechischen Epos. Eine empirische Methode und erste Ergebnisse, in: *Geschichte und Fiktion in der homerischen Odyssee*, hg. v. A. Luther, München (Beck), 19-46.

Büchler, M. (2013): *Informationstechnische Aspekte des Historical Text Re-use*. PhD thesis, Universität Leipzig.

Csajkas, P. (2002): *Die singulären Iterata der Ilias. Bücher 11--15*, München (Saur).

Dunning, T. (1993): Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61–74.

Janko, R. J. (1982): *Homer, Hesiod and the Hymns*, Cambridge (Cambridge University Press).

Moore, R. C. (2004): On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 333–340. ACL.

Pavese, C. O. / Boschetti, F. (2003): *A Complete Formular Analysis of the Homeric Poems*, Vol. I--III, Lexis' Research Tools.

Ramersdorfer, H. (1981): *Singuläre Iterata der Ilias. Alpha–Kappa*, Königstein/Ts. (Hain).

Rayson, P., Berridge, D., and Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004)*, pages 926–936. Presses universitaires de Louvain (PUL).

Roth, P. (1989): *Singuläre Iterata der Ilias: Phi--Omega*, Frankfurt a.M. (Athenäum).

Strasser, F. X. (1984): *Zu den Iterata der frühgriechischen Epik*, Königstein/Ts. (Hain).