Das artifizielle Manuskriptkorpus TASCFE

Armin Hoenen

13. Januar 2015

1 Abstrakt

In diesem Paper soll das *Teheran Artificial Shahname Corpus with Frankfurt Extension (TASCFE)* digitalisierter handschriftlicher Texte zur Evaluation automatisch generierter Stemmata vorgestellt werden. Ein *Stemma codicum* oder kurz Stemma ist eine Visualisierung der genealogischen Zusammenhänge innerhalb eines Manuskriptkorpus oder einfacher gesagt ein Manuskriptstammbaum. Die Generierung solcher Stammbäume verfolgt generell zwei Hauptziele: ein genaueres Verständnis der Überlieferungsgeschichte und die Rekonstruktion eines Urtextes. Bis in die 90er Jahre hinein wurden Stemmata vornehmlich manuell erstellt, sind aber seitdem zunehmend auch automatisch generiert und analysiert worden, siehe u.a. Spencer et al. (2004), Roos and Heikkilä (2009) und Roelli and Bachmann (2010).[1] Technologisch ist die bio-informatische Phylogenie Donordisziplin, wie die Nutzung phylogenetischer Programme und Algorithmen zur automatischen Manuskriptstammbaumerstellung zeigt. Dabei fehlt es in der biologischen Phylogenie an Möglichkeiten, erzeugte Stammbäume zu evaluieren, da die Aufspaltungsvorgänge der Spezies, die durch die Verzweigungen symbolisiert werden nicht beobachtet und aufgezeichnet werden konnten, lagen sie doch z.T. Millionen von Jahren in der Vergangenheit. Im Gegensatz dazu ist es in der Stemmatologie durchaus möglich, sowohl die Vorlage als auch die Kopie im Korpus vorzufinden. Mehr noch, es ist möglich neue Korpora zu erzeugen und gleichzeitig die Kopiergeschichte der Manuskripte aufzuzeichnen. Diese Daten können dann in einem klassisch informatischen Evaluationsszenario der Beurteilung von stem-

[1] Für eine detaillierte Darstellung der historischen Entwicklung der Stemmatologie siehe O'Hara (1996), Robinson and O'Hara (1996), van Reenen et al. (1996) und van Reenen et al. (2004).

Text	Sprache	Anzahl Manuskripte	Anzahl Worte	Publikation
Parzival	Englisch	21	957	Spencer et al. (2004)
Notre Besoin	Französisch	13	1029	Ph.V. Baret (2004)
Heinrichi	Altfinnisch	64	1208	Roos and Heikkilä (2009)
Shahname	Persisch	50	107	Hoenen (hic ipsum)

Abbildung 1: Die artifiziellen Traditionen

magenerierenden Methoden genutzt werden. Artifizielle Korpora wurden bisher drei Mal erzeugt, siehe Ph.V. Baret (2004), Spencer et al. (2004) und Roos and Heikkilä (2009). Nur das letztgenannte Paper evaluierte mehrere stemmagenerierende Algorithmen, darunter auch die händische Rekonstruktion, mittels einer Distanzfunktion zwischen dem echten Stemma und den erzeugten. Diese Distanz nannten die Autoren Average Sign Distance (ASD). Sie misst die Ähnlichkeit der Topologien des korrekten und des erzeugten Stammbaums anhand der Ähnlichkeit der inneren Abstände aller Knotentripel (von vorhandenen Manuskripttexten) im erzeugten mit deren *shortest-path* Abständen im echten Stammbaum. Abbildung 1 fasst Kennwerte der drei artifiziellen Korpora zusammen.

Alle bisher bekannten artifiziellen Traditionen sind im lateinischen Alphabet verfasst. Hier wird das TASCFE Korpus vorgestellt, welches in persischer Sprache (Farsi) im arabischen Alphabet vorliegt. Neben der Sprache besteht seine Besonderheit für eine Bereicherung der Landschaft artifizieller Korpora darin, orale Variation zu approximieren. Orale Variation ist solche Variation, die nicht aufgrund von Fehlern im Kopierprozess, sondern aufgrund der Dynamik mündlicher Überlieferung entstanden ist und die zum Teil stark von erstgenannter Variation abweicht. Die Oral Formulaic Theory (OFT) wurde in den 30er bis 60er Jahren des vorigen Jahrhunderts durch Parry and Parry (1987) und Lord (1960) im Zusammenhang mit der *Homerischen Frage* erarbeitet. Ergebnis dieser Theorie war u.a. die Erkenntnis, dass Texte wie die Odyssee keinen Urtext, d.h. keine Originalversion besitzen. In der Zeit vor Erfindung der Schrift wurden Texte ausschließlich oral tradiert. Dabei war zur konkreten Textmanifestation ein Aufführender und (mindestens ein) Zuhörer notwendig. Da die Umstände jeder Aufführung jedoch unterschiedlich waren, war es so auch der Text selbst. Z.B. nutzte ein Barde bei derselben Geschichte viele Ausschmückungen (z.B. Adjektive), wenn er viel Zeit hatte, erzählte sie jedoch ein anderes Mal, wo die Zeit drängte, ohne Ausschmückungen. Dazu kommen Fehler des menschlichen Erinnerungsapparates, der andersartige Variationen erzeugt, als solche, die beispiels-

weise durch Buchstabenverwechslung beim Abschreiben zu Stande kommen. Zu Beginn der Schrifteinführung wurden Texte via "Pseudo-Aufführungen" vor einem Schreiber (Diktate) erstmals in schriftliche Form überführt. Da derselbe Text mehrfach in solchen Diktaten aufgezeichnet worden sein kann, da weiterhin jede Aufführung ganz wie in der rein oralen Welt dieselbe Geschichte in unterschiedlicher Textform (mehr Ausschmückungen/weniger Ausschmückungen u.a. Arten oraler Varianz) hervorbrachte, können am Beginn mancher (meist der frühesten) Manuskripttraditionen Varianten stehen, die sich nicht mit den Arten an Variation aus rein literarisch überlieferten (d.h. als schriftliche Texte entstandenen) Texten decken. Solch eine Variation ist für die Stemmagenerierung wichtig, da sie determiniert, ob ein einziges oder mehrere Stemmata und ob mehrere oder nur ein Urtext angenommen werden müssen. Das TASCFE Korpus trägt dieser Art der Variation zuminndest teilweise Rechnung, da an seinem Anfang vier verscheidene Versionen stehen. Neben der Sprache ist dies die zweite stemmatologierelevante Besonderheit des TASCFE.

Der Text ist ein Auszug (Strophe) aus dem persischen Nationalepos Shahname (*Buch der Könige*), (l Qasim Ferdoussi, 1967, p.55). Es entstand um das Jahr 1000. Die Autorenschaft wird generell *Abu l-Qasim Ferdoussi* zugerechnet, wobei orale Einflüsse im Werk bereits seit längerem diskutiert werden, siehe u.a. Yamamoto (2003) und Rubanovich (2011) . Die ca. 6.500 Token des Korpus wurden 2014 in Teheran (43 Manuskripte) und in Frankfurt (7 Manuskripte) von Freiwilligen entweder von einer gedruckten oder einer handschriftlichen Vorlage durch Abschreiben produziert. Anschließend wurde das Korpus digitalisiert und aligniert. Ein des Persischen nicht mächtiger Freiwilliger kopierte zusätzlich eines der handgeschriebenen Manuskripte, um der These nachzugehen, dass in historischer Zeit Analphabeten oder Schreiber anderer Schriften Manuskripte kopiert haben könnten, was sich aber deshalb als unwahrscheinlich erwies, da es einer persischen Muttersprachlerin aufgrund der partiellen Unlesbarkeit nicht möglich war von dem so kopierten Manuskript eine weitere Kopie anzufertigen.

Kein einziges Manuskript entsprach genau der Vorlage. Eine qualitative Analyse der Phänomene, die denen historischer Korpora ähnelten (so z.B. Zeilensprünge oder Wortsprünge, aber auch synonymische Ersetzungen) konnte zeigen, dass aufgrund des Schriftsystems und der daraus teils zur lateinischen Schrift unterschiedlichen Fehler eine andere Differenzierung in Fehlerklassen je nach

Schriftsytsem notwendig sein kann. Dies knüpft an Andrews and Macé (2013) an, die zeigen konnten, dass Variationsklassen je nach Sprache variieren können. Das digitale Zeitalter eröffnet dahingehend die Möglichkeiten einer schriftsystemübergreifenden Analyse von durch Abschreibefehler verursachter Variation, die dann zur Abstraktion der dort wirkenden universalen Prinzipien beitragen wird, siehe Abbildung 2. Dies setzt die Schaffung geeigneter Ressourcen voraus.

Auf die Daten wurden im Weiteren stemmatologische Algorithmen angewandt (die dann mittels der oben angesprochenen ASD evaluiert wurden). Hierbei konnte gezeigt werden, dass ein Ansatz zur Feststellung von Oralität durch Gruppenbildung im Stemma besteht, wobei die Levenshtein Distanz, Levenshtein (1965), bei hoher Gewichtung von Lücken im Alignment eine besonders geeignete und gleichzeitig leicht zugängliche algorithmische Basis darstellt, siehe Abbildung 3. Dabei wurde ein wortpaar-basierter Vergleich aller Manuskriptpaare durchgeführt. Die Levenshtein Distanz aller Wortpaare des jeweiligen Manuskriptpaares wurde (auch als Baseline für den Vergleich mit weiteren Algorithmen) zu einer Manuskriptpaargesamtdistanz aufsummiert.

Die Matrix der Manuskriptpaardistanzen wurde mittels des Neighbor Joining Algorithmus, Saitou and Nei (1987), wiederum eine geeignete Baseline, aus dem 'ape' Packet (Paradis et al. (2004), Paradis (2012)) der Programmiersprache R in einen Stammbaum überführt, der dann visualisiert und evaluiert wurde, siehe Abbildung 4.[2]

[2]www.r-project.org/

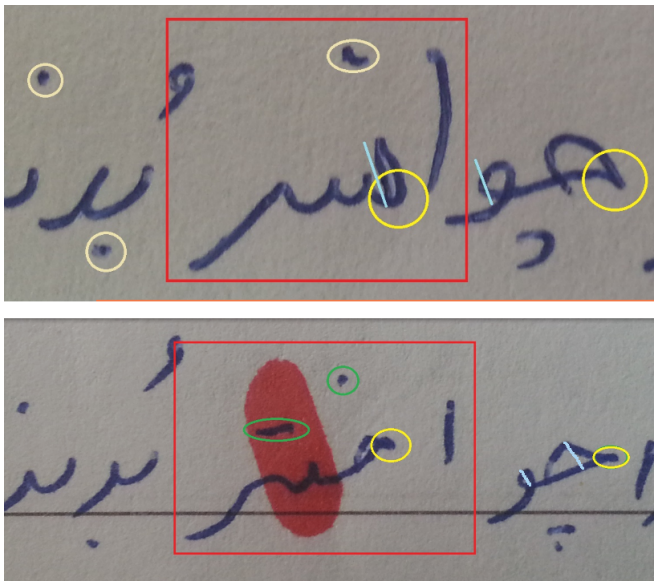

Abbildung 2: Die durch ungewöhnliche Buchstabenform ausgelöste Fehlkopie von افسر (oben) nach اختر (unten) in roten Rechtecken. Die Kennstellen, die den Abschreibefehler ausgelöst haben, sind farblich markiert. Das ف in افسر ist nicht so rund wie und länger als erwartet (blau). Zudem ist der einzelne Punkt auf dem ف versehentlich breiter (hautfarben). Dennoch ist im oberen Rechteck eindeutig nur ein Punktmuster erkennbar, unten jedoch zwei (grün). Des Weiteren hat das ـسـ zwei deutliche Hacken, wobei der mit rotem Stift unterlegte untere entsprechende Buchstabe nur einen aufweist. Außerdem ist das ف in der unteren rechten Ecke rund, was auf das vorausgehende چ jedoch nicht zutrifft (gelb). Obgleich der Abschreibefehler höchstwahrscheinlich durch die ungewöhnliche Form des ف ausgelöst wurde, passte die Ersetzung gut in den Kontext, vielleicht sogar besser als das Original. Genau diese Interaktion von kontextuellem Priming und ungewöhnlichen Buchstabenformen ist ein idealer Kandidat für schriftsystemübergreifende Prozesse beim Abschreiben.

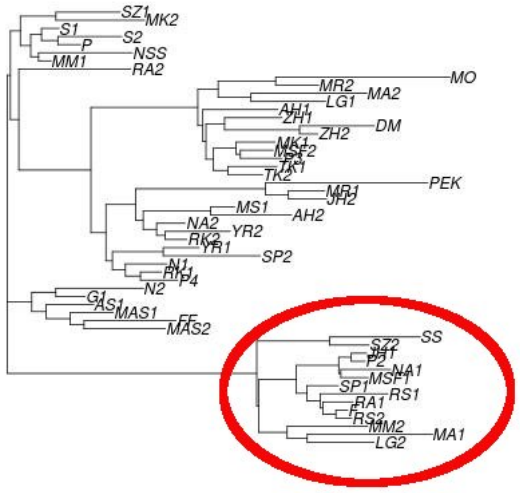

Abbildung 3: Automatische Erkennung einer durch orale Variation gekennzeichneten Gruppe.

Version	**ASD**
Shahname(V1)	55, 59
Shahname(V2)	55, 13
Shahname(V3)	57, 93
Shahname(V4)	55, 83
Shahname(Durchschnitt V1-V4)	56, 12
Shahname(als eine Tradition)	38, 31

Abbildung 4: Evaluation der erzeugten Stemmata (ASD). Das Stemma der Gesamttradition unter der Annahme nur einer Wurzel evaluiert mit diesen Algorithmen deutlich schlechter als der Durchschnitt der einzelnen Versionen.

Literatur

Andrews, T. L. and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmas. *Literary and Linguistic Computing*, 28(4):504–521.

l Qasim Ferdoussi, A. (1966-1967). *The Shahname - the book of kings.* The Great Islamic Encyclopaedia.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. english in: Soviet Physics Doklady 10 (8) (1966) 707–710.

Lord, A. B. (1960). *The Singer of Tales.* Harvard University Press.

O'Hara, R. J. (1996). Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano*, 27(1):81–88.

Paradis, E. (2012). *Analysis of Phylogenetics and Evolution with R.* Springer, New York, 2nd edition.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20:289–290.

Parry, M. and Parry, A. (1987). *The Making of Homeric Verse: The Collected Papers of Milman Parry.* Oxford University Press.

Ph.V. Baret, C.Macé, P. (2004). Testing methods on an artificially created textual tradition. In *Linguistica Computationale XXIV-XXV*, volume XXIV-XXV, pages 255–281, Pisa-Roma. Instituti Editoriali e Poligrafici Internationali.

Robinson, P. M. and O'Hara, R. J. (1996). Cladistic analysis of an old norse manuscript tradition. *Research in Humanities Computing (4).*

Roelli, P. and Bachmann, D. (2010). Towards generating a stemma of complicated manuscript traditions: Petrus alfonsi's dialogus. *Revue d'histoire des textes*, 5(4):307–321.

Roos, T. and Heikkilä, T. (2009). Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24:417–433.

Rubanovich, J. (2011). *Medieval Oral Literature*, chapter Orality in Medieval Persian Literature, pages 653–680. De Gruyter.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Spencer, M., Davidson, E. A., Barbrook, A., and Howe, C. J. (2004). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 227:503–511.

van Reenen, P., den Hollander, A., and van Mulken, M. (2004). *Studies in Stemmatology II*. Studies in Stemmatology. John Benjamins Publishing Company.

van Reenen, P., van Mulken, M., and Dyk, J. (1996). *Studies in Stemmatology I*. Studies in Stemmatology. John Benjamins Publishing Company.

Yamamoto, K. (2003). *The Oral Background of Persian Epics*. Brill, Leiden.