Abschlussbericht

1 Allgemeine Angaben

1.1 DFG-Geschäftszeichen

LI 2583/1-1 (705 762/806 766)

1.2 Antragssteller

Dr. Johann-Mattis List
* 16.07.1981
Kahlaische Str. 10
07743 Jena
Tel. 03641 686822

mattis.list@shh.mpg.de

1.3 Institut

Max-Planck-Institut für Menschheitsgeschichte
Abteilung für Sprach- und Kulturevolution
Kahlaische Straße 10
07743 Jena

1.4 Gastinstitute während des Aufenthalts

Centre de recherches linguistiques sur l'Asie Orientale
École des Hautes Études en Sciences Sociales
2 Rue de Lille
75007 Paris

Team Adaptation, Integration, Reticulation, Evolution
Université Pierre et Marie Curie
9 quai St Bernard
75005 Paris

1.5 Thema des Projekts

Untersuchung vertikaler und lateraler Aspekte der chinesischen Dialektgeschichte im Rahmen eines interdisziplinären Ansatzes der neue Methoden aus der Biologie für die historische Linguistik adaptiert und gewinnbringend anwendet.

1.6 Berichtszeitraum, Förderungszeitraum insgesamt

01/01/2015 – 31/12/2016

1.7 Liste der wichtigsten Publikationen aus diesem Projekt[1]

1.7.1 Aufsätze in Fachzeitschriften

1. List, J.-M., J. S. Pathmanathan, N. W. Hill, E. Bapteste, and P. Lopez (forthcoming): **Vowel purity and rhyme evidence in Old Chinese reconstruction**. *Lingua Sinica*.

2. *List, J.-M. (forthcoming): **Using network models to analyze Old Chinese rhyme data**. *Bulletin of Chinese Linguistics* 9.2.

3. *List, J.-M., S. Greenhill, and R. Gray (2017): **The potential of automatic word comparison for historical linguistics**. *PLOS ONE* 12.1. 1-18.

4. List, J.-M., J. Pathmanathan, P. Lopez, and E. Bapteste (2016): **Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics**. *Biology Direct* 11.39. 1-17.

5. *List, J.-M. (2016): **Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction**. *Journal of Language Evolution* 1.2. 119-136.

6. Chacon, T. and J.-M. List (2015): **Improved computational models of sound change shed light on the history of the Tukanoan languages**. *Journal of Language Relationship* 13.3. 177-204.

7. *List, J.-M. (2015): **Network perspectives on Chinese dialect history**. *Bulletin of Chinese Linguistics* 8. 42-67.

1.7.2 Aufsätze in Konferenzproceedings

1. **List, Johann-Mattis (2017): **A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets**. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*.

2. *Jäger, G., J.-M. List, and P. Sofroniev (2017): **Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists**. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Long Papers*.

3. *List, J.-M., P. Lopez, and E. Bapteste (2016): **Using sequence similarity networks to identify partial cognates in multilingual wordlists**. In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics. 599-605.

4. **List, J.-M., M. Cysouw, and R. Forkel (2016): **Concepticon. A resource for the linking of concept lists**. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. 2393-2400.

5. *Jäger, G. and J.-M. List (2016): **Investigating the potential of ancestral state reconstruction algorithms in historical linguistics**. In: *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*.

6. Jäger, G. and J.-M. List (2015): **Factoring lexical and phonetic phylogenetic characters from word lists**. In: *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*. Eberhard-Karls University.

[1] Studien, die zusammen mit größeren Datensätzen (Gold-Standards, Testdaten, Code) veröffentlicht wurden, sind durch einen vorangestellten Stern markiert. Sofern diese Daten auch unabhängig auf einer Webseite veröffentlicht wurden, ist dies mit zwei vorangestellten Sternen markiert worden. Auf die Daten wird jeweils im detaillierten Bericht genauer eingegangen.

1.7.3 Aufsätze in Sammelbänden

1. List, J.-M. (2017): **Fāngyán** 方言. In: Sybesma, R. (ed.): *Encyclopedia of Chinese language and linguistics*. Brill: Leiden and Boston. 219-225.

2. List, J.-M. (2017): **Contraction**. In: Sybesma, R. (ed.): *Encyclopedia of Chinese language and linguistics*. 672-675.

1.8 Liste der wichtigsten Vorträge aus diesem Projekt

1. List, J.-M. (2016): **Vowel purity and rhyme evidence in Old Chinese reconstruction**. Talk, held at the "29th Meeting on East Asian Linguistics"; (2016/07/24, ParisCentre des Recherches Linguistiques sur l'Asie Orientale.).

2. List, J.-M., P. Lopez, and E. Bapteste (2016): **Using sequence similarity networks to identify partial cognates in multilingual wordlists**. Paper, presented at the conference "Annual Meeting of the Association of Computational Linguistics"; (2016/08/07-12, BerlinAssociation of Computational Linguistics.).

3. List, J.-M. (2016): **Non-tree-like processes in language evolution**. Talk, held at the "EVOLUNET Summer School on Networks"; (2016/07/04-08, RoscoffUPMC.).

4. List, J.-M. (2016): **Modeling language change for the purpose of phylogenetic reconstruction**. Talk, held at the "Laboratoire Dynamique du Langage"; (2016/02/02, LyonUniversité Lumière Lyon 2.).

5. List, J.-M. (2016): **Historical relations between words and their implication for phylogenetic reconstruction**. Talk, held at the "School Of Oriental and African Studies"; (2016/10/13, LondonUniversity of London.).

6. List, J.-M. (2016): **Handling word formation in historical-comparative linguistics**. Paper, presented at the workshop "Workshop on Kiranti Languages"; (2016/12/01-02, ParisCRLAO.).

7. List, J.-M. (2016): **EDICTOR: A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets**. Talk, held at the "School Of Oriental and African Studies"; (2016/10/10, LondonUniversity of London.).

8. List, J.-M. (2016): **Computer-assisted language comparison. Ideas, tools, applications**. Talk, held at the "EVOLAMEP Project"; (2016/01/20, TübingenEberhard-Karls University.).

9. Hill, N. and J.-M. List (2016): **Challenges of representing and analyzing etymological data of South-East Asian languages**. Talk, held at the "46th Poznań Linguistic Meeting"; (2016/09/15-17, PoznańAdam Mickiewicz University.).

10. List, J.-M. (2016): **CLICS 2016. Chances and challenges**. Paper, presented at the workshop "Lexical Semantic Networks and Language Change"; (2016/03/17-18, Santa FeSanta Fe Institute.).

11. List, J.-M. (2016): **Auf dem Weg zu einer computer-gestützten historischen Sprachforschung** [On the way to a computer-assisted approach in historical linguistics]. Talk, held at the "Linguistic Colloqium [Linguistisches Kolloquium]"; (2016/11/09, JenaFriedrich-Schiller-Universität.).

12. List, J.-M. (2016): **Analogies, transfer, and adaptation. Interdisciplinary research on evolutionary dynamics in biology and linguistics**. Talk, held at the "Centre de Recherche"; (2016/05/09, ParisMusée de l'Homme.).

13. List, J.-M. (2015): **The future of the comparative method**. Paper, presented at the conference "Integrating inferences about our past - New findings and current issues in the peopling

of the Pacific and SouthEast Asia"; (2015/06/22/23, JenaMax Planck Institute for the Science of Human History.).

14. List, J.-M. (2015): **Using network models to analyze Old Chinese rhyme data**. Talk, held at the workshop "Recent Advances in Old Chinese Historical Phonology"; (2015/11/05-06, LondonSchool of Oriental and African Studies.).

15. List, J.-M. and T. Chacon (2015): **Towards a cross-linguistic database for historical phonology? A proposal for a machine-readable modeling of phonetic context**. Paper, presented at the workshop "Historical Phonology and Phonological Theory [organized as part of the 48th annual meeting of the SLE]"; (2015/09/04, LeidenSocietas Linguistica Europaea.).

16. List, J.-M. (2015): **Similarities and differences between evolutionary processes in linguistics and biology**. Talk, held at the "Séminaire du LBBE"; (2015/05/27, LyonLaboratoire de Biométrie et Biologie Évolutive.).

17. Jäger, G. and J.-M. List (2015): **Investing the potential of ancestral state reconstruction algorithms in historical linguistics**. Paper, presented at the workshop "Capturing Phylogenetic Algorithms for Linguistics"; (2015/10/26-30, Leiden).

18. List, J.-M. (2015): **Handling phonological and etymological relations in computer-based and computer-assisted frameworks. Theoretical aspects**. Talk, held at the workshop "Towards a Global Language Phylogeny"; (2015/02/23-26, Waiheke IslandMax Planck Institute for the Science of Human History.).

19. Jäger, G. and J.-M. List (2015): **Factoring lexical and phonetic phylogenetic characters from word lists**. Paper, presented at the conference "6th Conference on Quantitative Investigations in Theoretical Linguistics"; (2015/11/04-06, TübingenEberhard-Karls University.).

20. List, J.-M. (2015): **Datasets and software tools for computer-assisted language comparison**. Paper, presented at the workshop "Databases in Historical Linguistics"; (2015/08/20/21, Santa FeSanta Fe Institute.).

21. List, J.-M., M. Cysouw, and R. Forkel (2015): **Concepticon: A resource for the linking of concept lists**. Talk, held at the workshop "Language Comparison with Linguistic Databases"; (2015/04/30, LeipzigMax Planck Institute for Evolutionary Anthropology.).

22. List, J.-M. (2015): **Automatic identification of historically related words**. Talk, held at the workshop "Strings and Structures – Codes of Sense and Function"; (2015/05/20-21, Cologne-University of Cologne.).

23. Chacon, T. and J.-M. List (2015): **A sound-change-based phylogeny of the Tukanoan language family. Using ordered multistate models for phylogenetic reconstruction**. Paper, presented at the workshop "3rd Workshop Towards a Global Language Phylogeny"; (2015/10/21-23, JenaMax Planck Institute for the Science of Human History.).

1.9 Liste der Datenbanken und Softwareapplikationen aus diesem Projekt

1. List, J.-M., M. Cysouw, and R. Forkel (2016): **Concepticon**. Version 1.0. http://concepticon.clld.org.

2. List, J.-M. and R. Forkel (2016): **LingPy**. Version 2.6. http://lingpy.org.

3. List, J.-M (2016): **EDICTOR**. Version 0.1. http://edictor.digling.org.

2 Arbeits- und Ergebnisbericht

2.1 Ausgangsfragen und Zielsetzung des Projekts

Trotz einer langen Forschungstradition wissen wir nach wie vor wenig über den Ursprung und die Diversifizierung der chinesischen Dialekte. Der Grund liegt in einer komplexen Interaktion zwischen konvergenten und divergenten Kräften, die deren Entwicklung charakterisiert. Wir blicken, wenn wir die chinesischen Dialekte untersuchen, auf eine Sprachgeschichte, die von Phasen intensiven Sprachkontakts sowohl zwischen den Dialektvarietäten untereinander als auch zwischen den Dialekten und anderen Sprachen, von zahllosen Migrationsbewegungen und von einer sehr speziellen sprachlichen Tradition geprägt ist, die ihre Besonderheit vor allem auch der chinesischen Schrift verdankt, welche einen nicht unerheblichen Beitrag zur Wahrnehmung der chinesischen Sprache durch ihre Sprecher geleistet hat. Die Entwicklung der Dialekte wird daher unter Linguisten kontrovers diskutiert, und obwohl Fortschritte auf dem Gebiet der Untergruppen gemacht wurden, ist die Forschung nach wie vor weit davon entfernt, eine *communis opinio* etabliert zu haben (Norman 2003).

Das Ziel dieses Projektes war es, Licht in den dunklen Wald der chinesischen Dialektgeschichte zu bringen. Grundidee war es dabei, die chinesische Sprachgeschichte mit Hilfe interdisziplinärer und datenbasierter Ansätze von einem neuen Blickwinkel zu erforschen. Dabei sollten aktuelle Methoden, die in den letzten Jahren in der Bioinformatik zur Untersuchung lateralen Gentransfers entwickelt wurden für die historische Linguistik erschlossen und auf die chinesischen Dialektdaten angewendet werden. Ein Hauptaugenmerk lag dabei auf der Unterscheidung vertikaler, aus sprachlicher Abstammung resultierender, und lateraler, auf sprachlichem Kontakt beruhender, Beziehungen zwischen den chinesischen Dialekten.

Aufgrund des interdisziplinären Charakters wurde das Projekt an zwei Host-Institutionen durchgeführt: Für Detailfragen zur historischen Linguistik und zur chinesischen Sprachgeschichte kooperierte ich mit Sinologen des *Centre des Recherches Linguistiques sur l'Asie Orientale* (CRLAO) an der École des Hautes Études en Sciences Sociales. Fragen zum lateralen Gentransfer und zur nicht-vertikalen (nicht-baumhaften) Evolution im Allgemeinen konnte ich in der Équipe *AIRE* (*Adaption, Intégration, Réticulation, Évolution*) an der Université Pierre et Marie Curie nachgehen.

Methodologisch sollten generelle *Workflows* zur Untersuchung sprachlicher Daten in der historischen Linguistik entwickelt und auf chinesische Sprachdaten angewandt werden. Konkret sollten bestehende Methoden zur automatischen Erkennung von Kognaten und zur automatischen phonetischen Alinierung dabei ergänzt werden um (1) neue Methoden zur Identifizierung komplexer etymologischer Relationen mit Hilfe explorativer Datenanalyse, (2) neue Ansätze zur Modellierung komplexer etymologischer Beziehungen in historischen Szenarios (Bäumen und Netzwerken). Diese neuen Methoden sollten es schließlich ermöglichen, (3) *evolutionäre Netzwerke* der chinesischen Sprachen zu rekonstruieren, die im Gegensatz zu Netzwerken zur Datenexploration gewurzelt sind, also eine klare gerichtete Entwicklung aufzeigen, die aber nicht nur Vererbungs-, sondern auch Entlehnungsprozesse modelliert. Abgerundet werden sollte das Ganze durch die Erstellung eines umfangreichen lexikalischen Datensatzes grundlegender chinesischer Dialektvarietäten, welcher im Rahmen des Projektes veröffentlicht werden sollte.

Theoretisch sollte eine intensive Einarbeitung und Weiterbildung im Bezug auf die Schnittpunkte zwischen Biologie und historischer Linguistik auf der einen, und im Bezug auf die chinesische Sprachgeschichte auf der anderen Seite erfolgen. An beiden Host-Institutionen standen mir dabei namhafte Wissenschaftler zur Seite, die mich theoretisch und praktisch unterstützten. Auf biologischer Seite waren dies Philippe Lopez und Eric Bapteste, Leiter der Équipe *AIRE*. Auf linguistisch-sinologischer Seite waren dies Guillaume Jacques und Laurent Sagart, beide Seniorwissenschaftler am CRLAO.

2.2 Entwicklung der durchgeführten Arbeiten

Die grundlegenden Aspekte meiner Arbeiten lassen sich den vier Hauptaspekten (a) Daten, (b) Methoden, (c) Analysen und (d) Schnittstellen unterordnen. *Daten* bezeichnen hierbei die grundlegende Arbeiten zu Formaten und Standards, sowie die Erstellung konkreter Datensätze. *Methoden* bezeichnen die Algorithmen, die entwickelt wurden um entweder Daten automatisch zu analysieren, oder um mit Hilfe voranalysierter Daten neue Analysen durchzuführen. *Analysen* bezeichnen die konkrete Anwendung der Methoden und eine wissenschaftliche Bearbeitung derselben im Rahmen einer Publikation. *Schnittstellen* bezeichnen Tools und Interfaces die entweder die rasche Annotation der Daten erleichtern, oder es ermöglichen, die Ergebnisse der Analysen interaktiv zu betrachten. Tabelle 1 zeigt einen Überblick über die wichtigsten Arbeiten, die in diesem Zusammenhang während des Projekts in Angriff genommen worden, inklusive von Ergebnissen in Form von Publikationen.

Während die Projektziele im Großen erfolgreich in Angriff genommen werden konnte, taten sich im Kleinen zuweilen Schwierigkeiten auf. Probleme entstanden vor allem in der Sammlung der Daten, aber auch in der Anwendung von Netzwerkalgorithmen auf neuartig kodierte Datensätze. Im Folgenden werde ich gesondert auf die drei im Nachhinein von mir als wichtigste erachteten Punkte meiner Arbeit eingehen, die Entwicklung von Algorithmen, die Entwicklung von Datensätzen, die Entwicklung von Interfaces, spezifische Arbeiten zum Altchinesischen, sowie interdisziplinäre Arbeit und interdisziplinärer Transfer. Wie aus der Tabelle 1 ersichtlich wird, überschneiden sich die Kernfragen mit den Aspekten, so dass die Frage der Kognazität unter Daten-, Methoden-, Analysen- und Schnittstellengesichtspunkten betrachtet wurde. Daher ist jeder der einzelnen im Folgenden detaillierten Punkte immer im Zusammenhang mit anderen Aspekten zu betrachten, was ich versucht habe, in der Tabelle entsprechend darzustellen.

Aspekt	Daten	Methoden	Analysen	Schnittstellen	Publikationen
Kognazität	Formate und Testdatensätze	netzwerkbasierte Entdeckung von verwandten Wörtern	Evaluierungs-studien	EDICTOR-Tool zur Annotation partieller Kogna-ten	List 2017, List und Forkel 2016, List u. a. 2016c, List u. a. 2017
Direktionale Pro-zesse im Sprach-wandel	Formate und Testdatensätze	Rekonstruktion von Entwick-lungsszenarien aus multi-state-Merkmalen mit Hilfe von Step-matritzen	lexikalische Ent-wicklung von Substantiven in chinesischen Dialekten	Interaktive Visua-lisierung lexikali-scher Evolution	Chacon und List 2015, List 2016, List u. a. 2016a
Rekonstruktion des Altchinesi-schen	Reimannotationen	Netzwerkanalyse von Reimdaten	quantitativer Vergleich von Rekonstruktions-systemen	interaktiver Reim-browser	List u. a. 2017
interdisziplinärer Transfer	–	Vergleich auf Ba-sis von prozess-basierten Analo-gien	Revision be-stehender und Entwicklung neuer potentiell fruchtbarer Ana-logien in Biologie und Linguistik	–	List u. a. 2016b

Tabelle 1: Grundlegende Arbeiten im Projekt, dargestellt anhand der vier Hauptaspekte *Daten*, *Methoden*, *Analysen* und *Schnittstellen*. Die letzte Spalte verweist auf die wichtigsten Publikationen, in denen die Aspekte behandelt wurden.

2.2.1 Entwicklung von Algorithmen und neuen Verfahren

Eines der großen Ziele des Projektes war es, die bestehenden Algorithmen, die ich vor Projektbeginn im Rahmen verschiedener Publikationen (List 2014) entwickelt und veröffentlicht hatte, weiter auszubauen. Konkret hieß das, das vor allem in Bezug auf die Methoden zur automati-

schen Lehnworterkennung neue Verfahren entwickelt werden sollten, die Wörter nicht nur als unteilbare Einheiten betrachten, sondern Verwandtschaft zwischen Wörtern (*Kognazität*) auch partiell erfassen können. Die Erfassung von Teilkognazität ist fundamental für den Vergleich von Sprachen, in denen die Komposition eine große Rolle in der Wortbildung spielt. Dies gilt bei den indogermanischen Sprachen vor allem auch für das Deutsche, wo schon kleine Kinder sehr früh lernen, neue Wörter aus alten zu bilden, indem sie diese aneinanderreihen. Im Chinesischen ist diese Tendenz sogar noch stärker, da monomorphemische Wörter gewöhnlich nur eine Silbe lang sind, wogegen in der Umgangssprache eine starke Tendenz besteht (deren Gründe noch nicht gut erforscht sind), bisyllabische Wörter zu benutzen. Während im Deutschen Basiswörter wie *Sonne* und *Mond*, monomorphemisch sind, bevorzugen die chinesischen Dialekte hingegen bisyllablische Konstruktionen, und wir finden *yuèliàng* 月亮 (wörtl. “Mondschein”) im Mandarinchinesischen gegenüber *yuèguāng* 月光 (wörtl. “Mondglanz”) in vielen anderen Dialekten, vorwiegend im Süden Chinas. Da, wie wir an der Aussprache im Mandarinchinesischen sehen können, beide Wörter das erste Element teilen, können wir nicht sagen, dass beide Wörter nicht verwandt sind. Wir können aber auch nicht sagen, dass sie beide auf denselben Ursprung zurückgehen, da es ja offensichtlich ist, dass beiden Wörtern eine unterschiedliche Konzeptualisierung zugrunde liegt. Diese Teilkognazität ist ein großes Problem in der historischen Linguistik, da sie bisher sehr selten erforscht wurde, und keiner der modernen und populären Algorithmen zur Rekonstruktion von Sprachphylogenien mit dieser Form von Daten umgehen kann.

Gleich zu Beginn meines Projektes stellte ich fest, dass der Entwicklung von Algorithmen, die Teilkognazität in phylogenetischen Analysen berücksichtigen, die Entwicklung von Verfahren, um diese Teilkognazität zu erkennen, vorausgehen muss, da es nur so möglich ist, ausreichend große Datensätze zu erstellen, die dann auch analysiert werden können. Die Entwicklung eines solchen Verfahrens mit Hilfe von Netzwerkmethoden aus der Biologie verzögerte sich eine Weile, da zunächst Daten zum Testen erstellt werden mussten, und diese Erstellung voraussetzte, Interfaces zum schnellen und vor allem konsistenten annotieren von Daten zur Verfügung zu haben (siehe die folgenden Abschnitte 2.2.2 und 2.2.3 für eine detaillierte Beschreibung dieser beiden Schritte). Ferner erforderte die Entwicklung des Algorithmus zur Erkennung von partiellen Kognatheitsbeziehungen auch eine intensive Einarbeitung in die Methoden des biologischen Labors. Die Entwicklung des Algorithmus, der direkt auf den im biologischen Labor entwickelten Ähnlichkeitsnetzwerken (*similarity networks*) aufbaute und diese soweit modifizierte, dass sie gewinnbringend in der Linguistik verwendet werden können, verlief danach aber erfolgreich, und nachdem ein erster Prototyp des Algorithmus gegen Ende des Jahres 2015 erstellt worden war, konnten wir uns erfolgreich für die prestigeträchtige Konferenz der Association of Computational Linguistics bewerben, wo wir den Algorithmus im Rahmen eines Short Papers vorstellten (List u. a. 2016c). Das Verfahren mitsamt den Daten wurde im Rahmen dieser Publikation online ebenfalls zur Verfügung gestellt, und der Algorithmus wurde als Teil der Version 2.5 der Python Softwarebibliothek LingPy (http://lingpy.org), deren erste Version ich vor fast sechs Jahren im Rahmen meiner Dissertation veröffentlichte, frei zur Verfügung gestellt (List und Forkel 2016).

Dadurch, dass die automatische Entdeckung partieller Kognaten wie auch deren konsistente formelle Annotation sauber gelöst werden konnten, wurde es möglich, endlich auch Algorithmen zu testen, die partielle Kognazität in der Berechnung von Sprachphylogenien miteinbeziehen. Leider konnten im Gegensatz zu dem, was ich noch im Antrag erhofft hatte, hierbei jedoch nur vorläufige Ansätze entwickelt werden, da es sich als zu schwierig herausstellte, komplette phylogenetische Netze, die sowohl Vererbung als auch Entlehnung sprachlichen Materials handhaben, automatisch zu rekonstruieren. Meine Arbeiten zur phylogenetischen Rekonstruktion mit Hilfe von Daten, die partielle Kognazität widerspiegeln, sind somit ein erster Schritt hin zu realistischeren Modellen. Sie verharren nach wie vor im linguistischen Stammbaummodell, ermöglichen es aber, die retrogressiven Prozesse der Komposition vielversprechend zu modellieren. Als grundlegende Idee, die ich im Antrag noch nicht entwickelt hatte, stellte sich die

sogenannte auf multiplen *states* basierte Merkmalskodierung heraus. In der Biologie wurde und wird sie heute vorwiegend praktiziert, indem beispielsweise für morphologische Merkmale mehrere Stufen angesetzt werden, die sich ausprägen können (bspw. *langer* vs. *mittlerer* vs. *kurzer Schnabel*). Dabei wird in der Biologie durch sogenannte Stepmatritzen vorgegeben, in welche Richtung sich diese Merkmale entwickeln können, wobei bestimmte Entwicklungen Zwischenstufen erfordern (bspw. würde der Weg von einem kurzen zu einem langen Schnabel immer einen mittleren Schnabel voraussetzen). In der Linguistik wurden die Daten seit den frühen Arbeiten von Gray und Atkinson (2003) und Atkinson und Gray (2006) fast ausschließlich binär kodiert, wodurch jegliche Form der Abstufung in den phylogenetischen Modellen nicht modelliert werden konnte, und stattdessen jeglicher Wandel immer nur als ein Prozess des Gewinns oder des Verlustes aufgefasst wurde (sogenannte *birth-death*- oder *gain-loss*-Modelle).

Das Problem der multi-state-Kodierung ist, dass es wenig Softwareapplikationen gibt, die es ermöglichen sie ohne Beschränkungen auf linguistische Daten anzuwenden. Daher war ich gezwungen, meine eigenen Implementationen zu schreiben, die allerdings, was auch von Gutachtern und Kollegen kritisiert wurden, in dem Parsimonieparadigma verharren, Evolution von linguistischen Merkmalen also nach dem Prinzip der Sparsamkeit der evolutionären Prozesse (im gewissen Sinne *Okhams Rasiermesser*) inferieren. Es war jedoch meiner Ansicht nach weniger wichtig, die linguistischen Prozesse im Rahmen von komplexen bayesianischen Modellen vollständig abzubilden (was ohnehin problematisch wäre, da mir Kollegen versicherten, dass meine Daten nicht ausreichen würden, um sinnvolle Ergebnisse mit diesen Ansätzen, die aus der Masse an Daten lernen, zu finden), als vielmehr zu zeigen, was prinzipiell möglich ist. In dieser Hinsicht schließlich, waren die Ansätze zur multi-state-Kodierung erfolgreich, und ich konnte nicht nur zeigen, wie sich prinzipiell die lexikalische Entwicklung chinesischer Dialekte unter Berücksichtigung der Komposition als Hauptprozess des lexikalischen Wandels modellieren lassen (List 2016), sondern auch zeigen, dass Ähnliches auch in Bezug auf Lautwandelprozesse möglich ist. Dafür kollaborierte ich mit Thiago Chacon (Universität Brasília) zu Daten der Tukanofamilie, und es gelang uns, mit Hilfe der konsequenten multi-state-Kodierung von Lautwandelprozesse, eine neue Klassifikation dieser Familie zu entwickeln, die zentrale Erkenntnisse aus bestehende vereint.

2.2.2 Entwicklung von Datensätzen

Bei der Entwicklung eines großen, chinesische Dialekte in der Breite und Tiefe umfassenden Datensatzes kristallisierten sich bereits in den ersten Monaten gewissen Probleme heraus. Zum einen erwiesen sich die meisten veröffentlichten Daten über die chinesischen Dialekte als nicht brauchbar, da diese nicht auf den onomasiologischen Character einer lexikostatistischen Datenbank zugeschnitten sind und oftmals nur sporadische Übersetzungen von Dialektwörtern ins Mandarinchinesische aufweisen, was eine gezielte Bearbeitung oder Digitalisierung unmöglich macht. Zum anderen zeigte sich, dass ohne spezifische Infrastruktur die sehr diversen Datensätze nicht verglichen werden können. Daher konzentrierte ich mich zuerst auf die grundlegenden Aspekte der Infrastruktur, indem ich mit Kollegen aus Deutschland die Entwicklung des Concepticon-Projektes weiter vorantrieb, das dann auch 2016 offiziell in Version 1.0 veröffentlicht wurde. Das Concepticon ist eine Sammlung von inzwischen fast 200 verschiedenen Questionnaires, die alle auf einen einheitlichen Katalog von Konzepten verlinkt wurden, und somit direkt verglichen werden können, unabhängig von Sprache und Zweck des ursprünglichen Questionnaires. Der Vorteil dieses Vorgehens ist, dass bei der Erschließung großer Datensätze die Glossen in den Questionnaires (also die Labels für die Bedeutungen der Wörter in den verschiedenen Sprachen) einheitlich verglichen werden können, wobei das Concepticon als *tertium comparationis* funktioniert. Mehr Informationen zum Concepticon sind auf der Projektwebseite (http://concepticon.clld.org) sowie in List u. a. (2016a) zu finden.

Im Rahmen der kollaborativen Arbeit an der Concepticon-Initiative mit Forschern des Max-Planck-Instituts Jena fiel auf, dass umfassende Formate, die zur Pflege und zum Austausch von sprachübergreifenden Daten (*cross-linguistic data*) dienen, derzeit nicht vorhanden sind, aber

dringend benötigt werden. Mit dem Concepticon war ein erster Schritt zur Vergleichbarkeit von Daten erstellt worden, auch wenn es ohne Zweifel noch dauern wird, bis dieses vollständig in die wissenschaftliche Arbeit integriert und etabliert ist. Ähnlich hatte sich in den letzten Jahren die Glottolot-Initiative (http://glottolog.org, Hammarström u. a. 2015) mehr und mehr zum Quasistandard und zur mächtigen Alternative gegenüber Ethnologue (http://www.ethnologue.com, Lewis und Fennig 2013) entwickelt. Vor allem in Bezug auf die Dokumentation von sprachlichen Varietäten wie Dialekten ist Glottolog sehr viel flexibler, zumal ich mit Hilfe der Max-Planck-Gesellschaft die Möglichkeit hatte, direkt mit den Herausgebern von Glottolog zusammenzuarbeiten. Was nach wie vor fehlte, war neben verbindlichen Kriterien zur Darstellung sprachlicher Formen (Wörter, Morpheme) in Datensätzen, vor allem eine übergreifende Formatinitiative, die Vorschriften und Empfehlungen herausgibt, wie lexikalische sprachübergreifende Daten möglichst einheitlich erstellt werden können. Diese Initiative wurde im Rahmen der Zusammenarbeit mit Kollegen des Max-Planck-Instituts für Menschheitsgeschichte unter dem Label *CLDF: Cross-Linguistic Data Formats* begründet (Forkel u. a. 2015) und umfasst neben der offiziellen Webpräsentation der Standards (http://cldf.clld.org) und aktuellen Diskussionen und Beispielen, sowie der Software-API (https://github.com/glottobank/cldf/), vor allem auch konkrete Beispiele, wie Daten konsistent erstellt und bearbeitet werden können. Dabei lag mein Beitrag vor allem auf den Bereichen, die für mein Forschungsprojekt unabdingbar waren, also der konsistenten Kodierung von partiellen Kognazitätsbeziehung zwischen Wörtern, einschließlich deren Alinierung, sowie in der Einbindung dieser Formate in meine Forschungssoftware (siehe Abschnitt 2.2.1) und die Interfaces (siehe Abschnitt 2.2.3) zur Vor- und Nachbearbeitung der Forschungsdaten.

Meine Mitarbeit an der Etablierung von Standardformaten für die Pflege und den Austausch lexikalischer sprachübergreifender Daten ermöglichte es mir schließlich, verschiedenste Testdatensätze zu erstellen, mit deren Hilfe ich die Algorithmen, die im Projekt entwickelt werden sollte, direkt testen konnte. Ohne diese Daten wäre die Entwicklung kaum möglich gewesen, da erst manuell kuratierte Testdaten uns zeigen, wie gut Algorithmen mit bestimmten Problemen umgehen können. Diese Datensätze umfassen einen umfassenden Goldstandard zur allgemeinen Kognatenerkennung (List u. a. 2017), einen Goldstandard zur partiellen Kognatenerkennung (List u. a. 2016c), einen von Ben Hamed und Wang (2006) modifizierten Datensatz handannotierter partieller Kognaten im Chinesischen, der vor allem für phylogenetische Experimente wichtig ist (List 2016). Vor allem die Kodierung partieller Kognaten, die ohne das EDICTOR-Interface, welches ich während des Projekts entwickelte (siehe Abschnitt 2.2.3), und die im Rahmen der CLDF-Initiative neu entwickelten Formate zur Repräsentation dieser Daten, stellen einen großen Fortschritt für die historische Linguistik im Allgemeinen dar, was vor allem auch daran deutlich wird, dass die Frage partieller Kognazität weder in der quantitativen noch der qualitativen Linguistik bisher genauer behandelt wurde.

Ein weiterer Datensatz, der sich allerdings nicht mit den chinesischen Dialekten, sondern mit ihrer Vorgängervarietät, dem Altchinesischen befasst, wurde im Rahmen meiner Arbeiten zur Reimanalyse im Altchinesischen erstellt (siehe Abschnitt 2.2.4). Dieser Datensatz umfasst eine digitalisierte Form der Reimannotationen, die (Baxter 1992) für die berühmte Gedichtkollektion des Shījīng 詩經 “Buch der Lieder” erstellte. In dieser Analyse annotierte Baxter, welche Wörter in welchen Versen der Gedichte jeweils miteinander reimten. Die erweiterte Analyse der Reimokkurrenzen erlaubt uns Rückschlüsse auf die Aussprache des Altchinesischen, welches die Sprache ist, aus der später alle chinesischen Dialekte hervorgingen. Dies Analysen wurden traditionell manuell ausgeführt, und wurden bisher noch nie digital zugänglich gemacht, weshalb die Resource in vielfacher Hinsicht als wertvoll für die historische Linguistik des Chinesischen erachtet werden kann. Diese Resource wurde zusammen mit einer exemplarischen Netzwerkanalyse der Reime (List u. a. 2017), die sich derzeit im Druck befindet, online zur Verfügung gestellt, und auch um ein Interface ergänzt, welches es Interessierten ermöglicht, die Reime interaktiv zu untersuchen. Die Daten entstanden im Rahmen einer Kollaboration mit Laurent Sagart, meinem Gastgeber am sinologischen Institut in Paris. Laurent Sagart stellte mir

auch die rekonstruierten Aussprachen seiner neuen Rekonstruktion des Altchinesischen zur Verfügung (Baxter und Sagart 2014), die in den Datensatz aufgenommen wurden. Die Daten wurden in einer weiteren kollaborativen Studie mit den Kollegen aus dem biologischen Labor (siehe genauer im Abschnitt 2.2.4) um weitere Rekonstruktionen namhafter Linguisten (Karlgren 1954, SCHUESSLER, Starostin 1989, Pān 2000, Zhèngzhāng 2003) erweitert (List u. a. im Erscheinen).

Ursprünglich war geplant, zu Ende des Projektes alle Daten im Rahmen einer Onlinedatenbank zu sammeln und öffentlich zu machen. Leider war es nicht möglich, dieses Projekt in der vorgesehenen Zeit zu beenden, auch weil zwei neuere Datensätze (CIHUI, Liú Lìlǐ 刘俐李 u. a. 2007) erst vor Kurzem vollständig digitalisiert werden konnten. Ein Prototyp dieser Datenbank ist derzeit auf GitHub einsehbar, unter dem Arbeitstitel *Chinese Dialect Database* (https://github.com/digling/cddb). Diese Datenbank wird voraussichtlich noch in diesem Jahr veröffentlicht und wird neben den beiden neuen Datensätzen, die noch nicht vollständig analysiert worden sind (CIHUI, Liú Lìlǐ 刘俐李 u. a. 2007), sowie weitere Teildatensätze (Coblin 2015, Norman 2003) enthalten. Ein wichtiger Teil der Arbeit bestand auch in der Standardisierung von Daten und der Normalisierung von Bezeichnungen für Dialektpunkte (geographische Koordinaten, Literatur, alternative Namen, Untergruppen), sowie der Postulierung grober Klassifikationen, die von einer breiten Mehrheit chinesischer Linguisten unterstützt werden. Die Ergebnisse dieser Arbeit wurden dem Glottolog-Projekt zur Verfügung gestellt und werden die derzeit nicht zufriedenstellende und zuweilen fehlerhafte Klassifikation chinesischer Varietäten im Projekt mit dem für Mitte dieses Jahres geplanten neuen Release von Glottolog ablösen.

2.2.3 Entwicklung von Interfaces

Unter der Entwicklung von Interfaces verstehe ich Schnittstellen, die in einem computergestützten Framework Menschen die Kommunikation mit Computern erleichtern. Das mag zunächst trivial scheinen, jedoch ist die Bedeutung von Schnittstellen für die Wissenschaft nicht zu unterschätzen. Biologen benutzen beispielsweise schon seit langer Zeit spezielle Alinierungseditoren, mit denen sie Alinierungen, die zunächst automatisch erstellt wurden, korrigieren können. Obwohl bereits Dixon und Kroeber (1919) in ihrer Analyse indigener Sprachfamilien in Kalifornien explizite Alinierungen verwenden, wurden Alinierungen, welche aufzeigen, wie genau zwei oder mehr Wörter verwandt sind, in der historischen Linguistik bisher kaum verwendet. Während die Alinierung in der Linguistik bei sehr nah verwandten Wörtern trivial ist (vgl. bspw. English *hand* und Deutsch *Hand*, wo klar ist, welche Laute sich entsprechen), kann es bei entfernteren oder den Linguisten nicht vertrauten Sprachen schnell zu Problemen und Missverständnissen führen. Während ein Sinologe beispielsweise problemlos sieht, wo sich Wörter in Chinesischen Dialekten entsprechen, hätten selbst auf historische Sprachwissenschaft spezialisierte Linguisten, die sich nicht mit sinotibetischen Sprachen auskennen, sicher Probleme, die Alinierung von Wörtern wie Hǎikǒu [zit^{3} hau^{31}] und Yínchuān [ʐʅ^{13} t^{h}əu^{0}] vorzunehmen, und würden eventuell annehmen, dass das [t] in Hǎikou dem [t^{h}] in Yínchuān entspricht. Dies stimmt jedoch nicht, da zwischen Silben im Chinesischen immer strikte Morphemgrenzen herrschen, weshalb die Gegenüberstellung von [h] und [t^{h}] die korrekte Alinierung darstellt.

Schon vor Beginn meines Projektes in Paris hatte ich angefangen, an einem Editor für etymologische Wörterbücher zu arbeiten. Die Idee des *EDICTOR*-Tools (http://edictor.digling.org) war es von vornherein, ein auf einfachen Textformaten basiertes Interface zu schaffen, welches plattformunabhängig von Linguisten verwendet werden kann und unterschiedliche Module zur Alinierung, zur Annotation kognater Wörter, und zur Analyse von Morphemen enthält. Während meiner Zeit in Paris konnte ich das Programm entscheidend weiterentwickeln, und neben handlichen und intuitiven Modulen zur Annotaton von Kognaten und zur manuellen Korrektur automatischer Alinierungen auch um ein Modul zur Annotation partieller Kognaten erweitern. Dieses Modul war entscheidend für einen Großteil meiner Arbeit in Paris, insofern, als es mir ermöglichte, die Testdatensätze, die ich so dringend brauchte, um die Algorithmen weiterzuentwickeln, rasch fertigzustellen.

Neben der Programmierung des EDICTORs, welcher in Form einer JavaScript-Applikaton als webbasiertes Tool vorliegt, und somit plattformübergreifend funktioniert, da die Benutzer lediglich die URL aufrufen müssen, erforderte vor allem auch die Konzeption der Formate und der Repräsentation der Daten. Man mag denken, dass dies kein so schweres Problem darstellt, jedoch sollte hierbei nicht vergessen werden, dass die formale Repräsentation partieller Kognaten bisher weder in der klassischen noch in der computerbasierten historischen Linguistik wirklich aufgegriffen wurde. Rekonstruktion im Chinesischen wie auch in vielen anderen sinotibetischen Sprachen ist so gut wie nie eine Rekonstruktion von Lexemen, sondern lediglich eine Rekonstruktion von Morphemen, da sowohl formale Verfahren zur Darstellung komplexer etymologischer Beziehungen zwischen Wörtern fehlen, wie auch anerkannte Methoden, um aus dem Sprachvergleich auf die komplexen Wörter in den Ursprachen zu schließen.

Während meiner Zeit in Paris konnte ich nach langem Ringen schließlich einen Weg finden, sowohl partielle Kognaten formal in einer Form darzustellen, die auch in Computerprogramme eingelesen werden kann, wie es mir auch gelang, das EDICTOR-Tool so zu erweitern, dass ein rasches Analysieren von komplexen etymologischen Beziehungen möglich ist. Die Arbeitserleichterung ist hierbei tatsächlich beachtlich. Ich schätze, dass die manuelle Annotation ohne automatische Voranalyse nun in einem Viertel der Zeit gemacht werden kann. Nicht zu unterschätzen ist auch die Bedeutung der Konsistenz, da die Annotation automatisch eine Textdatei erzeugt, die von Programmen und Software wie LingPy gelesen werden kann. Menschen unterschätzen oft ihre Konsistenz, wenn es um die Erzeugung von Daten geht. Der Vorteil von Schnittstellen ist, dass diese im gleichen Moment, wo der Nutzer die Analyse vornimmt, die Dateneingabe auf Fehler prüfen können. Dies minimiert Fehlannotationen, die in der Vergangenheit in vielen linguistischen Datenbanken festgestellt werden konnten.

Das EDICTOR-Tool wurde nun in Version 0.1 veröffentlicht und wird inzwischen nicht nur von mir, sondern auch von mehr und mehr Kollegen. Ein Artikel, der das Tool vorstellt, wurde für die *Systems Demonstratons* der Konferenz des *European Chapter of the Association of Computational Linguistics* angenommen und vor Kurzem veröffentlicht (List 2017). Das Tool wie auch der Quellcode, der auch in Zukunft weiterentwickelt werden wird, sind online verfügbar, und auch ein Demovideo kann unter https://www.youtube.com/watch?v=IyZuf6SmQM4 abgerufen werden.

Abgesehen von meiner Arbeit am EDICTOR, wurden weitere Schnittstellen während des Projektes von mir entwickelt. Gemein ist allen diesen Projekten, dass sie webbasiert sind, was den Vorteil hat, das die Nutzer, die Webapplikationen von Smartphones und Computern gewohnt sind, sich schnell und auf ihre Intuition vertrauend mit den Tools vertraut machen können. Ein weiterer grundlegender Bestandteil der Schnittstellen ist, dass sie *interaktiv* sind. Das heißt, es geht nicht um statische Visualisierungen von Daten, sondern um interaktive Angebote an Nutzer, Daten und Analysen Schritt für Schritt zu untersuchen. Zwei Beispiele wurden im Zusammenhang mit Arbeiten zum Lautwandel (Chacon und List 2015) und zum lexikalischen Wandel (List 2016) veröffentlicht und ermöglichen es, lautlichen oder lexikalischen Wandel Schritt für Schritt in allen Formen, die von den Algorithmen inferiert wurden, an einer sprachlichen Phylogenie nachzuvollziehen.[2].

Als weitere Schnittstelle, die allerdings nicht mit den chinesischen Dialekten sondern mit der Analyse ihrer Vorgängerspräche, dem Altchinesischen, zu tun hat, sei noch auf den *Shījīng-Browser* (http://digling.org/shijing/) verwiesen, der als Supplement zu einer in Kürze erscheinenden Publikation veröffentlicht wurde (List im Erscheinen). Dieses Tool ermöglicht es, die Reimanalysen von Baxter (1992), welche eine Grundlage für die Rekonstruktion der altchinesischen Aussprache darstellen (siehe auch 2.2.4), interaktiv zu untersuchen, sie mit alternativen Analysen zu vergleichen, und dabei auch die Rekonstruktionen von Baxter und Sagart (2014) und Pān (2000) miteinander zu vergleichen.

[2] Diese Applikationen in diesem Zusammenhang lediglich zu beschreiben wird den Schnittstellen leider nicht gerecht, weshalb ich, anstatt hier mehr zu schreiben, lieber auf die Webseite unter http://digling.github.io/beyond-cognacy-paper/ verweise, auf welcher eine der Schnittstellen abgerufen werden kann

2.2.4 Arbeit am Altchinesischen

Das Altchinesische stellt eine Varietät des Chinesischen dar, die im ersten Jahrtausend v. Chr. gesprochen wurde. Bezeugt ist die Sprache in einer Vielzahl schriftlicher Quellen, jedoch macht es die spezifische, nicht primär phonetische, Form der chinesischen Schrift unmöglich, die Aussprache mit Sicherheit zu ermitteln. Um die Aussprache zu rekonstruieren, bedienen sich Forscher unterschiedlicher Indizien. Hierzu gehört neben der Struktur der Schrift vor allem auch die Analyse der Reime, vor allem der Reime, die im *Buch der Lieder*, dem *Shījīng* 詩經, einer Gedichtsammlung, die zwischen 1050 und 600 v. Chr. verfasst wurde, verwendet wurden. Obwohl bereits mit Baxter (1992) erste stochastische Verfahren entwickelt wurden, um die Reime im Shījīng zu analysieren, wurden bisher nur explizite Hypothesen getestet und keine explorativen Analysen gemacht. Durch die Kollaboration mit den Kollegen im biologischen Labor gelang es uns, neue Netzwerkanalysen der Reime durchzuführen, welche bestimmte Aspekte der neuen altchinesischen Rekonstruktion von Baxter und Sagart (2014) bestätigen konnten, unter anderem die Rekonstruktion von Reimen mit der Koda -r (List im Erscheinen), die auf Arbeiten von Starostin (1989) zurückgeht, wie auch die Bestätigung, dass die Rekonstruktionen, die sechs Vokale ansetzen, *purer* sind, da in ihnen äußerst selten Reime mit unterschiedlichen Vokalen auftreten (List u. a. im Erscheinen)

2.2.5 Interdisziplinarität und interdisziplinärer Wissenstransfer

Im biologischen Labor führten wir viele Diskussionen über die Unterschiede zwischen biologischer und linguistischer Evolution, wie auch über mögliche Ähnlichkeiten in den Prozessen. Diese Gespräche waren überaus fruchtbar, da sie mir einen tiefen Einblick in die theoretischen Aspekte der Evolutionsbiologie gewährten, und mir auch halfen, meine eigene Disziplin aus einer anderen Perspektive zu betrachten. Diese Arbeiten zu möglichen und unmöglichen Analogien mündeten schließlich in einer Publikation, in der wir vorschlagen, um problematische Analogien zwischen den Disziplinen zu vermeiden, den Schwerpunkt auf prozess-basierte Analogien zu legen, also solche Analogien, die zwischen Prozessen und nicht zwischen Objekten gemacht werden (List u. a. 2016b). In diesem Zusammenhang schlugen wir auch neue, oder bisher nur selten postulierte Analogien vor, die es ermöglichen könnten, neue gemeinsame Analyseverfahren in Biologie und Linguistik zu entwickeln, oder bestehende Methoden zu übertragen. Hierzu gehörten neben der Verwendung von Similaritätsnetzwerken zur Entdeckung etymologisch verwandter Wörter (welche wir erfolgreich anwenden konnten, vgl. List u. a. 2016c), insbesondere auch die Parallele zwischen Domainassemblierung in der Biologie und Wortkompositon in der Linguistik. Weitere Arbeiten werden erforderlich sein, um de Zweckmäßigkeit dieser Analogien zu untersuchen.

2.3 Darstellung der erreichten Ergebnisse

Wie bereits in der Einleitung und in der Tabelle 1 angedeutet wurde, lassen sich die Ergebnisse unterschiedlichen Aspekten zuordnen, wobei auch die Ergebnisse neben dem klassischen Output in Form von Publikationen wichtige und jeweils auch frei verfügbare Ressourcen zu *Daten*, *Methoden*, *Analysen* und *Schnittstellen* enthalten. Ich werde dies im Folgenden genauer präzisieren und dabei dem Schema in Tabelle 1 folgen.

2.3.1 Untersuchung von Kognazität

Das Projekt führte zu einer inhaltlichen Neufassung von *Kognazität* (Verwandtschaft linguistischer Elemente) in der historischen Linguistik, welche eine direkte Implikation für zukünftige Datenbanken und auch etymologische Wörterbücher hat. Die Neufassung von *Kognazität*, die vor allem auch die Möglichkeit der unvollständigen oder partiellen Kognazität zwischen sprachlichen Elementen betont, ermöglichte (a) die Entwicklung neuer automatischer Verfahren zur

Entdeckung partieller Kognaten (*Methoden*, List u. a. 2016c), (b) die Entwicklung von Schnittstellen zur Korrektur automatischer Analysen (*Schnittstellen*, List 2017), (c) die Erstellung umfangreicher Testdatensätze für neue algorithmische Verfahren (*Daten*, List u. a. 2016c), und wurde (d) in umfangreichen Evaluierungsstudien vorgestellt (*Analysen*, List u. a. 2016c, List u. a. 2017).

2.3.2 Untersuchung direktionaler Prozesse im Sprachwandel

Basierend auf der Neufassung von *Kognazität* ermöglichte das Projekt die Entwicklung von Modellen, die direktionale Prozesse im Sprachwandel im Rahmen phylogenetischer Analysen beschreiben können. Die Entwicklung dieser neuen Modelle, welche diese Prozesse grundlegend als multi-state im Gegensatz zu den bisher gängigen binary-state Modellen beschreiben, ermöglichte (a) die Entwicklung erster Prototypen, die die Prozesse im Rahmen eines Parsimonie-Frameworks mit für Merkmale individuellen Stepmatritzen beschreiben und somit ein Verfahren zur *ancestral state reconstruction*, wie auch zur Rekonstruktion von Phylogenien darstellen (*Methoden*, Chacon und List 2015, List 2016), (b) die Entwicklung von Daten, mit denen die Verfahren getestet werden können (*Daten*, List 2016), (c) die Veröffentlichung von Analysen, die die neuen Verfahren auf die Testdaten anwenden (*Analysen*, Chacon und List 2015, List 2016), sowie (d) die Entwicklung von interaktiven Applikationen, die es erlauben, die Analysergebnisse im Detail zu inspizieren (*Schnittstellen*, Chacon und List 2015, List 2016).

2.3.3 Rekonstruktion des Altchinesischen

Um die Rekonstruktion des Altchinesischen auf ein solideres Fundament zu stellen, wurde im Rahmen des Projektes eine Modellierung der Reimevidenz als Netzwerk entwickelt, die es ermöglicht, konkrete Rekonstruktionsvorschläge direkt zu überprüfen. Dieser neue, datengetriebene Ansatz ermöglichte es, (a) Verfahren zur Netzwerkanalyse von chinesischen Reimdaten zu entwickeln (*Methoden*, List im Erscheinen), welche (b) einen quantitativen Vergleich von Rekonstruktionssystemen ermöglichten (*Analysen*, List u. a. im Erscheinen). Grundlage für das Testen der Methoden waren (c) Datensätze zu Reimannotationen (List im Erscheinen), sowie umfangreiche Datensätze zum Vergleich von Rekonstruktionssystemen (List u. a. im Erscheinen, *Daten*). Mit Hilfe (d) eines interaktiven Reimbrowsers wurde ferner eine *Schnittstelle* geschaffen, die es ermöglicht, Reimanalysen schnell und detailliert zu untersuchen und zu vergleichen.

2.3.4 Interdisziplinärer Transfer

Durch die intensive Diskussion über Disziplingrenzen im Rahmen des Projekts wurde eine *Methode* zur Entdeckung möglicherweise fruchtbarer Analogien auf Basis von Prozessen entwickelt, welche als *Analyse* zur Revision bestehender und zur Postulation potentiell neuer Analogien zwischen Linguistik und Biologie führte (List u. a. 2016b).

2.3.5 Konkrete Erkenntnisse

Neben diesen Abstrakten Ergebnissen führte das Projekt auch zu konkreten Resultaten, von denen ich die wichtigsten im Folgenden kurz auflisten möchte:
Klassifikation der chinesischen Dialekte: Im Rahmen verschiedener Studien wurden unterschiedliche Modelle zur Klassifikation der chinesischen Dialekte getestet, und in all diesen Studien erwies sich der Vorschlag von Sagart (2011) als der Robusteste (List 2015, List 2016). Dies lässt vermuten, dass die postulierte Dreiteilung der chinesischen Dialekte in eine nördliche eine südliche und eine mittlere Gruppe (Norman 2003) historisch falsch ist, und wir vielmehr

von einer verschachtelten Verzweigung ausgehen müssen. Teile dieser Arbeiten haben zu einer Revision und Präzision der Dialektdaten in der neuesten Version von Glottolog beigetragen (Hammarström u. a. 2017).
Rekonstruktion des Altchinesischen: Die Studien zu den Netzwerken der chinesischen Reime liefern zusätzliche Evidenz für die Koda -r im Altchinesischen, wie auch für die Annahme von sechs unterschiedlichen Vokalen (List im Erscheinen, List u. a. im Erscheinen).
Neue Analogien im interdisziplinären Transfer: Während die Wort-Gen-Analogie (Pagel 2009) als überholt und irreleitend angesehen werden sollte, scheint insbesondere de Analogie zwischen der Assemblierung neuer Proteine aus Proteindomänen und der Komposition neuer Wörter aus bestehenden Wörtern in der Wortbildung eine fruchtbare Analogie zu sein, von deren genauere Erforschung sowohl die Biologie als auch die Linguistik stark profitieren könnten.

2.4 Wirtschaftliche Verwertbarkeit der Ergebnisse

Die Ergebnisse sind momentan nicht wirtschaftlich verwertbar. Aufgrund der starken akademischen Ausrichtung der Forschung ist es im Moment eher unwahrscheinlich, dass die Ergebnisse wirtschaftlich Verwendung finden werden.

2.5 Kooperationspartner im In- und Ausland

Eine Vielzahl von Kollegen unterstützten mich während des Projektes konkret als Koautoren, und auch indirekt, in vielfältigen Diskussionen. Insbesondere mit Forschern des Max-Planck-Instituts für Menscheitsgeschichte hielt ich regen Kontakt. Im Folgenden liste ich nur die Kollegen, mit denen ich während der Zeit am intensivsten zusammengearbeitet habe in alphabetischer Reihenfolge.

- Eric Bapteste (UPMC Paris)
- Gerhard Jäger (Universität Tübingen)
- Guillaume Jacques (CRLAO Paris)
- Laurent Sagart (CRLAO Paris)
- Philippe Lopez (UPMC Paris)
- Robert Forkel (MPI Jena)
- Russell Gray (MPI Jena)
- Simon Greenhill (MPI Jena)
- Thiago Chacon (Universität Brasília)

2.6 Qualifikation des wissenschaftlichen Nachwuchses

Wissenschaftlicher Nachwuchs wurde im Rahmen des Projektes in Frankreich selbst nicht direkt betreut. Ich hatte jedoch die Gelegenheit, mit Genehmigung der DFG, ein Blockseminar an der Universität Düsseldorf zum Thema „Python und Javascript für Linguisten” zu geben, in welchem ich Bachelor- und Masterstudenten in die Grundlagen der beiden Programmiersprachen einführen konnte.

3 Zusammenfassung

Im Zuge des Forschungsprojektes „Vertikale und laterale Aspekte der chinesischen Dialektgeschichte" sollte untersucht werden, wie sich die chinesischen Dialekte seit ihrer Entstehung in ihre heutige Form entwickelt haben. Dabei sollten Lösungsansätze für das klassische Problem der chinesischen Sprachwissenschaft, Dialektgruppen lediglich statisch zu definieren, aber Beziehungen unter diesen Dialektgruppen zu ignorieren, mit Hilfe von Netzwerkansätzen, die vertikale und laterale Ansätze der Sprachklassifikation vereinen, entwickelt werden. Die Netzwerkansätze, die sich derzeit in der Evolutionsbiologie großer Beliebtheit erfreuen, sollten ferner als mögliche Alternative zu derzeit vorherrschenden Baummodellen in der quantitativ orientierten historischen Sprachwissenschaft etabliert werden.

Um diese Untersuchung zu ermöglichen, wurden neue Methoden zur automatischen Analyse und Schnittstellen zur manuellen Korrektur maschineller Fehler entwickelt. Diese wurden an Dialektdaten, die für das Projekt eigens digitalisiert wurden, getestet. Basierend auf der quantitativen und qualitativen Analyse dieser Daten konnten erste neue Erkenntnisse, nicht nur über die Entwicklung der chinesischen Dialekte, sondern auch über die Verlässlichkeit von Sprachklassifikationsmethoden gewonnen werden.

Hier zeigte sich insbesondere, dass die derzeit gängigen und beliebten automatischen Methoden zur Ermittlung sprachlicher Phylogenien, die sprachliche Entwicklung meist als baumartige Verzweigung darstellen, nicht annähernd der Komplexität der aktuellen Prozesse gerecht werden, und daher auch oft falsche oder zumindest fragwürdige Ergebnisse produzieren. Lösungsansätze für dieses Problem konnten im Projekt entwickelt werden, liegen derzeit aber lediglich als Prototypen in sehr vereinfachten algorithmischen Frameworks for. Es ist zu hoffen, dass sie in Zukunft weiterentwickelt werden können, um die Ansätze zum automatischen Sprachvergleich und die klassischen Methoden der historischen Linguistik näher zusammenzuführen.

Durch den interdisziplinären Austausch zwischen Biologen und Linguisten, der dadurch gefördert wurde, dass ich in zwei Labors, einem biologischen und einem sinologischen, Arbeiten konnte, gelang es uns ferner, erste Ideen über neue fruchtbare Analogien zwischen evolutionären Prozessen in Biologie und Linguistik zu entwickeln, welche in Zukunft nicht nur methodologischen Transfer anregen könnten, sondern eventuell zu gemeinsamen Forschungsprojekten führen könnten. Als prägnantestes Beispiel ist in diesem Zusammenhang die Ähnlichkeit zwischen der Bildung neuer Proteine durch die Komposition von Proteindomänen in der Biologie und der Bildung neuer Wörter durch die Komposition bestehender Wörter in der Linguistik zu nennen. Da die genaueren Mechanismen dieser kompositionellen Prozesse in beiden Bereichen wenig erforscht sind, könnte eine interdisziplinäre Herangehensweise womöglich wichtige neue Erkenntnisse liefern.

Die Ergebnisse des Projektes wurden in vielfältiger Form veröffentlicht, wobei insbesondere auch Wert auf die Verwendung interaktiver Schnittstellen gelegt wurde, um es interessierten Forschern zu erleichtern, die Ergebnisse von Analysen oder die Daten selbst transparent und bequem zu inspizieren.

Als überraschend stellte sich heraus, dass die klassischen Netzwerkmethoden auch sehr gut für die automatische Analyse von Reimstrukturen geeignet sind. Da wir die Aussprache des Altchinesischen vor allem durch die Reime in alten Gedichten erschließen, erlaubte dieses vor Projektbeginn noch nicht in Betracht gezogene Anwendungsfeld von Netzwerkmethoden, erste quantitative Ansätze zu entwickeln, mit denen bestehende Rekonstruktionen der altchinesischen Phonologie verglichen und evaluiert werden können.

References

Atkinson, Q. D. und R. D. Gray (2006). “How old is the Indo-European language family? Illumination or more moths to the flame?” In: *Phylogenetic methods and the prehistory of languages*. Hrsg. von P. Forster und C. Renfrew. Cambridge, Oxford und Oakville: McDonald Institute for Archaeological Research, 91–109.

Baxter, W. H. (1992). *A handbook of Old Chinese phonology*. Berlin: de Gruyter.

Baxter, W. H. und L. Sagart (2014). *Old Chinese. A new reconstruction*. Oxford: Oxford University Press.

北京大学, B. D., Hrsg. (1964). *Hànyǔ fāngyán cíhuì* 汉语方言词汇 [Chinese dialect vocabularies]. Běijing 北京: Wénzì Gǎigé 文字改革.

Ben Hamed, M. und F. Wang (2006). “Stuck in the forest: Trees, networks and Chinese dialects”. *Diachronica* 23, 29–60.

Chacon, T. C. und J.-M. List (2015). “Improved computational models of sound change shed light on the history of the Tukanoan languages”. *Journal of Language Relationship* 13.3, 177–204.

Coblin, W. S. (2015). *A study of comparative Gàn. In memory of Jerry Norman*. Language and Linguistics Monograph Series 58. Taipei: Institute of Linguistics, Academia Sinica.

Dixon, R. B. und A. L. Kroeber (1919). *Linguistic families of California*. Berkeley: University of California Press.

Forkel, R., M. Dunn, S. Greenhill und J.-M. List (2015). *Cross-linguistic data formats*. GlottoBank Working Group. URL: http://github.com/glottobank/cldf/.

Gray, R. D. und Q. D. Atkinson (2003). “Language-tree divergence times support the Anatolian theory of Indo-European origin”. *Nature* 426.6965, 435–439.

Hammarström, H., R. Forkel, M. Haspelmath und S. Bank (2015). *Glottolog*. Version 2.5. URL: http://glottolog.org.

Hammarström, H., R. Forkel und M. Haspelmath (2017). *Glottolog*. Version 3.0. URL: http://glottolog.org.

Karlgren, B. (1954). “Compendium of phonetics in ancient and archaic Chinese”. *Bulletin of the Museum of Far Eastern Antiquities* 26, 211–367.

Lewis, M. P. und C. D. Fennig, Hrsg. (2013). *Ethnologue. Languages of the world*. URL: http://www.ethnologue.com.

List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.

– (2015). “Network perspectives on Chinese dialect history”. *Bulletin of Chinese Linguistics* 8, 42–67.

– (2016). “Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction”. *Journal of Language Evolution* 1.2, 119–136.

– (2017). “A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.

– (im Erscheinen). “Using network models to analyze Old Chinese rhyme data”. *Bulletin of Chinese Linguistics* 9.2.

List, J.-M. und R. Forkel (2016). *LingPy. A Python library for historical linguistics*. Version 2.5. URL: http://lingpy.org.

List, J.-M., M. Cysouw und R. Forkel (2016a). “Concepticon. A resource for the linking of concept lists”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. “LREC 2016” (Portorož, 23.–28. 05. 2016). Hrsg. von N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk und S. Piperidis. European Language Resources Association (ELRA), 2393–2400.

List, J.-M., J. S. Pathmanathan, P. Lopez und E. Bapteste (2016b). “Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics”. *Biology Direct* 11.39, 1–17.

List, J.-M., P. Lopez und E. Bapteste (2016c). “Using sequence similarity networks to identify partial cognates in multilingual wordlists”. In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics. Berlin, 599–605.

List, J.-M., S. J. Greenhill und R. D. Gray (01/2017). “The potential of automatic word comparison for historical linguistics”. *PLOS ONE* 12.1, 1–18.

List, J.-M., J. S. Pathmanathan, N. W. Hill, E. Bapteste und P. Lopez (im Erscheinen). “Vowel purity and rhyme evidence in Old Chinese reconstruction”. *Lingua Sinica*.

Liú Lìlǐ 刘俐李, Wáng Hóngzhōng 王洪钟 und Bǎi Yíng 柏莹 (2007). *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí* 现代汉语方言核心词・特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng 南京: Fènghuáng 凤凰.

Norman, J. (2003). “The Chinese dialects. Phonology”. In: *The Sino-Tibetan languages*. Hrsg. von G. Thurgood und R. J. LaPolla. London und New York: Routledge, 72–83.

Pagel, M. (2009). “Human language as a culturally transmitted replicator”. *Nature Reviews. Genetics* 10, 405–415.

Sagart, L. (2011). *Classifying Chinese dialects/ Sinitic languages on shared innovations*. Paper, presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28). URL: https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations.

Schuessler, A., Hrsg. (2007). *ABC Etymological dictionary of Old Chinese*. Honolulu: University of Hawai’i Press.

Starostin, S. A. (1989). *Rekonstrukcija drevnekitajskoj fonologičeskoj sistemy (Reconstruction of the phonological system of Old Chinese)*. Moscow: Nauka.

潘悟云, P. W. (2000). *Hànyǔ lìshǐ yīnyùnxué* 汉语历史音韵学 [Chinese historical phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.

郑张尚芳, Z. S. (2003). *Shànggǔ yīnxì* 上古音系 [Old Chinese phonology]. Shànghǎi 上海: Shànghǎi Jiàoyù 上海教育.