# Quantitative historische Linguistik

## 1 Automatische Alinierung

### 1.1 Arbeitsstand

- Sound-Class-Based Phonetic Alignment-Methode (SCA) zur paarweisen und multiplen Alinierung (List 2012b).
- Implementiert in LingPy (Version 2.1.dev, List und Moran forthcoming) für Python3.
- Grundprinzip der Methode ist die Reduktion des IPA-Alphabets auf ein kleineres Set von "Lautklassensymbolen" (verschiedene Modelle werden angeboten, können aber auch frei vom User bestimmt werden). Diese werden kombiniert mit spezifischen Scoring-Matrizzen.
- Paarweise Alinierung unterstützt alle gängigen Alinierungsmodi (lokal, global, semi-global, diagonal).
- Basierend auf dem Prinzip von sonoritätsbasierten Gewichtungen (Sonoritätsprofil), mit deren Hilfe Kontextinformationen als String (prosodischer String) dargestellt werden, können kontextspezifische Gap-Penalties bei der Alinierung verwendet werden.

| **Phonetic Sequence** | j | a | b | ə | l | k | a |
|---|---|---|---|---|---|---|---|
| **SCA Model** | J | A | P | E | L | K | A |
| **ASJP Model** | y | a | b | I | l | k | a |
| **DOLGO Model** | J | V | P | V | R | K | V |
| **Sonority Profile** | 6 | 7 | 1 | 7 | 5 | 1 | 7 |
| **Prosodic String** | # | v | C | v | c | C | > |
| **Relative Weight** | 2.0 | 1.5 | 1.5 | 1.3 | 1.1 | 1.5 | 0.7 |

### 1.2 Goldstandard

- PhonAlign („Phonetic Alignment Database", Arbeitstitel) enthält derzeit 750 MSA-Dateien (siehe Tabelle).
- Alle Daten liegen in IPA-Transkription vor.
- LingPy bietet gängige Evaluierungsmethoden an.

| **Dataset** | **Languages** | **PSA** | **MSA** | **Words** | **Taxa** | **PID** | **Source** |
|---|---|---|---|---|---|---|---|
| Andean | Andean dialects (Aymara, Quechua) | 619 | 76 | 883 | 20 | 55 | SAL |
| Bai | Bai dialects | 889 | 90 | 1416 | 17 | 32 | BDS, Wang 2006 |
| Bulgarian | Bulgarian dialects | 1515 | 152 | 32418 | 197 | 48 | Prokić u. a. 2009 |
| Dutch | Dutch dialects | 500 | 50 | 3024 | 62 | 44 | MAND |
| French | French dialects | 712 | 76 | 3810 | 62 | 41 | TPPSR |
| Germanic | Germanic languages and dialects | 1110 | 111 | 4775 | 45 | 32 | LOE |
| Japanese | Japanese dialects | 219 | 26 | 224 | 10 | 40 | Shirō 1973 |
| Norwegian | Norwegian dialects | 501 | 51 | 2183 | 51 | 46 | NORDAVINDEN |
| Ob-Ugrian | Uralic languages | 444 | 48 | 689 | 21 | 45 | GLD |
| Romance | Romance languages | 297 | 30 | 240 | 8 | 37 | LOE |
| Sinitic | Chinese dialects | 200 | 20 | 20 | 40 | 35 | YINKU |
| Slavic | Slavic languages | 120 | 20 | 81 | 5 | 38 | DERKSEN |

# 2 Automatische Kognatenerkennung

## 2.1 Arbeitstand

- LexStat (List 2012a) ermittelt Kognaten in multiplen Wortlisten, basierend auf zuvor ermittelten regulären Lautkorrespondenzen.
- LexStat ist als Teil von LingPy (Version 2.1.dev) implementiert.
- Wie bei der SCA-Methode sind es wieder Lautklassen, die zur internen Sequenzrepräsentation verwendet werden. LexStat kombiniert hier prosodische Strings und Lautklassen, um kontextbasierte Lautkorrespondenzen erfassen zu können.
- Neben der LexStat-Methode bietet LingPy auch verschiedene Baseline-Methoden an: die Methode von Turchin u. a. (2010), einfache Kognatenzuweisung mit Hilfe der normalisierten Levenshtein-Distanz (NED, „normalized edit distance"), Kognatenzuweisung mit Hilfe von SCA-Distanzscores, wobei SCA-Ähnlichkeitswerte mit Hilfe der Formel von Downey u. a. (2008) in Distanzen umgewandelt werden.

## 2.2 Goldstandard

- CoBeDB („Cognate Benchmark Database", Arbeitstitel) besteht aus 6 lexikostatistischen Datensätzen, für die manuelle Kognatenzuweisungen vorliegen.
- Alle Einträge (Wörter) sind in IPA transkribiert.
- Weitere Wortlisten liegen für spezifische Tests vor.

| **Dataset** | **Languages** | **Items** | **Words** | **Cog.-S.** | *D* | **Taxa** | **Source** |
|---|---|---|---|---|---|---|---|
| BAI | Bai dialects | 110 | 1028 | 205 | 0.10 | 9 | Wang 2006 |
| IEL | Indo-European Languages | 207 | 4393 | 1778 | 0.38 | 20 | IELex |
| JAP | Japanese dialects | 200 | 1985 | 458 | 0.14 | 10 | Shirō 1973 |
| OUG | Uralic languages | 110 | 2055 | 239 | 0.07 | 21 | GLD |
| PAN | Austronesian languages | 210 | 4358 | 2730 | 0.61 | 20 | ABVD |
| SIN | Chinese dialects | 140 | 2789 | 1025 | 0.33 | 15 | YINKU |

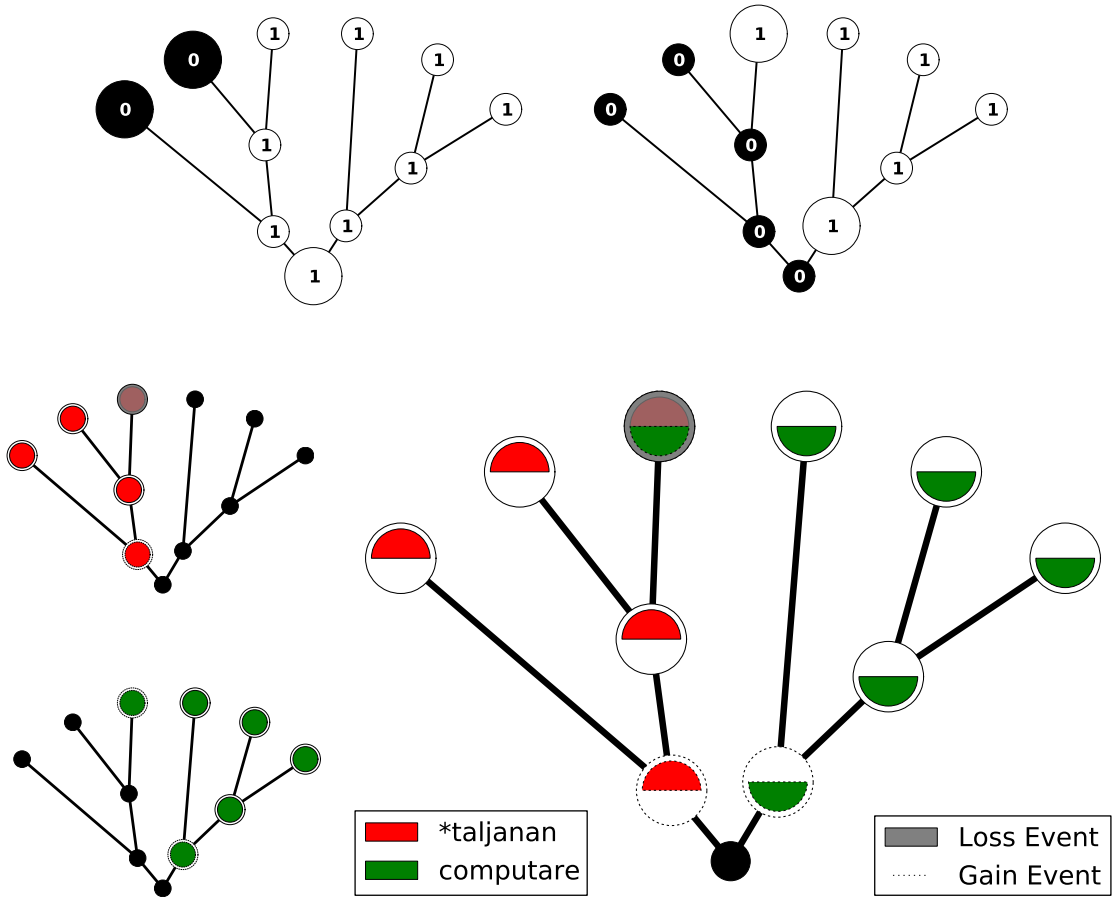
# 3 Automatische Entlehnungserkennung

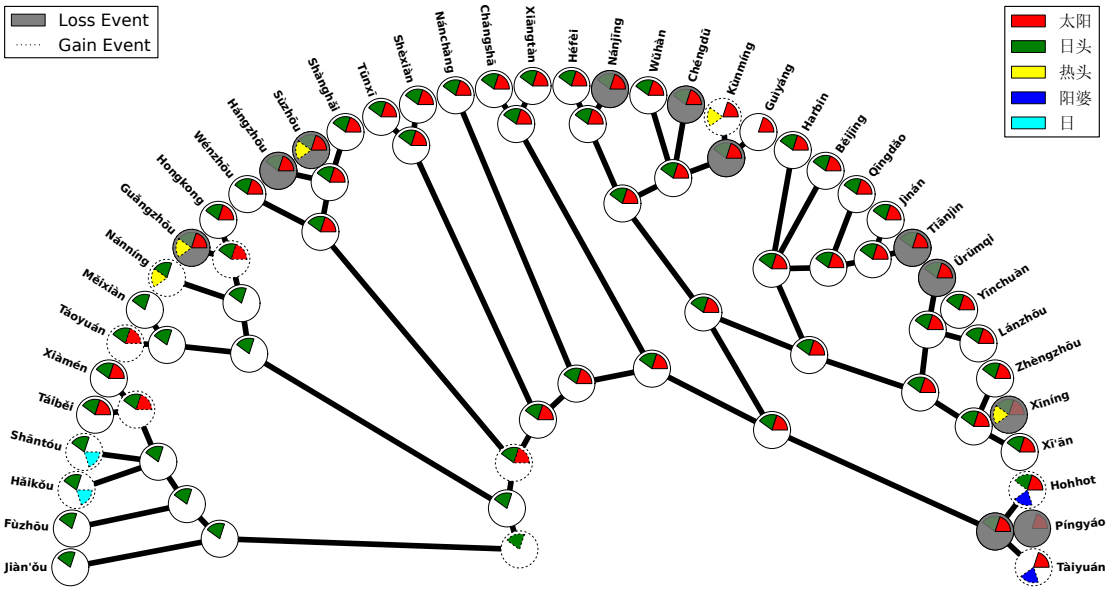
## 3.1 Arbeitsstand

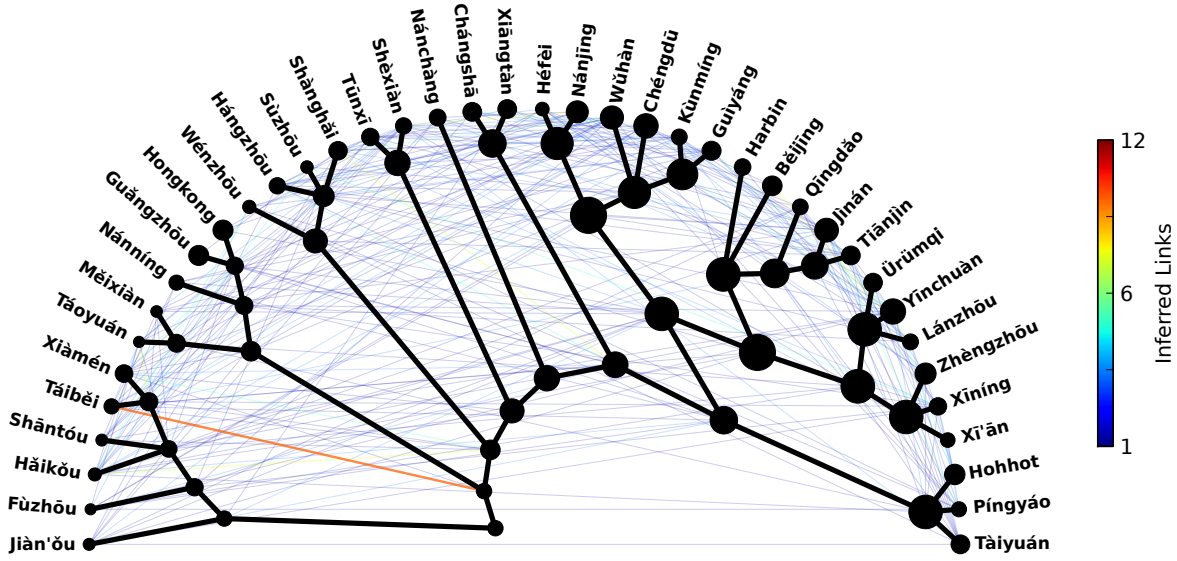
- PhyBo (phylogeny-based borrowing detection, List u. a. forthcoming, Nelson-Sathi u. a. 2011) ist ein Modul von LingPy, das auf Grundlage phyletischer Patterns (binärer Präsenz-Absenz-Matrizzen) und eines vom User definierten Referenzbaumes mögliche Entlehnungen ermittelt. Grundlegende Idee ist dabei, dass Entlehnungsereignisse dadurch entdeckt werden können, dass sie zu Mustern führen, die nicht mit einem traditionellen Baummodell kompatibel sind.

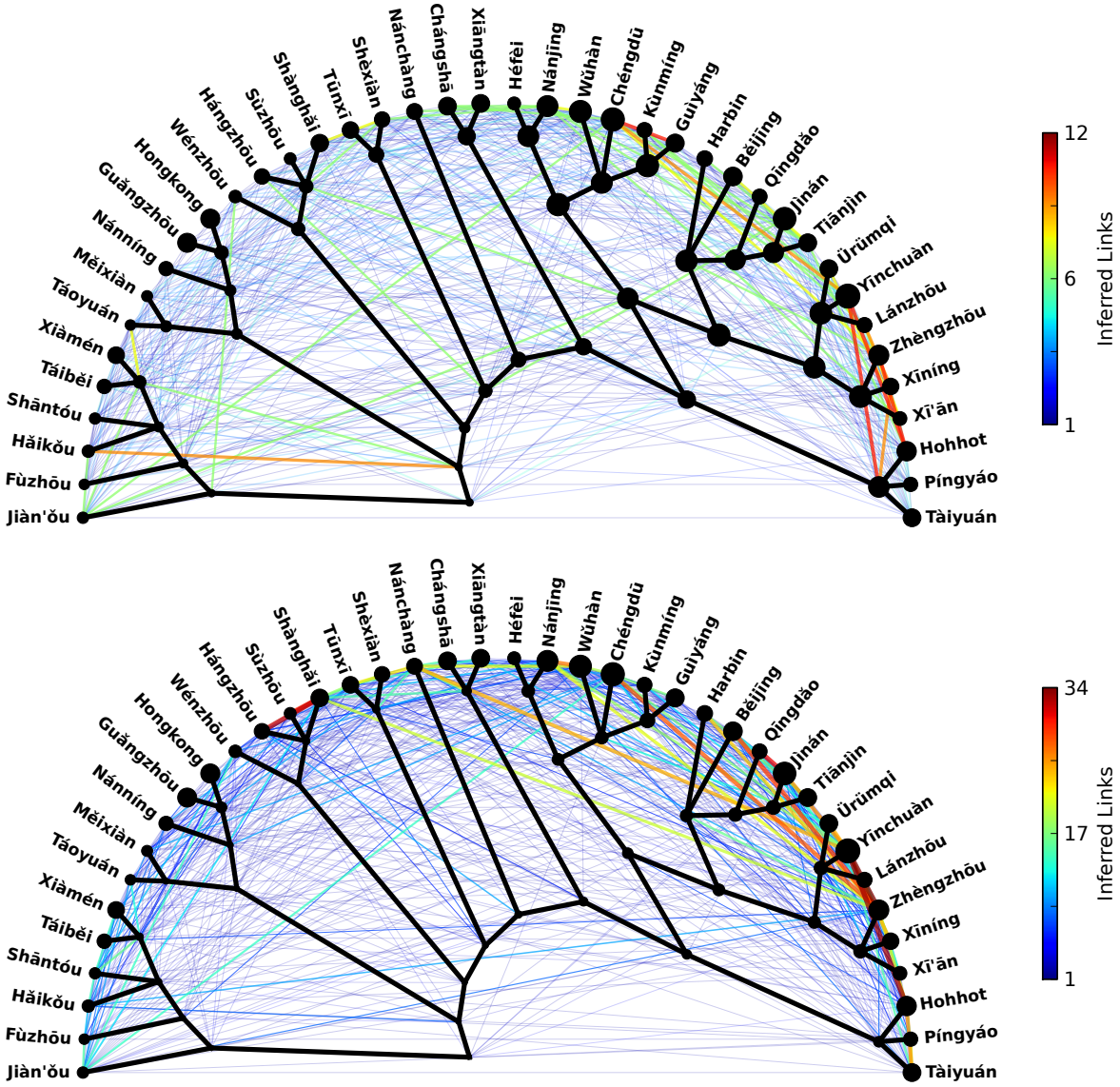
- Die Muster werden mit Hilfe von Gain-Loss-Mapping-Techniken aus der Biologie untersucht. Beim Gain-Loss-Mapping (Cohen u. a. 2010, Mirkin u. a. 2003) wird ein evolutionäres Szenario für die Entwicklung eines Kognatensets von der Wurzel des Referenzbaums bis zu den Blättern ermittelt. Ein Szenario erlaubt dabei vereinfachend nur die Ereignisse *Loss Event* (Schwund eines zuvor vorhandenen Kognatensets) und *Gain Event* (Ursprung eines Kognatensets).
- Wenn mehr als ein Gain-Event für ein Kognatenset postuliert wird, kann oft von einer Entlehnung ausgegangen werden.
- Das grundlegende Kriterium zur Auswahl der Gain-Loss-Modelle, die (normalerweise parsimoniebasiert) die besten Gain-Loss-Szenarien ermitteln, ist das „Vocabulary Size Distribution"-Kriterion, d.h. die Annahme, dass in den Ursprachen eines Referenzbaumes ähnlich viele Wörter vorhanden sein sollten wie in seinen Blättern.

| **Variety** | Spanish | French | Italian | English | German | Danish |
|---|---|---|---|---|---|---|
| **"to count"** | *contar* | *compter* | *contare* | *count* | *zählen* | *tælle* |
| **Latin** *computare* | 1 | 1 | 1 | 1 | 0 | 0 |
| **Proto-Germanic** **taljan-* | 0 | 0 | 0 | 0 | 1 | 1 |




Loss Event
Gain Event
太阳
日头
热头
阳婆
日
Nánchāng
Chángshā
Xiāngtán
Héfēi
Nánjīng
Wǔhàn
Chéngdū
Kūnmíng
Guìyáng
Harbin
Běijīng
Qīngdǎo
Jīnán
Tiānjīn
Ürümqi
Yínchuān
Lánzhōu
Zhèngzhōu
Xīníng
Xī'ān
Hohhot
Píngyáo
Tàiyuán
Shànghǎi
Sūzhōu
Hángzhōu
Wēnzhōu
Hongkong
Guǎngzhōu
Nánníng
Měixiàn
Tāoyuán
Xiàmén
Tāiběi
Shāntóu
Hǎikǒu
Fùzhōu
Jiàn'ōu
Tūnxī
Shèxiàn


Nánchāng
Chángshā
Xiāngtán
Héfēi
Nánjīng
Wǔhàn
Chéngdū
Kūnmíng
Guìyáng
Harbin
Běijīng
Qīngdǎo
Jīnán
Tiānjīn
Ürümqi
Yínchuān
Lánzhōu
Zhèngzhōu
Xīníng
Xī'ān
Hohhot
Píngyáo
Tàiyuán
Shànghǎi
Sūzhōu
Hángzhōu
Wēnzhōu
Hongkong
Guǎngzhōu
Nánníng
Měixiàn
Tāoyuán
Xiàmén
Tāiběi
Shāntóu
Hǎikǒu
Fùzhōu
Jiàn'ǒu
Tūnxī
Shèxiàn
12
6
1
Inferred Links




## 3.2 Goldstandard

- Derzeit liegt als Goldstandard nur ein Subset von Micheal Dunn's *Indo-European lexical cognacy database (IELex)* vor, das 40 Sprachen umfasst, und 186 bekannte Entlehnungen enthält, die als Kognaten „getarnt" sind.
- Die 186 bekannten Entlehnungen verteilen sich auf 100 Kognatensets.
- Der „Golstandard" ist relativ problematisch, da zu erwarten ist, dass in diesem noch viel mehr unentdeckte Entlehnungen enthalten sind, was zu durchgängig schlechten *precision*-Werten führt, die aber wohl eher der Unvollständigkeit des Goldstandards denn minderen Qualität der Methoden geschuldet sind.

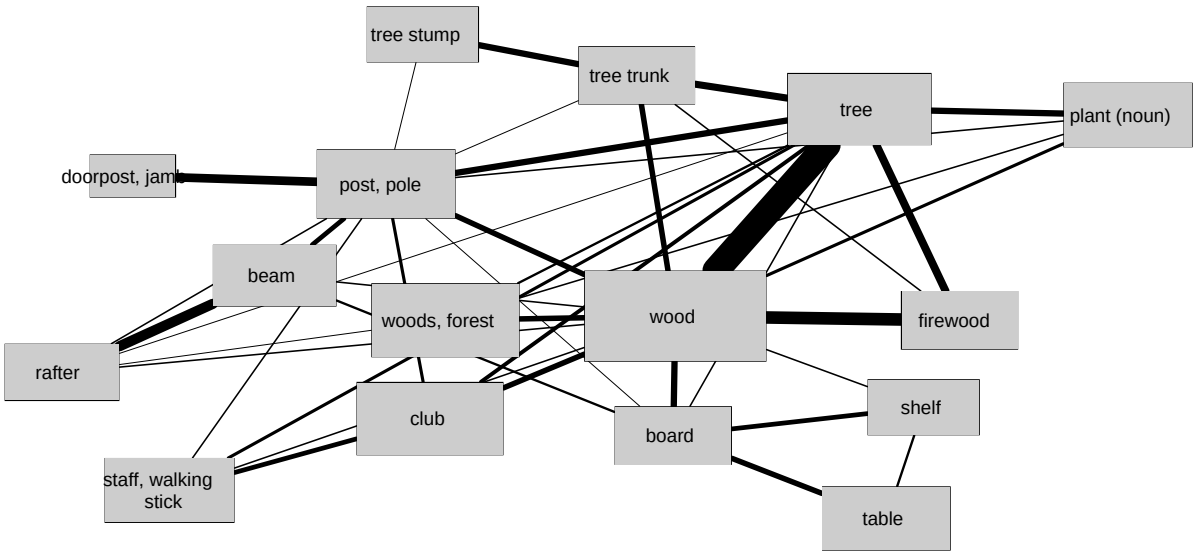
# 4 Quantitative Ansätze zum Umgang mit semantischem Wandel

## 4.1 Arbeitsstand

- Tendenzen des semantischen Wandels können sehr sinnvoll sein, um Heuristiken zur Kognatenerkennung zu unterstützen und über die Wortlistenebene hinaus nach Kognaten zu suchen.
- *Semantic Maps* (Cysouw 2010, Steiner u. a. 2011) bieten eine einfache Methode, um mit Hilfe polysemer Konzepte semantische Ähnlichkeit zwischen Konzepten zu definieren.
- Dabei wird, ausgehend von einer großen Wortliste, deren Glossen in viele Sprachen übersetzt wurden, gezählt, wie oft in jeweils einer Sprache ein Konzept durch die gleiche Form ausgedrückt wird. Es wird also die Anzahl von Polysemien in einem Datensatz gezählt.
- Eine Erweiterung derartiger Analysen um neue Methoden der Netzwerkanalyse (Erstellung gewichteter Polysemienetzwerke) ermöglicht es, die semantische Ähnlichkeit über die item-basierte Aufzählung von Polysemien hinauszuheben.
- Algorithmen zur Entdeckung von Community-Strukturen (Girvan und Newman 2002, Newman 2006) haben sich als relativ nützlich dafür erwiesen, semantische Beziehungen, die bedeutungsvoll zu sein scheinen, zu entdecken (List u. a. 2013).

## 4.2 Datensätze

- Es gibt keinen Goldstandard, jedoch eine Datenbank, die sowohl zur manuellen als auch zur automatischen Erforschung cross-linguistischer Polysemien verwendet werden kann.
- CLiPs („Database of Cross-Linguistic Polysemies”, `http://www.quanthistling.info/clips/`) listet automatisch entdeckte Polysemien für 200 Wortlisten, die aus der *IDS – The Intercontinental Dictionary Series*, *World Loanword Database*, und Daten der Logos Group entnommen und vereinheitlicht wurden.
- Jede Wortliste liefert Übersetzungen für bis zu 1300 Konzepte.
- Die Datenbank kann zur einfachen Abfrage möglicher Synonymbeziehungen verwendet werden.
- Eine Netzwerkrepräsentation der Daten mit erweiterter Community-Struktur-Analyse könnte es ermöglichen, bestimmte „Nähewerte” für Konzeptpaare zu bestimmen. Basierend auf solchen Werten könnten Kognatenerkennungsmethoden erweitert und bewusst zur Suche nach Kognaten mit Reflexen unterschiedlicher Bedeutung verwendet werden.




## Literatur

Cohen, O., H. Ashkenazy, F. Belinky, D. Huchon und T. Pupko (2010). "GLOOME: gain loss mapping engine". In: *Bioinformatics* 26.22, 2914–2915.

Cysouw, M. (2010). "Drawing Networks from Recurrent Polysemies". Comment on 'Polysemous Qualities and Universal Networks' by Loïc-Michel Perrin (2010). In: *Linguistic Discovery* 8.1, 281–285.

Downey, S. S., B. Hallmark, M. P. Cox, P. Norquest und S. Lansing (2008). "Computational Feature-Sensitive Reconstruction of Language Relationships: Developing the ALINE Distance for Comparative Historical Linguistic Reconstruction". In: *Journal of Quantitative Linguistics* 15.4, 340–369.

Girvan, M. und M. E. Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12, 7821–7826.

Haspelmath, M. und U. Tadmor, Hrsg. (2009). *World Loanword Database*. URL: `http://wold.livingsources.org`.

List, J.-M., S. Nelson-Sathi, W. Martin und H. Geisler (forthcoming). "Using phylogenetic networks to model Chinese dialect history". In: *Language Dynamics and Change*.

List, J.-M. (2012a). "LexStat. Automatic Detection of Cognates in Multilingual Wordlists". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. (Avignon, France, 23.–24. Apr. 2012). Association for Computational Linguistics, 117–125.

– (2012b). "SCA. Phonetic alignment based on sound classes". In: *New directions in logic, language, and computation*. Hrsg. von M. Slavkovik und D. Lassiter. LNCS 7415. Berlin und Heidelberg: Springer, 32–51.

List, J.-M. und S. Moran (forthcoming). "An open source toolkit for quantitative historical linguistics". In: *Proceedings of the ACL 2013 System Demonstrations*. (Sofia, Bulgaria, 4.–9. Aug. 2013). Association for Computational Linguistics.

List, J.-M., A. Terhalle und M. Urban (2013). “Using network approaches to enhance the analysis of cross-linguistic polysemies”. In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) - Short Papers*. (Potsdam, Germany, 19.–22. März 2013). Association for Computational Linguistics, 347–353.

Logos Group, Hrsg. (2008). *Logos Dictionary*. Url: `http://www.logosdictionary.org/index.php`.

Mirkin, B. G., T. I. Fenner, M. Y. Galperin und E. V. Koonin (2003). “Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes”. In: *BMC Evolutionary Biology* 3, 2.

Nelson-Sathi, S., J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin und T. Dagan (2011). “Networks uncover hidden lexical borrowing in Indo-European language evolution”. In: *Proc B* 278.1713, 1794–1803.

Newman, M. E. J. (2006). “Finding community structure in networks using the eigenvectors of matrices”. In: *Physical Review E* 74 (3). ISSN: 1539-3755, ISBN: 1539-3755, 1–19.

Prokić, J., M. Wieling und J. Nerbonne (2009). “Multiple sequence alignments in linguistics”. In: “Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education” (Athens, Greece, 30. März 2009). Stroudsburg, PA: Association for Computational Linguistics, 18–25. acm: `1642052`.

Shirō, H. (1973). “Japanese dialects”. In: *Diachronic, areal and typological linguistics*. Hrsg. von H. M. Hoenigswald und R. H. Langacre. Current Trends in Linguistics 11. The Hague und Paris: Mouton, 368–400.

Steiner, L., P. F. Stadler und M. Cysouw (2011). “A pipeline for computational historical linguistics”. In: *Language Dynamics and Change* 1.1, 89–127.

Turchin, P., I. Peiros und M. Gell-Mann (2010). “Analyzing genetic connections between languages by matching consonant classes”. In: *Journal of Language Relationship* 3, 117–126.

Wang, F. (2006). *Comparison of languages in contact. The distillation method and the case of Bai*. Taipei: Institute of Linguistics Academia Sinica.

## Quellen

ABVD S. J. Greenhill, R. Blust und R. D. Gray, Hrsg. (2008/). *The Austronesian Basic Vocabulary Database*. URL: `http://language.psy.auckland.ac.nz/austronesian/`.

BDS B. Allen (2007). *Bai Dialect Survey*. SIL International. PDF: `http://www.sil.org/silesr/2007/silesr2007-012.pdf`.

DERKSEN R. Derksen, Hrsg. (2008). *Etymological dictionary of the Slavic inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 4. Leiden und Boston: Brill.

GLD G. Starostin und P. Krylov, Hrsg. (2011). *The Global Lexicostatistical Database. Compiling, clarifying, connecting basic vocabulary around the world: From free-form to tree-*

*form*. URL: http://starling.rinet.ru/new100/main.htm.

IDS M. R. Key und B. Comrie, Hrsg. (2007). *IDS – The Intercontinental Dictionary Series*. URL: http://lingweb.eva.mpg.de/ids/.

IELex M. Dunn, Hrsg. (2012). *Indo-European lexical cognacy database (IELex)*. URL: http://ielex.mpi.nl/.

LOE C. Renfrew und P. Heggarty (2009). *Languages and Origins in Europe*. URL: http://www.languagesandpeoples.com/. (Besucht am 12. 06. 2012).

MAND G. de Schutter, B. van den Berg, T. Goeman und T. de Jong, Hrsg. (2007). *Mand. Morfologische Atlas van de Nederlandse Dialecten*. Meertens Instituut. URL: http://www.meertens.knaw.nl/mand/database/. (Besucht am 12. 06. 2012).

NORDAVINDEN NORDAVINDEN. *Nordavinden og sola. En norsk dialektprøvedatabase på nettet* [The North Wind and the Sun. A Norwegian dialect database on the web]. Aufnahmen und Transkriptionen von J. Almberg. Technische Implementierung von K. Skarbø. URL: http://www.ling.hf.ntnu.no/nos/.

SAL P. Heggarty (2006). *Sounds of the Andean languages*. URL: http://www.quechua.org.uk/. (Besucht am 12. 06. 2012).

TPPSR L. Gauchat, J. Jeanjaquet und E. Tappolet, Hrsg. (1925). *Tableaux phonétiques des patois suisses romands. Relevés comparatifs d'environ 500 mots dans 62 patois-types. Publiés avec introduction, notes, carte et répertoires*. Neuchâtel: Attinger.

YINKU Hóu Jīng 侯精, Hrsg. (2004). *Xiàndài Hànyǔ fāngyán yīnkù* 现代汉语方言音库 [Phonological database of Chinese dialects]. Shanghai: Shànghǎi Jiàoyù.