André Müller

Künstliche Sprachen im Rahmen einer computergestützten lexikostatistischen Untersuchung[1]

Gliederung

1 Vorüberlegungen

Um die Ähnlichkeit zweier Sprachen feststellen und vergleichen zu können, existieren bislang mehrere unterschiedliche Ansätze und ausgebaute Projekte. Während viele Jahre zuvor noch statistisch schwer nachweisbare Maßstäbe bzw. schlecht reproduzierbare Methoden angelegt und genutzt wurden, wie etwa die gegenseitige Verständlichkeit (Interintelligibilität) oder die bloße subjektive Einschätzung eines einzelnen Sprachforschers, wurden mit der Zeit zunehmend objektivere Methoden gefordert. Ansätze dazu gab es bereits im frühen 19. Jahrhundert. In den 50er und 60er Jahren des 20. Jahrhunderts entwickelte der Linguist Morris Swadesh ein Verfahren zum Sprachvergleich anhand von normierten Wortlisten. Durch die Entwicklung des Computers konnten die statistischen Berechnungen in einem Bruchteil der Zeit abgewickelt werden, was die Arbeit erheblich erleichterte und die Möglichkeit erlaubte, mehr Daten zu vergleichen, als von Hand berechnet werden könnten. Auch die Zugänglichkeit zu Sprachdaten hat sich in dieser Zeit erhöht.

Seit 2007 beschäftigt sich ein internationales linguistisches Projekt, ASJP (*Automated Similarity Judgment Program*) mit dem Thema und versucht, durch automatisierte Berechnungen die phonetisch-lexikalischen Distanzen aller Sprachen der Welt untereinander zu vergleichen, um so objektive Aussagen über die phonetisch-lexikalischen Ähnlichkeiten in einer bestimmten Stichprobe des Vokabulars zweier oder mehrerer Sprachen zu treffen und andere Daten und Analysen extrapolieren zu lassen. Ähnliche Projekte, allerdings in leicht anderer Ausführung und kleinerem Umfang, sind auch von Oswalt (1970), Villemin (1983), Nakhleh et al. (2005) und anderen umgesetzt worden, zum Teil für einzelne Sprachfamilien. Die Darstellung der Ähnlichkeiten erfolgt in numerischen Werten für jedes Sprachenpaar, kann aber als Baumdiagramm grafisch dargestellt werden. Die entstehenden Baumdiagramme ähneln vom Aufbau her stark den genetischen Bäumen, die in der historischen Sprachwissenschaft oft für Sprachverwandtschaften aufgestellt werden. Die vom ASJP-

[1] Mein Dank gilt Sebastian Sauppe, Sabine Fiedler, Helen Geyer und Søren Wichmann für wertvolle Korrekturen, Anmerkungen und Hinweise zu dieser Arbeit.

Projekt erzeugten Bäume stellen jedoch keine direkte Verwandtschaft dar, sondern nur die Ähnlichkeit der Sprachen untereinander.

Neben einigen tausend natürlichen, lebenden Sprachen, Pidgin- und Kreolsprachen sowie einigen Dutzend ausgestorbenen und rekonstruierten Sprachen und Sprachstufen sind in den Datensätzen auch Sprachdaten zu den wichtigsten konstruierten Sprachen enthalten.[2] Die vorliegende Arbeit konzentriert sich nach einer Einführung in die Methoden des ASJP-Projekts auf die Ergebnisse des Projekts in Bezug auf diese konstruierten Sprachen im Allgemeinen und im Einzelnen.

2 Das ASJP-Projekt

ASJP steht für *Automated Similarity Judgment Program*[3] und ist der Name eines 2007 ins Leben gerufenen linguistisch-statistischen Forschungsprojekts mit Mitarbeitern aus Dänemark, Deutschland, Russland, den Niederlanden, Großbritannien, den USA und anderen Ländern. Grundlage des Projekts ist die von Morris Swadesh (1955) entwickelte Liste von grundlegendem Vokabular, das nach Untersuchungen, Vergleichen und Einschätzungen als besonders stabil, d.h. resistent gegen Entlehnung, Ersetzung und Bedeutungsverschiebung, gilt. Die Wörter auf der so genannten *Swadesh-Liste* sollen ebenfalls möglichst kulturell neutral sein, wenig Synonyme haben und auf grundlegende Konzepte referieren. Swadesh entwickelte zwei Listen: eine mit 100 Wörtern und eine größere mit 200 Wörtern, die je nach Bedarf 100 bzw. 200 möglichst stabile, frequente und im Alltag saliente Wörter enthalten. An der wirklichen Stabilität und der einzelnen Lexeme sowie der Relevanz einer solchen Liste wurde in den vergangenen 55 Jahren viel Kritik geübt, doch waren die Swadesh-Listen seitdem ein wichtiges Hilfsmittel für die Berechnungen in der Lexikostatistik und später der Glottochronologie[4].

Beim ASJP-Projekt wurde anfangs die Swadesh-100-Liste als Grundlage zur Berechnung der phonetisch-lexikalischen Distanz benutzt, später aber durch eine kürzere Liste mit 40 Wörtern ersetzt, als durch statistische Verfahren gezeigt werden konnte, dass sich ab einer Anzahl von 40 Wörtern aufwärts die Ergebnisse der Ähnlichkeitsberechnungen nicht mehr signifikant ändern (Holman et al. 2008a, 8ff.). Die Nutzung von nur 40 Wörtern hat den Vorteil, eine schnellere Sammlung, Kodierung und Berechnung der Daten zuzulassen, des Weiteren erhält man bei schlecht dokumentierten Sprachen weniger fehlende Listeneinträge, da die stabilsten 40 Wörter ausgewählt wurden. Die 40-Wort-Liste enthält 3 Pronomen, 2 Numeralia, 5 Verben, 2 Adjektive sowie 28 Nomen,[5] die vor allem Tiere, Körperteile und Naturerscheinungen beschreiben. Es folgt eine alphabetische Auflistung dieser 40 Wörter:

[2] Für die Darstellung aller Sprachen der Welt im Sprachbaum werden Pidgin-, Kreolsprachen, künstliche, rekonstruierte und vor längerer Zeit ausgestorbene Sprachen sowie ältere Sprachstufen oftmals nicht betrachtet und aus den Berechnungen exkludiert, da normalerweise nur die heute lebenden, durch Sprachevolution natürlich entstandenen Sprachen betrachtet werden.

[3] Die offizielle Homepage des ASJP-Projekts ist http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm.

[4] Unter der Annahme, dass sich der Wortschatz einer Sprache im Laufe der Zeit mit einer konstanten Rate verändert, versucht man in der Glottochronologie die Zeit des Aufspaltens einer Sprache in zwei Tochtersprachen zu berechnen. Diese Berechnungen geben dann Aufschluss über das ungefähre Alter einer Sprachfamilie bzw. eines Zweiges oder einer Einzelsprache. Auch das ist eines der Forschungsschwerpunkte im ASJP-Projekt.

[5] Die angegebenen Wortarten gelten fürs Deutsche und Englische, manche Lexeme weichen jedoch in manchen Sprachen davon ab. In vielen Sprachen sind „voll (sein)“ und „neu (sein)“ beispielsweise Verben.

blood	eye	Knee	nose	sun
bone	fire	Leaf	one	tongue
breast	fish	Liver	path	tooth
to come	full	Louse	person	tree
to die	hand	mountain	to see	two
dog	hear	Name	skin	water
drink	horn	New	star	we
ear	I	Night	stone	you

Um eine Berechnung der Ähnlichkeit zu ermöglichen, muss für jede Sprache die gleiche Orthographie angewendet werden, und da die Phonetik bei dieser Berechnung eine große Rolle spielt, muss es sich dabei um ein phonetisches Alphabet handeln. Da viele der genutzten Programmiersprachen und verwendeten Programme zu Anfang noch kein Unicode unterstützten und die Verarbeitung von Daten im Internationalen Phonetischen Alphabet (IPA) so unmöglich machten, musste auf die auf der englischen Tastatur vorhandenen Zeichen zurückgegriffen werden. Der zu diesem Zweck entwickelte *ASJP-Code* (vergleichbar etwa mit dem verbreiteten SAMPA-System[6]) stellt also eine phonetische Umschrift auf Basis der ASCII-Zeichen dar; die Fülle an verschiedenen phonetischen Symbolen wurde hier auf 41 Symbole reduziert, von denen einige mehrere unterschiedliche Phoneme darstellen können. Später, als auch die Nutzung von IPA zur Kodierung der Daten möglich war, zeigten weitere Untersuchungen (Holman et al. 2008b), dass der Unterschied zwischen IPA und ASJP-Code so geringen Einfluss auf die Ergebnisse hatte (etwa 1 %), dass diese Verbesserung vernachlässigbar war und darauf verzichtet wurde, die Daten per Hand erneut in IPA zu kodieren.

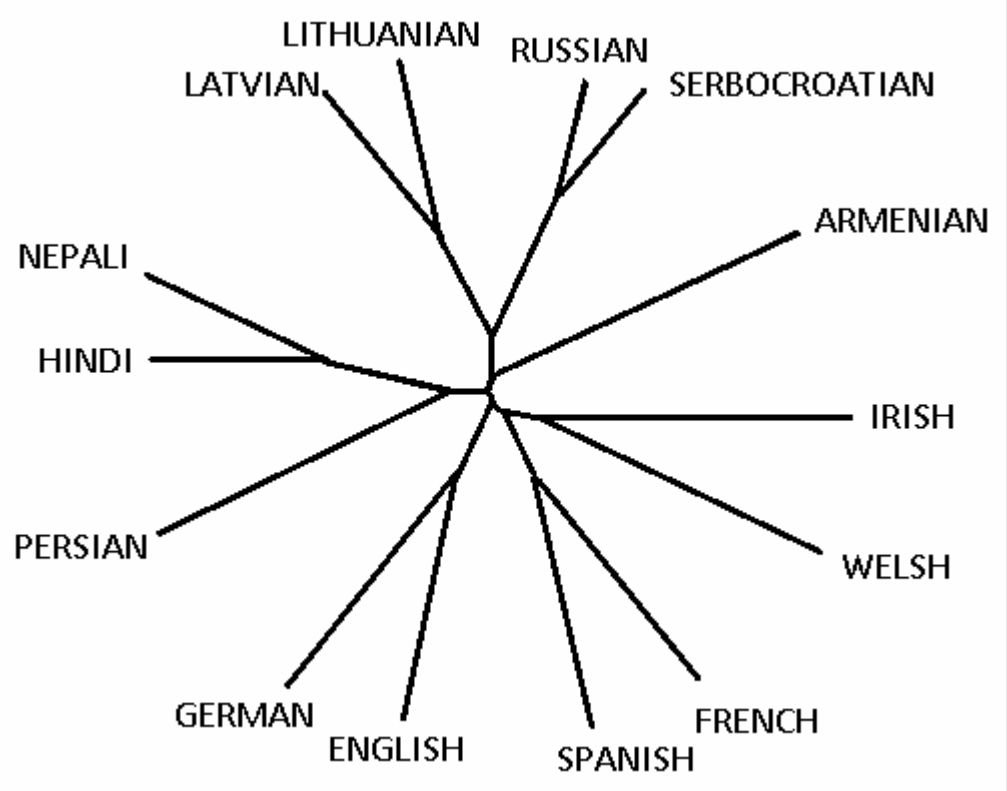
Komplizierter als die Datenkodierung erweisen sich die mathematischen Grundlagen für die Berechnung der intersprachlichen Ähnlichkeit: diese Grundlage bildete der von Levenshtein (1966) entwickelte Algorithmus, mit dem sich die Mindestzahl an erforderlichen Änderungen zwischen zwei Lexemen berechnen lässt. Die „Kosten“ für die Ersetzung, Hinzufügung oder Löschung eines einzelnen Zeichens (d.h. Phonems) betragen jeweils 1; die Summe aller erforderlichen Änderungen ist das Maß der Unähnlichkeit zwischen zwei Lexemen und wird als *Edit Distance* oder als *Levenshtein-Distanz* bezeichnet. Die Levenshtein-Distanz zwischen dem spanischen Wort *gato* (ASJP-Code: <gato>) und dem Esperanto-Wort *kato* (<kato>) beträgt in diesem Fall also 1 (Ersetzung von *g* durch *k*), während das englische *tongue* (<toN>) sich vom deutschen Wort *Zunge* (<cuN3>) durch 3 Änderungen unterscheidet: Ersetzung von <t> und <o> durch <c> und <u>, Hinzufügen von <3>. Die Richtung (*tongue* → *Zunge* oder *Zunge* → *Hund*) spielt dabei keine Rolle. Historisch verwandte Wörter (sog. *Kognate*) erzielen so normalerweise viel geringere Levenshtein-Distanzen als unverwandte Wörter (wie z.B. engl. *mountain* <maunt3n> und chin. *shan* <San> mit einer Distanz von 5). Um den Einfluss unterschiedlicher Phoneminventare und Wortlängen sowie auch den zum Teil daraus resultierenden Einfluss zufälliger Ähnlichkeiten auszugleichen, sind weitere mathematische Methoden nötig. Genaue Erklärungen dazu können Wichmann et al. (2010a) entnommen werden. Um den Einfluss der Wortlänge auszugleichen, wird pro Wortpaar die Levenshtein-Distanz (LD) durch die Länge des längeren der beiden Wörter dividiert, das Resultat ist die LDN (*Levenshtein Distance Normalized*). Bessere Resultate können erzielt werden, indem die LDN zusätzlich durch die durchschnittliche LDN aller möglichen Wortpaare, die nicht auf dasselbe Konzept referieren, in den beiden Sprachen, geteilt wird. So können zufällige

[6] Offizielle Homepage: http://www.phon.ucl.ac.uk/home/sampa/home.htm.

Ähnlichkeiten, die von ähnlichen Phoneminventaren verursacht werden, minimiert werden.[7] Das Ergebnis der zweiten Modifikation ist die LDND (*Levenshtein Distance Normalized Divided*). Die phonetisch-lexikalische Gesamtdistanz zweier Sprachen ergibt sich nun aus der Summe der LDND für alle Wortpaare.

Die für das ASJP-Projekt verwendeten Computerprogramme erstellen danach aus den verschiedenen modifizierten Levenshtein-Distanzen eine Distanzmatrix, d.h. eine Tabelle, deren Spalten- und Zeilenzahl der Anzahl der im Datensatz enthaltenen Sprachen entspricht. Da die Distanz in beiden Richtungen identisch ist (A-B = B-A) und die Distanz zwischen ein und derselben Sprache per Definition null ist, enthält die Tabelle allerdings „nur“ $\frac{n \times (n-1)}{2}$ verschiedene gefüllte Felder, wobei *n* die Anzahl der Sprachen im Datensatz ist. Der derzeitige Datensatz (Wichmann et al. 2010b) enthält bei 4762 Sprachen [8] demnach 11.335.941 zu vergleichende Sprachpaare — ein unmöglich per Hand zu bewerkstelligender Rechenaufwand.

Mithilfe von Computerprogrammen wie *MEGA* (Wichmann et al. 2010a, 10) oder *SplitsTree* (Brown et al. 2008, 7) und dem *Neighbor-Joining*-Algorithmus können die Daten aus der Distanzmatrix automatisch als Baumdiagramm dargestellt werden. Die erzeugten Baumdiagramme sind ungewurzelt, was bedeutet, dass keine Verzweigstelle entsteht, die als Mutterknoten (und damit eine hypothetische „Ursprache“) des gesamten Baumes betrachtet werden kann. Ein solcher Baum einiger der indo-europäischen Sprachen ist hier dargestellt:



[7] Bei Sprachen mit jeweils sehr kleinem (und oft ähnlichem) Phoneminventar ist die Wahrscheinlichkeit, dass sich Wörter zufällig ähneln größer als bei Sprachen mit großen Phoneminventaren.

[8] Stand: 13.08.2010. Eine Liste der bereits im Datensatz enthaltenen Sprachen ist einsehbar unter: http://email.eva.mpg.de/~wichmann/LanguagesOfTheWorld.pdf — das entspricht bereits über 68 % der laut dem Ethnologue (Lewis 2009) auf der Welt gesprochenen 6909 Sprachen.

In dieser Darstellung (Quelle: Brown et al. 2008, 15) verzweigen sich die einzelnen Sprachen teils direkt, teils indirekt aus der Mitte heraus. Eine alternative Darstellungsmethode reiht den Baum vertikal auf, was vor allem bei großen Datenmengen übersichtlicher ist. Auch in dieser Darstellung (siehe Müller et al. 2010) ist der Baum ungewurzelt. Die Verzweigungen sind im Allgemeinen binär, da die jeweiligen Distanzen zwischen drei Sprachen durch den Algorithmus selten einen exakt identischen Wert haben. In beiden Fällen ist der (in der zweiten Darstellungsmethode horizontale) Abstand zweier Knoten (Sprachen A und B) zu ihrem Mutterknoten (Proto-A-B) proportional zur phonetisch-lexikalischen Distanz dieser Sprachen — in der Glottochronologie entspricht das ebenfalls der angenommenen Zeit seit der Aufspaltung von Proto-A-B in die Tochtersprachen A und B, ein nicht unumstrittenes Verfahren. Vergleiche mit von Expertenhand erstellten Verwandtschaftsbäumen einzelner Sprachfamilien (Brown et al. 2008, 7ff.) zeigt eine relativ hohe Übereinstimmung der ASJP-Bäume mit den jeweiligen Expertenklassifikationen.

3 Künstliche Sprachen und ASJP

Seit Gaston Moch 1897 die Unterscheidung zwischen *a priori*- und *a posteriori*-Kunstsprachen prägte, gilt sie als vielleicht **das** Hauptkriterium zur Einordnung künstlicher Sprachen. Viele Kunstsprachen benutzen ein hybrides System mit beiden Eigenschaften (Schmidt-Radefeldt 1998, 680f.). Die für diese Arbeit wesentlichen Aspekte der beiden Systeme lassen sich bei Gledhill (1998, 19) finden. Relevant ist hier für den Typ a priori: „A schematic word is a combination of semantic signifiers and is unlikely to resemble any recognizable language“ und für den Typ a posteriori: „a naturalistic language project based on the common features of a series of existing natural languages. […] Naturalistic words are designed to be maximally recognizable“ (entnommen aus Gledhill 1998, 19). Obgleich sowohl Gledhill als auch Schmidt-Radefelt in dieser Unterscheidung beim Typ a priori eher von einer Symbol-Konzept-Zuweisung sprechen, so z.B. bei Kunstsprachen wie dem Solresol, bei der versucht wurde, die Konzepte der Welt phonetisch bzw. graphisch zu repräsentieren, konzentriere ich mich hier auf die rein lexikalische Unterscheidung.

A priori bedeutet hier also, dass die Lexeme der Sprache in keinem geplanten oder bewussten Bezug zu einer natürlichen Sprache stehen, das Lexikon kann also als „frei erfunden“ betrachtet werden. Beispiele dafür sind Klingonisch, Na'vi[9], sowie Tolkiens elbische Sprachen Sindarin und Quenya. Es ist der Untersuchungsgegenstand dieser Arbeit, ob lexikostatistisch nicht doch signifikante Ähnlichkeiten zu anderen Sprachen erkennbar sind, die eventuell auf unbewussten oder bewussten Einfluss der vom Erfinder gesprochenen natürlichen Sprachen rückschließen lassen könnten. A posteriori bezeichnet hier hingegen künstliche Sprachen, deren Lexeme eine oder mehrere natürliche Sprachen zum Vorbild haben. Die hier angesprochenen Vertreter dieses Typs sind Esperanto, Ido, Interlingua, Lojban, Slovio, Toki Pona und Volapük.

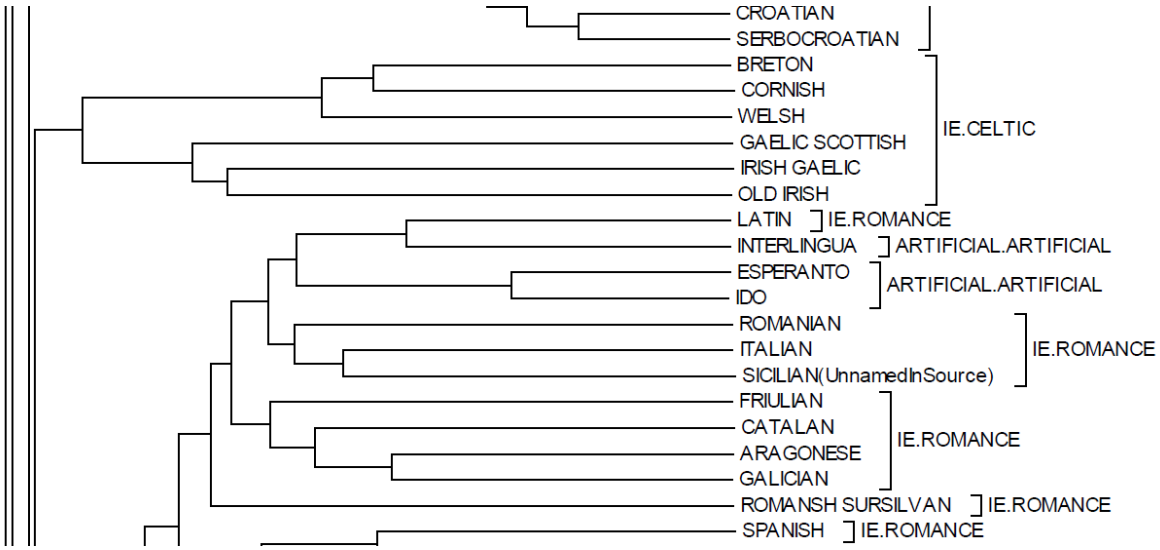
Des Weiteren gleichen laut Back (1980, 267) Plansprachen bei der Betrachtung ihrer „Etymologien“ natürlichen Sprachen, deren kompletter Wortschatz aus Lehnwörtern aufgebaut ist (im fiktiven Fall). Ein Vergleich zwischen Plan- und Pidgin- bzw. Kreolsprachen liegt also nahe und wurde an anderer Stelle schon vorgenommen.[10]

[9] Na'vi ist eine vom Linguisten Paul Frommer entwickelte künstliche Sprache für die Bewohner des Mondes Pandora im Film *Avatar*. Sie ist aufgrund zu geringen Vokabulars nicht in ASJP enthalten.

[10] Siehe hierzu z.B. Liu Haitao (2001): Pidgins, Creoles and planned languages. – Linguistic development under special conditions. In: Klaus Schubert (Hrsg.): Planned languages: From Concept to Reality. Brüssel: Hogeschool voor Wetenschap en Kunst, 121-177.

3.1 Esperanto

Laut Janton (1993, 51) basiert die Lexik des Esperanto zu 75 % auf romanischen und 20 % auf germanischen Wurzeln. Die restlichen 5 % entfallen dabei vor allem auf griechische und einige wenige slawische Wortwurzeln. Von den untersuchten 40 Wörtern der ASJP-Liste sind 32 romanisch-basiert, 7 germanisch-basiert und 1 aus dem Griechischen kommend (*hepato*, Leber). Bei einigen der Wörter ist die Zuordnung nicht leicht, da es nahe Kognate in beiden Sprachfamilien gibt (z.B. *nazo*, Nase). Die Liste spiegelt jedoch das von Janton ermittelte Verhältnis wider. Der folgende Ausschnitt aus dem ASJP-Baum mit der kompletten Liste aller Sprachen zeigt die Position Esperantos im Baum:



Die dem Esperanto ähnlichste Sprache in phonetisch-lexikalischer Sicht ist demnach das Ido (s. unten), diese ähneln weiterhin den beiden Sprachen Interlingua und Latein, deren weiter links liegende Gabelung eine im Vergleich zu Esperanto–Ido größere Differenz und damit weiter zurückliegende (natürlich nur fiktive) Aufspaltung der beiden Sprachen impliziert. Die diesen vier Sprachen nächstähnliche Gruppe besteht aus den Sprachen Rumänisch, Italienisch und Sizilianisch. Es schließen sich weitere romanische Sprachen und romanisch-basierte Kreolsprachen an.

Der vorwiegend romanische Teil des Esperanto-Lexikons schlägt sich also auch in den Resultaten des ASJP-Klassifikation nieder: Esperanto scheint für die automatische Einschätzung der Software eine romanische Sprache zu sein.

3.2 Ido

Die Veränderungen, die die Sprache Ido 1907 aus dem Esperanto entstehen ließen, betrafen die Morphologie, Syntax, Graphematik, Phonologie, aber auch großzügig die Lexik (vgl. Anton 2001). Der größtenteils romanische Charakter ist geblieben und wurde teilweise verstärkt. Die Wörter innerhalb der 40-Wort-Liste, deren Wurzel nicht nur verändert, sondern quasi ausgetauscht wurden, betreffen: Haut (*haŭto* > *pelo*) und du (*vi* > *tu*), wobei ersteres ein germanisches (< dt. *Haut*) durch ein romanisches (vgl. span. *piel*, ital. *pelle*) Wort ersetzt und ersteres eine innerromanische Ersetzung (vgl. franz. *vous* und franz. *tu*) betrifft.

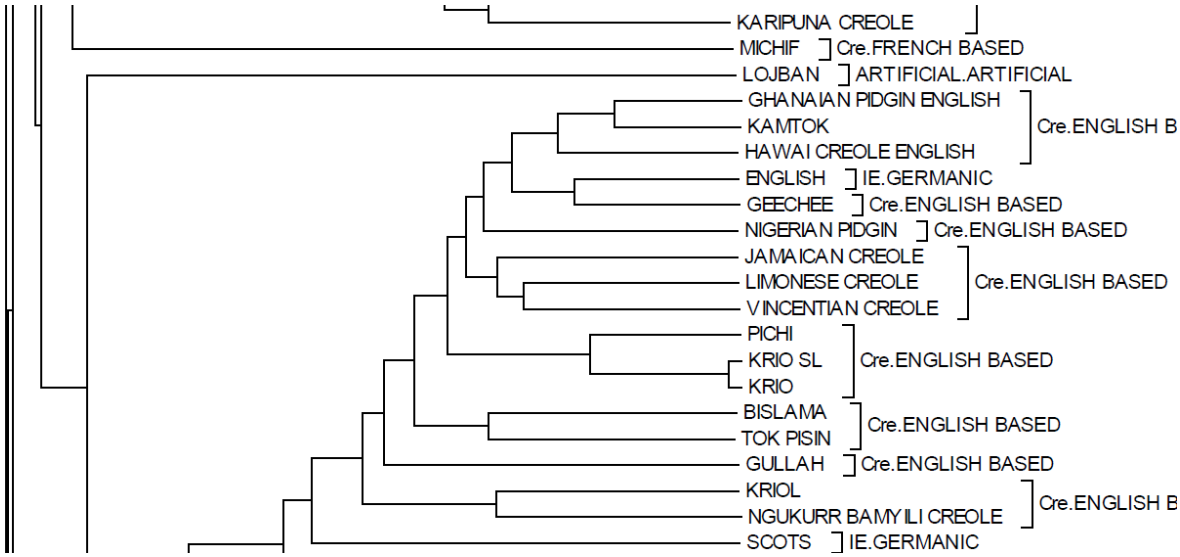
Die Sprache könnte daher unter Umständen als geringfügig „romanischer“ gelten als das Esperanto, doch verbindet die große Ähnlichkeit der beiden Sprachen Esperanto und Ido im Sprachbaum.

3.3 Interlingua

Der Wortschatz des 1951 von Alexander Gode entwickelten Interlingua basiert laut Schmidt-Radefeldt (1998, 684) vorrangig auf neulateinischen Wurzeln und auch bei näherer Betrachtung der 40-Wort-Liste sind alle darauf enthaltenen Wörter ihrem Ursprung nach zumindest romanisch. Die Nähe zum Lateinischen wird auch im ASJP-Baum deutlich, da die beiden Sprachen dort aus einer Verzweigung hervorgehen. Die phonetisch-lexikalische Differenz zwischen Interlingua und Lateinisch ist hier leicht größer als bei Esperanto und Ido und etwas geringer als die zwischen beispielsweise den natürlichen Sprachen Norwegisch (Bokmål) und Schwedisch.

3.4 Lojban

Die Lexik der „(prädikaten)logischen Sprache“ Lojban basiert nicht auf einer oder einigen wenigen Sprachfamilien, sondern wurde durch ein spezielles Computerverfahren aus sechs sprecherreichen Sprachen der Erde (Mandarin-Chinesisch, Englisch, Hindi, Spanisch, Russisch, Arabisch) erzeugt. Wichtige Faktoren waren dabei unter anderem die Vermeidung von Ambiguität, ein bestimmtes Silbenmuster pro Wortart und eine gewichtete möglichst große phonetische Ähnlichkeit zu den sechs Kontrollsprachen (Nicholas & Cowan 2003, II.4.17 und Fellmann 2001, 130f.). Als Resultat erinnern die Wörter des Lojban nur noch sehr entfernt an die jeweiligen Kognate aus dem Englischen oder anderen Sprachen. Da vier der sechs Kontrollsprachen zur indoeuropäischen Sprachfamilie (darin jedoch zu unterschiedlichen Zweigen) gehören, ist eine leichte Tendenz zu dieser Familie auch im Baum zu erwarten.

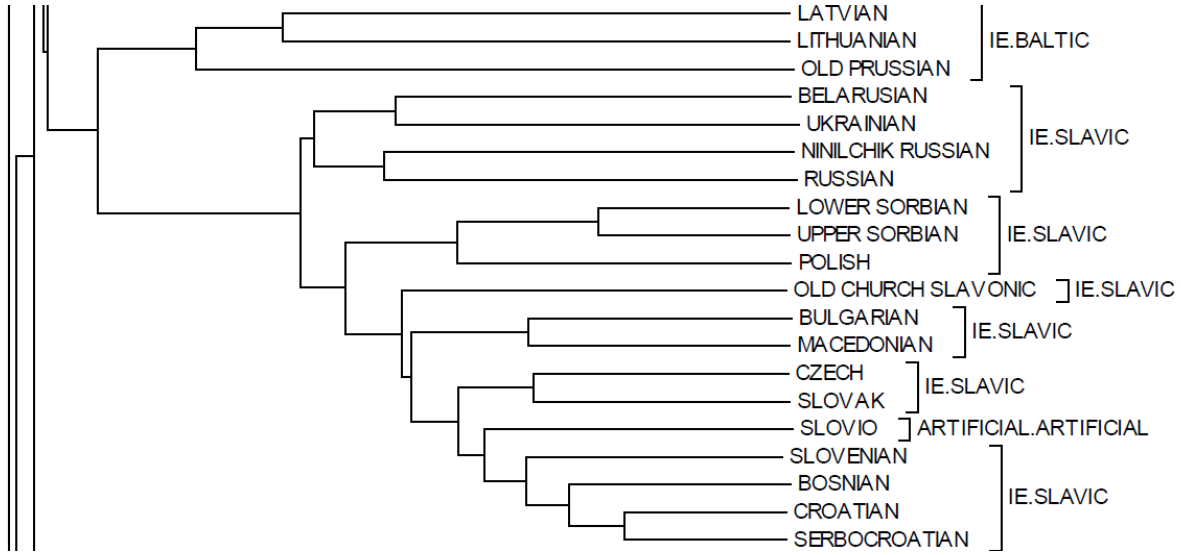


Wie in der Abbildung zu sehen, ist das tatsächlich der Fall. In der Abbildung nur schlecht erkennbar, bildet Lojban einen Schwesterknoten von einer ganzen Reihe von Sprachen, die größtenteils englisch-basierten Kreolsprachen angehören, unter ihnen auch das Englische selbst. Die große Entfernung des gemeinsamen Zweigpunktes und die große Anzahl der Schwestersprachen zeigen, dass die Ähnlichkeit zum Englischen bzw. den damit verwandten

Kreolsprachen zwar vorhanden, aber doch sehr gering ist. Die jeweilige Differenz zu den anderen Kontrollsprachen ist anscheinend viel größer, da sich Lojban andernfalls bei den sinotibetischen (Chinesisch) oder der afroasiatischen Sprachfamilie (Arabisch) wiedergefunden hätte. Bei näherer Betrachtung der Wortliste fällt auf, dass bei den Inhaltswörtern, die im Lojban die Silbenstruktur CVCCV oder CCVCV besitzen, oft die letzte Silbe dem chinesischen Kontrollwort ähnelt, die erste Silbe häufig dem Englischen, Russischen oder Spanischen (z.B. Sonne: *solri*, vgl. engl. *sun*, span. *sol*, chin. *rì* oder Nase: *nazbi*, vgl. engl. *nose*, russ. *nos*, chin. *bízi*).

3.5 Slovio

Slovio[11] ist eine im Jahre 1999 von dem Slowaken Mark Hucko entwickelte Plansprache, die vor allem auf slawischen Wortwurzeln und slawischer Grammatik basiert, die zum Teil noch in der Entwicklung begriffen ist. Die Lexeme ähneln sehr stark den aus dem Russischen oder Tschechischen bekannten Wörtern, vereinzelt wurde international bekannteren Wörtern ein Vorzug gegeben. Es ist zu erwarten, dass sich Slovio in die Gruppe der slawischen Sprachen einreiht. Interessanter ist die Frage, ob Huckos eigene slowakische Wurzeln die Wahl der Lexeme bewusst oder unbewusst prägten oder ob es z.B. eher eine Tendenz zum Russischen – der sprecherreichsten slawischen Sprache – gibt.

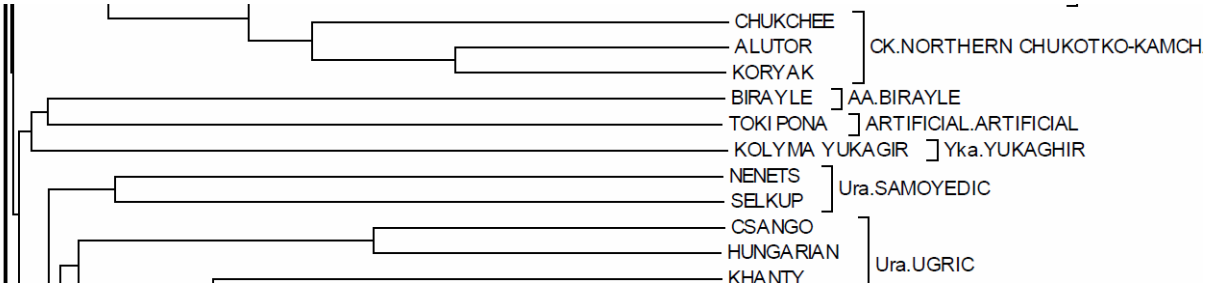


Im Baum nimmt Slovio die Position eines Schwesterknoten der südwestslawischen Sprachen Bosnisch, Serb(okroat)isch, Kroatisch und Slowenisch ein, welche zusammen mit Slovio einen Schwesterknoten zur Zweiergruppe Tschechisch–Slowakisch (zwei Vertretern des Westslawischen) bilden. Eine besonders große Ähnlichkeit speziell zum Slowakischen ist nicht zu erkennen, jedoch eine stärkere Ähnlichkeit zu den westlichen und südwestlichen slawischen Sprachen. Es ist jedoch anzumerken, dass die von ASJP erstellte Aufteilung des Slawischen nicht identisch ist mit den traditionellen genetischen Bäumen der slawischen Sprachen.[12] Das könnte auf gewisse *„shared innovations“* (d.h. identische Veränderungen in mehreren Sprachzweigen, die dieses Merkmal nicht von der gemeinsamen Ursprache ererbt haben) in der Phonetik der südwestlichen und westlichen slawischen Sprachen, eventuell bedingt durch Sprachkontakt, zurückzuführen sein.

[11] Äußerst wenig ist bisher über Slovio publiziert worden. Informationen zur Sprache findet man vorwiegend auf http://www.slovio.com. Vgl. auch Mannewitz (2009).
[12] Vgl. http://www.ethnologue.com/show_family.asp?subid=292-16 (Lewis 2009).

3.6 Toki Pona

Im Jahre 2001 entwickelte die Kanadierin Sonja Elen Kisa eine experimentelle Plansprache, die mit so wenig wie möglich Vokabular auskommen sollte: Toki Pona[13] (wörtl. „gute Sprache“), eine isolierende Plansprache mit einer Grammatik ähnlich vielen Pidginsprachen und einem Vokabular von rund 120 Wörtern. Mit diesen Wörtern könne jeder Begriff durch Komposition gebildet werden, ähnlich der Idee der universellen semantischen Primitiva von Wierzbicka.[14] So sind auch viele der 40 Wörter in der ASJP-Liste zusammengesetzte Begriffe, wie etwa Blut: *telo loje* (wörtl. rotes Wasser), wir: *mi mute* (ich viele) oder Ohr: *lupa kute* (Hörrohr). Viele Begriffe sind auch sehr vage, so steht *kasi* sowohl für Baum als auch für Blatt. Toki Pona kann zu den aposteriorischen Sprachen gezählt werden, da die Lexeme auf existierenden Sprachen basieren. Laut einem etymologischen „Wörterbuch“ auf der Toki-Pona-Webseite entstammt etwa die Hälfte des Vokabulars den Sprachen Tok Pisin, Finnisch, Kroatisch und Esperanto.[15] Auch aus vielen anderen Sprachen, wie Georgisch, Chinesisch, Akan und Tonganisch sind Wörter entlehnt und der Phonotaktik des Toki Pona angepasst worden. Wie schlagen sich diese Einflüsse im ASJP-Baum nieder?

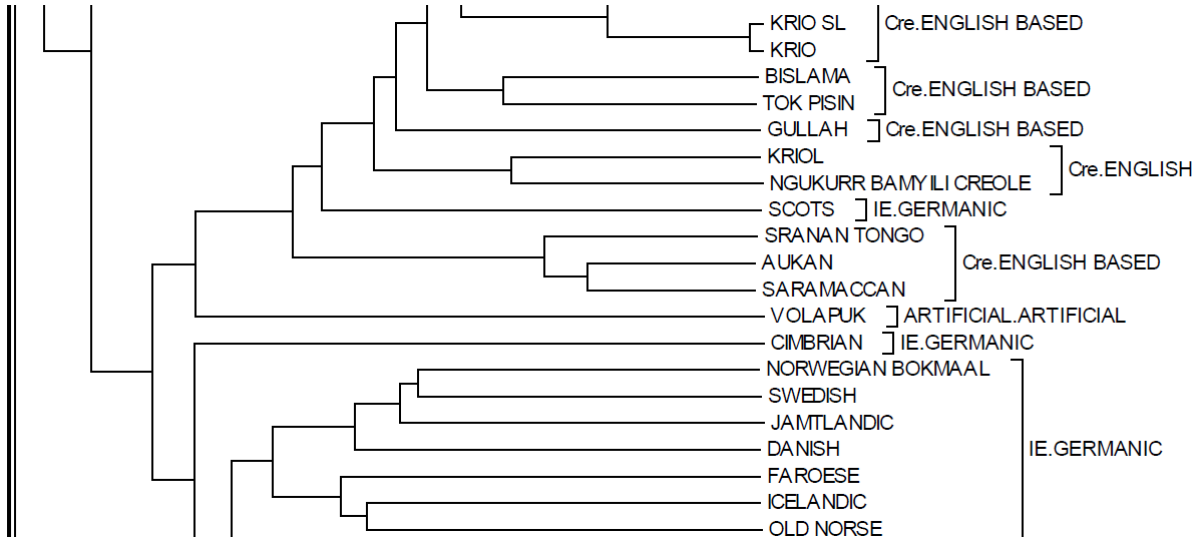


Toki Pona ist nicht unter den indoeuropäischen Sprachen zu finden, sondern wird auf einen Ast mit der Sprache Birayle bzw. Ongota (afroasiatisch, Äthiopien) gelegt, deren genaue Zuordnung in den afroasiatischen Sprachen noch unklassifiziert ist.[16] Die nächsthöhere Verknüpfung findet mit dem Jukagirischen statt, einer isolierten sibirischen Sprache mit zwei Varietäten (nur eine ist davon derzeit im ASJP enthalten). Dann folgen die uralischen Sprachen. Die verbindenden Knotenpunkte im Baum befinden sich äußerst weit links, was für eine große Unähnlichkeit der Lexeme spricht. Je weiter links sich diese Knotenpunkte befinden, desto größer wächst die Wahrscheinlichkeit einer eher zufälligen Zuordnung (vgl. Brown et al. 2008, 10). Ein Vergleich der 40 Wörter dieser beiden Sprachen zeigt zwei auffällig ähnliche Wörter: TP *kala* = Bir. *kara* (Fisch), TP *kasi* = Bir. *haši* (Blatt), während bei den anderen augenscheinlich keine Ähnlichkeit vorliegt. Die beiden genannten Toki-Pona-Wörter entstammen laut Kisa jedoch einer anderen Sprache: sie entsprechen den finnischen Wörtern *kala* (Fisch) und *kasvi* (Pflanze). Keines der Wörter im Toki Pona entstammt einer afroasiatischen Sprache, daher kann man von einer zufälligen Einordnung ausgehen.[17]

[13] Weiterführende Informationen über Toki Pona auf http://en.tokipona.org sowie auf http://en.wikipedia.org/wiki/Toki_Pona.
[14] Vgl. die Homepage der „*Natural Semantic Metalanguage*“: http://www.une.edu.au/bcss/linguistics/nsm/.
[15] Das Tok Pisin ist eine auf dem Englischen basierende Kreolsprache, während Esperanto lexikalisch vor allem an die romanischen und germanischen Sprachen angelehnt ist, wie in 3.1 bereits erwähnt. Das könnte das Bild zugunsten der germanischen oder romanischen Sprachen theoretisch verzerren.
[16] Vgl. http://www.ethnologue.com/show_language.asp?code=bxe (Lewis 2009).
[17] Für die Berechnung der Wahrscheinlichkeit zufälliger Übereinstimmung von Wörtern in zwei Sprachen siehe Meillet (1925) und Nichols (1997).

3.7 Volapük

Die Lexik des Volapük basiert zum größten Teil auf den europäischen Sprachen Englisch, Deutsch und Französisch, zu einem etwas kleineren Teil aber auch auf apriorischen Wortwurzeln wie etwa den in der 40-Wort-Liste enthaltenen Zahlwörtern *bal* (eins), *tel* (zwei) und den Pronomen *ob* (ich), *ol* (du) und *obs* (wir), die jeweils einem bestimmten sprachinternen System folgen. Das Volapük lässt sich also als Beispiel für eine Hybridsprache aus den beiden Typen a priori und a posteriori ansehen. Die aposteriorischen Wörter sind zum Teil durch die Phonotaktik der Sprache, wie auch durch (scheinbar?) willkürliche Veränderungen oft nicht mehr wiederzuerkennen, wie die Beispiele *vol* (< engl. *world*) und *pük* (< engl. *speak*) zeigen. Neben den 5 erwähnten Apriorismen sind 21 Wörter englischer, 10 deutscher und 4 französischer Herkunft. Die eindeutige Zuordnung wird jedoch durch die mögliche Mehrfachzuordnung einiger Wörter erschwert (z.B. kann *hon* sowohl engl. *horn* als auch dt. *Horn* zugeordnet werden). Ein Vorherrschen englischbasierter Wurzeln ist jedoch deutlich und repräsentativ für die Sprache selbst.[18]



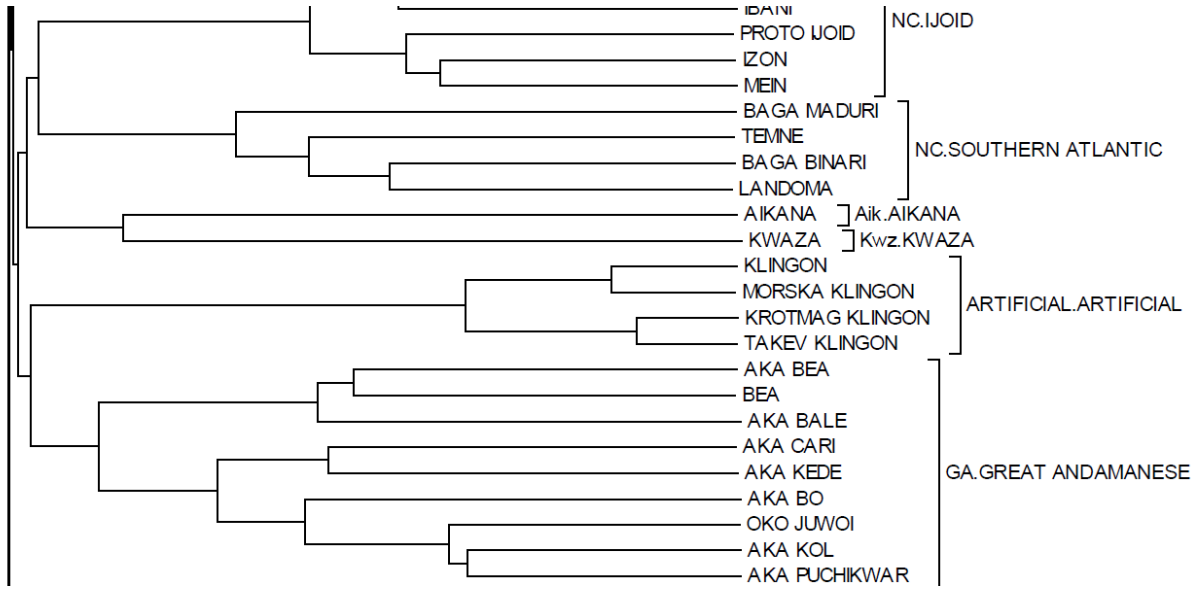
Und auch Volapük bildet zusammen mit den englischbasierten Kreolsprachen einen eigenen Zweig; die Neigung in Richtung des Englischen ist also auch mit objektiven Methoden messbar.

3.7 Klingonisch

Bei den apriorischen Kunstsprachen beziehe ich mich lediglich auf zwei große und bekannte Vertreter: das Klingonische und die beiden Elbensprachen Sindarin und Quenya (s. unten). Klingonisch gehört nicht zu den Plan- oder Welthilfssprachen, da es nicht dem Zweck der internationalen Verständigung dient. Entwickelt wurde die Sprache in den 80er Jahren des 20. Jahrhunderts vom amerikanischen Linguisten Marc Okrand für den Film *Star Trek III: The Search for Spock* und wurde dann auch in weiteren Filmen genutzt und ausgebaut. Heute gibt es zum Klingonischen eine komplette Grammatik (Okrand 1992) sowie ein Vokabular von insgesamt knapp 3000 Wörtern (eigene Recherche). Da Okrand die Sprache mit dem Ziel entwarf, sie möglichst exotisch, fremdartig und z.T. sehr harsch klingen zu lassen, basiert

[18] Sprague (1888) äußerte sich dazu: „The radicals have been taken principally from the following languages: English, Latin, German and French. More material has been taken from the English than from any other language. The English words are, however, much modified in adapting them to Volapük."

auch das Lexikon nicht auf existenten Sprachen. Einige Ausnahmen wie die Wörter *Human* (Mensch), *tera'* (Erde) und *ghotI'* (Fisch)[19] sind vorhanden, bilden aber nur einen äußerst geringen Teil dex Lexikons. Auch die Phonetik ist möglichst unterschiedlich zum Englischen gehalten und enthält viele uvulare und glottale Laute. Neben dem „Standardklingonischen" hat Okrand (1997, 14ff.) noch einige grammatikalische, lexikalische und vor allem phonetische Eigenschaften anderer klingonischer Dialekte entwickelt. Diese Änderungen habe ich in die Erstellung weiterer klingonischer Wortlisten einfließen lassen, sodass quasi die „klingonische Sprachfamilie" bzw. ein Dialektkontinuum in ASJP repräsentiert ist. Dass die Abweichungen der Dialekte untereinander eher gering sind, zeigen die Position und die Abstände zueinander im Sprachbaum:



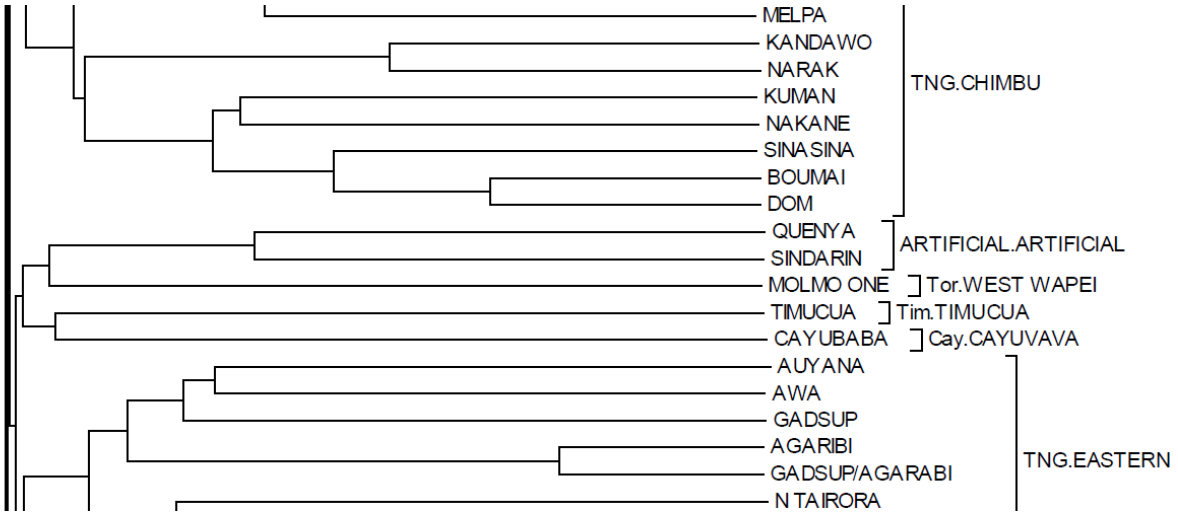
So ist der so genannte *Morska*-Dialekt dem Standardklingonischen (*tlhIngan Hol* bzw. *ta' Hol*) phonetisch-lexikalisch am nächsten, während die Dialekte der fiktiven Regionen *Krotmag* und *Tak'ev* sich untereinander sehr ähneln. Diese Ähnlichkeit ist geplant, wie Okrand (1997, 22) selbst ausführt: „Speaking in a manner that is sort of between that of the Krotmag region and **ta' Hol** are the people of Tak'ev (**taq'ev**) […]". Eine sehr weitläufige Verbindung scheint laut ASJP zu den Sprachen der Großen Andamanen (einer Inselgruppe im Indischen Ozean) zu bestehen, hier sind die Ähnlichkeiten zwischen Klingonisch und z.B. Aka Bea noch viel geringer und es sind nicht einmal „Scheinkognate" zu erkennen. Offenbar ist das Klingonische aber allen anderen Sprachen im Datensatz noch unähnlicher. Etwaige zu erwartende Ähnlichkeiten mit dem Englischen (Okrands Muttersprache) oder der kalifornischen Sprache Mutsun aus der Penuti-Familie, mit der sich Okrand in seiner linguistischen Karriere beschäftigt hat, bleiben aus. In phonetisch-lexikalischer Hinsicht scheint die Sprache also deutlich dem Typ a priori anzugehören.

3.8 Elbisch

Mit „Elbisch" fasse ich die Sprachen Sindarin und Quenya zusammen, die in J.R.R. Tolkiens Büchern *The Lord of the Rings*, *The Little Hobbit* u.a. Verwendung finden. Es ist ebenfalls eine fiktive Sprache, die sich jedoch zumindest grammatikalisch wie auch phonetisch mehr

[19] Letzteres geht auf das bekannte englische Wortspiel zurück, nach dem sich das Wort <ghoti> laut englischen Ausspracheregeln *fish* aussprechen lässt, wenn man die Phonem-Graphem-Korrespondenzen aus *enough*, *women* und *nation* als Grundlage nimmt.

oder weniger stark an europäische Sprachen anlehnt. Grammatikalisch gesehen ist Quenya an das Finnische angelehnt (Pesch 2003, 27), das Sindarin eher an die keltischen Sprachen (Pesch 2003, 30). Ähnlichkeiten zu europäischen Sprachen oder dem *Standard Average European* (Haspelmath 2001) wurden auch von anderen Seiten festgestellt (Julian Jarosch, persönliche Korrespondenz). Lexikalisch gesehen scheinen sich die von Tolkien erfundenen Wörter jedoch nicht eindeutig irgendeiner existierenden Sprache zuzuordnen. Sindarin und Quenya gruppieren sich auf einem Ast zusammen mit dem Molmo One[20], einer kleinen Sprache der Toricelli-Sprachfamilie, die in Papua Neuguinea gesprochen wird:



Ein Vergleich der 40 Wörter in den drei Sprachen zeigt große Ähnlichkeiten zwischen Sindarin und Quenya (die im Universum von Mittelerde auch einen gemeinsamen Vorfahren haben: das „Primitive Quendisch“ (Pesch 2003, 41)), jedoch kaum zum Molmo One. Mehr als eine ähnliche Silbenstruktur und eine grob ähnliche Verteilung der Vokale ist nicht festzustellen. Weitere Ähnlichkeiten mit anderen „Ausreißersprachen“ im Baum und mit Sprachen des Trans-Neuguinea-Phylums haben nur sehr geringe Signifikanz, sodass auch hier davon ausgegangen werden kann, dass das Lexikon der beiden Elbensprachen – im Gegensatz zu ihrer Grammatik und Phonotaktik – tatsächlich apriorisch ist. Betrachtet man die Sprache als Ganzes, könnte man auch hier von einem hybriden Typ ausgehen.

4 Zusammenfassung

In dieser Arbeit wurden das lexikostatistische Projekt des *Automated Similarity Judgment Program (ASJP)* und die computergestützten Methoden zur Berechnung der phonetisch-lexikalischen Ähnlichkeiten zwischen den Sprachen der Welt vorgestellt. Es wurde nur ein grober Einblick gegeben. Auf viele weitere Faktoren und Unterprojekte, wie etwa die automatische Erkennung von Lehnwörtern, die Einbeziehung von Synonymen, die Berechnung von Urheimaten oder Aufspaltungszeiten konnte nicht eingegangen werden.

Des Weiteren konnte ich zeigen, wie mit einer solchen objektiven Methode der Unterschied zwischen apriorischen und aposteriorischen Typen (auf deren Lexikon bezogen) bei Kunstsprachen deutlich gemacht werden kann: während sich Sprachen mit aposteriorisch konstruierter Lexik in den mit ASJP erzeugten Sprachbäumen meist eng verbunden mit ihren Kontrollsprachen zeigen, findet man die apriorischen Sprachen „irgendwo“ im Baum, oft sehr weitläufig verbunden mit anderen isolierten oder in ihren jeweiligen Familien stark

[20] Vgl. http://www.ethnologue.com/show_language.asp?code=aun (Lewis 2009).

abweichenden Sprachen. Sprachen, deren Lexikon auf sehr vielen unterschiedlichen Sprachen der Welt fußt, wie das Toki Pona, verhalten sich ähnlich wie apriorische Kunstsprachen. Da keine der Kontrollsprachen des Toki Pona stark überwiegt, und diese selbst an vielen verschiedenen Stellen im Baum zu finden sind, ordnet sich auch die Sprache selbst nicht einer einzelnen der Kontrollsprachen zu — die lexikalische Neigung ist dafür einfach zu gering.

Bibliografie

Anton, Günter (2001): Über die Struktur und Entwicklung des Ido im Vergleich zum Esperanto. In: Detlev Blanke (Hrsg.): Interlinguistische Informationen, Beiheft 7: Zur Struktur von Plansprachen, 30-47. Berlin: GIL.

Back, Otto (1980): Pri planlingva etimologio. In: Szerdahelyi István (Hrsg.): Miscellanea Interlinguistica, 266-276. Budapest: Tankönyvkiadó. [Online-Version: http://ling.cuc.edu.cn/htliu/plliufina.pdf].

Brown, Cecil H., Eric W. Holman, Søren Wichmann, Viveka Velupillai (2008): Automated classification of the World's languages: A description of the method and preliminary results. STUF – Language Typology and Universals 61.4, 285-308.

Fellmann, Ulrich (2001): Loglan: Sprache, Logik und soziale Realität. In: Blanke, Detlev (Hrsg.): Interlinguistische Informationen, Beiheft 7: Zur Struktur von Plansprachen, 118-140. Berlin: GIL.

Gledhill, Christopher (1998): The Grammar of Esperanto – A corpus-based description. München: LINCOM EUROPA.

Haspelmath, Martin (2001): The European linguistic area: Standard Average European. In: Language Typology and Language Universals (Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 20.2). Berlin: De Gruyter, 1492-1510.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Dik Bakker (2008a): Explorations in automated lexicostatistics. In: Folia Linguistica 42.2, 331-354. [Online-Version: http://email.eva.mpg.de/~wichmann/Explorations.pdf].

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, Dik Bakker (2008b): Advances in automated language classification. In: Antti Arppe, Kaius Sinnemäki, Urpu Nikanne (Hrsg.): Quantitative Investigations in Theoretical Lingustics, 40-43. Helsinki: University of Helsinki.

Janton, Pierre (1993): Esperanto: Language, literature, and community (Edited by Humphrey Tonkin, translated by Humphrey Tonkin, Jane Edwards, and Karen Johnson-Weiner). Albany: State University of New York Press.

Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In. Soviet Physics Doklady 10, 707-710.

Lewis, M. Paul (2009): Ethnologue: Languages of the World, 16th edition. Dallas, Texas: SIL International. [Online-Version: http://www.ethnologue.com]

Liu Haitao (2001): Pidgins, Creoles and planned languages – Linguistic development under special conditions. In: Klaus Schubert (Hrsg.): Planned Languages: From Concept to Reality. Brüssel: Hogeschool voor Wetenschap en Kunst, 121-177.

Mannewitz, Cornelia (2009): Sprachplanung im Internet: Das Projekt Slovio. In: Fiedler, Sabine (Hrsg.): Esperanto und andere Sprachen im Vergleich. (Interlinguistische Informationen. Beiheft 16) Berlin: GIL, 157-164.

Meillet, Antoine (1925): La méthode comparative en linguistique historique. Paris: Édouard Champion.

Moch, Gaston (1897): La question de la langue internationale et sa solution par l'Espéranto. Paris.

Müller, André, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, Anthony Grant

(2010): ASJP World Language Tree of Lexical Similarity: Version 3 (Juli 2010). [Online Version: http://email.eva.mpg.de/~wichmann/WorldLanguageTree-003.pdf]
Nakhleh, Luay, Don Ringe, Tandy Warnow (2005): Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. In: Language 81, 382-420.
Nicholas, Nick, John Cowan (2003): What is Lojban? – .i la lojban. mo. Online-Version: http://lojban.org/publications/level0/brochure/lingissues.html#AEN12599 [Zugriff: 29.06.2010].
Nichols, Johanna (1997): Modelling ancient population structures and population movement in linguistics and archeology. In: Annual Review of Anthropology 26, 359-384.
Okrand, Marc (1992): The Klingon Dictionary. New York: Pocket Books.
Okrand, Marc (1997): Klingon for the Galactic Traveler. New York: Pocket Books.
Oswalt, Robert L. (1970): The detection of remote linguistic relationship. In: Computer Studies in the Humanities and Verbal Behavior 3, 117-129.
Pesch, Helmut W. (2003): Elbisch. Bergisch Gladbach: Bastei Lübbe.
Schmidt-Radefeldt, Jürgen (1998): Plan- und Kunstsprachen auf romanischer Basis I. Allgemeine Problematik. In: Günter Holtus, Michael Metzeltin, Christian Schmitt (Hrsg.): Lexikon der Romanistischen Linguistik (LRL), Band VII, 680-686. Tübingen: Max Niemeyer Verlag.
Sprague, Charles E. (1888): Hand-Book of Volapük. New York: The Office Company. [Online-Version: http://personal.southern.edu/~caviness/Volapuk/HBoV/hbv.htm]
Swadesh, Morris (1955): Towards a greater accuracy in linguistic dating. In: International Journal of American Linguistics 21, 121-137.
Villemin, François (1983): Un essai de détection des origines du japonais à partir de deux méthodes statistiques. In: B. Brainerd (Hrsg.): Historical Linguistics, 116-135.
Wichmann, Søren, Eric W. Holman, Dik Bakker, Cecil H. Brown (2010a): Evaluating linguistic distance measures. In: Physica A 389, 3632-3639. [Online-Version: http://email.eva.mpg.de/~wichmann/WichmannReplyRevised8.pdf]
Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Sebastian Sauppe, Oleg Belyaev, Matthias Urban, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck, Helen Geyer (2010b). The ASJP Database (version 13). [Online-Version: http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm]