

GOEDOC – Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen

2017

Digitale Erschließung und systematische Annotation kolonialer Wörterbücher am Beispiel der Mayasprache K'iche'

Frauke Sachse, Michael Dürr, Christian W.R. Klingler

DARIAH-DE Working Papers

Nr. 22

Sachse, F.; Dürr, M.; Klingler, Christian W. R.: Digitale Erschließung und systematische Annotation kolonialer Wörterbücher am Beispiel der Mayasprache K'iche'
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2017
(DARIAH-DE working papers 22)

Verfügbar:

PURL: <http://resolver.sub.uni-goettingen.de/purl/?dariah-2017-2>

URN: <http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2017-2-3>

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: Die korpusorientierte Erfassung kolonialzeitlicher Wörterbücher und kulturwissenschaftlichen Forschung amerindischer Sprachen geht mit besonderen Problemstellungen einher. Die Textquellen sind multilingual und unstandardisiert verschriftet, die Lexikoneinträge morphologisch komplex und Bedeutungskorrelationen oft uneinheitlich. Die Überführung solcher Texte in maschinenlesbare Korpora setzt die orthographische Vereinheitlichung, morphologische Analyse und lexikalische Bedeutungszuordnung der Wortformen voraus, die sich als Analyseschritte nicht systematisch trennen lassen. Am Beispiel von kolonialen Wörterbüchern (Lexikographien) der Mayasprache K'iche' wurde an der Abteilung für Altamerikanistik der Rheinischen Friedrich-Wilhelms-Universität Bonn der Prototyp eines Annotationswerkzeugs entwickelt, das den Prozess von Transkription, Glossierung und Lemmatisierung im Rahmen eines halbautomatisierten Auszeichnungsverfahrens unterstützt. Ausgehend von diesem Prototyp soll in der XML-basierten Forschungsumgebung TextGrid ein digitales Werkzeug zur einheitlichen Erfassung kolonialer Lexikographien entstehen, die für die Auszeichnung vergleichbarer Wörterbücher zu anderen Sprachen nutzbar wird.

Keywords: XML-Annotation, koloniale Lexikographie, Wörterbücher, Mayasprachen, Korpuslinguistik, unstandardisierte Orthographie

XML-annotation, colonial lexicographies, dictionaries, Mayan languages, corpus linguistics, unstandardized orthographies

Digitale Erschließung und systematische Annotation kolonialer Wörterbücher am Beispiel der Mayasprache K'iche'

Frauke Sachse¹

Michael Dürr²

Christian W.R. Klingler¹

¹Rheinische Friedrich-Wilhelms-Universität Bonn

²Freie Universität Berlin



Frauke Sachse, Michael Dürr, Christian W.R. Klingler: „Digitale Erschließung und systematische Annotation kolonialer Wörterbücher am Beispiel der Mayasprache K'iche'“. *DARIAH-DE Working Papers* Nr. 22. Göttingen: DARIAH-DE, 2017. URN: [urn:nbn:de:gbv:7-dariah-2017-2-3](https://nbn-resolving.org/urn:nbn:de:gbv:7-dariah-2017-2-3).

Dieser Beitrag erscheint unter der
Lizenz [Creative-Commons Attribution 4.0 \(CC-BY\)](https://creativecommons.org/licenses/by/4.0/).

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm,
Thomas Kollatz, Stefan Schmunk und Christof Schöch
herausgegeben.



Zusammenfassung

Die korpusorientierte Erfassung kolonialzeitlicher Wörterbücher und kulturwissenschaftlichen Forschung amerindischer Sprachen geht mit besonderen Problemstellungen einher. Die Textquellen sind multilingual und unstandardisiert verschriftet, die Lexikoneinträge morphologisch komplex und Bedeutungskorrelationen oft uneinheitlich. Die Überführung solcher Texte in maschinenlesbare Korpora setzt die orthographische Vereinheitlichung, morphologische Analyse und lexikalische Bedeutungszuordnung der Wortformen voraus, die sich als Analyseschritte nicht systematisch trennen lassen. Am Beispiel von kolonialen Wörterbüchern (Lexikographien) der Mayasprache K'iche' wurde an der Abteilung für Altamerikanistik der Rheinischen Friedrich-Wilhelms-Universität Bonn der Prototyp eines Annotationswerkzeugs entwickelt, das den Prozess von Transkription, Glossierung und Lemmatisierung im Rahmen eines halbautomatisierten Auszeichnungsverfahrens unterstützt. Ausgehend von diesem Prototyp soll in der XML-basierten Forschungsumgebung TextGrid ein digitales Werkzeug zur einheitlichen Erfassung kolonialer Lexikographien entstehen, die für die Auszeichnung vergleichbarer Wörterbücher zu anderen Sprachen nutzbar wird.

Schlagwörter

XML-Annotation, koloniale Lexikographie, Wörterbücher, Mayasprachen, Korpuslinguistik, unstandardisierte Orthographie

Keywords

XML-annotation, colonial lexicographies, dictionaries, Mayan languages, corpus linguistics, unstandardized orthographies

Inhaltsverzeichnis

1	Einleitung	4
2	Koloniale Wörterbücher des K'iche'	4
3	Kriterien und Probleme der Korpusfassung	7
4	Tool for Systematic Annotation of Colonial K'iche'* (TSACK)	12
5	Auszeichnungsverfahren	14
6	Ergebnis und Ausblick	26
7	Literaturverzeichnis	28
8	Abkürzungsverzeichnis linguistischer Glossierungen	30
9	Anhang	30

1 Einleitung

Das Thema dieses Beitrags ist die Suche nach digitalen Lösungen für die lexikographische Erschließung von unstandardisiert verschrifteten historischen Dokumentationen amerindischer Sprachen. Ein besonderes Problem stellt hierbei die korpusorientierte Erfassung kolonialzeitlicher Wörterbücher dar. Die Überführung solcher multilingualer Texte in maschinenlesbare Korpora setzt Verfahren zur orthographischen Vereinheitlichung, Lemmatisierung und Morphologisierung voraus, die sich mit bestehenden Lösungen und Werkzeugen bisher nicht umsetzen lassen.

Am Beispiel kolonialer Lexikographien der Mayasprache K'iche' werden an der Abteilung für Altamerikanistik der Rheinischen Friedrich-Wilhelms-Universität Bonn grundlegende Erfassungskriterien erarbeitet und darauf basierend eine Software-Anwendung entwickelt, mit der sich die kolonialzeitlichen Wörterbuchtexte digital annotieren lassen. Diese Ergebnisse bilden die Grundlage für den Aufbau eines digitalen Werkzeugs, das eine zentrale und effiziente Erfassung kolonialer Lexikographien ermöglichen soll. Dieses Tool soll in die XML-basierte Forschungsumgebung TextGrid (<https://textgrid.de/>) eingebettet werden und neben einer Software zur effizienteren Erfassung und Auszeichnung der Wörterbuchtexte in XML/TEI auch eine Oberfläche zur gezielten Korpusabfrage bereitstellen. Die Grundlagen werden zunächst an den kolonialen Wörterbüchern des K'iche' entwickelt und getestet; das digitale Werkzeug soll langfristig aber auch für lexikographische Quellen anderer Sprachen mit ähnlichen strukturellen Anforderungen zur Verfügung stehen sowie für andere Dokumenttypen (z.B. indigensprachige monolinguale Texte, koloniale Grammatikbeschreibungen) nachnutzbar sein.

2 Koloniale Wörterbücher des K'iche'

Während der Kolonialzeit verfolgte die Kirche in Lateinamerika phasenweise eine Strategie der Evangelisierung in den indigenen Sprachen. In diesem Kontext entstanden ab dem 16. Jahrhundert zu verschiedenen amerindischen Verkehrssprachen vielfältige und umfangreiche Sprachdokumente, die neben doktrinalen Texten (Katechismen, Predigten, Bibelübersetzungen) auch missionarslinguistische Grammatiken und Wörterbücher umfassen. Das K'iche' im Hochland von Guatemala gehörte zu den ersten Sprachen Amerikas, die in der Mission verwendet und systematisch dokumentiert wurden. Zur Zeit der spanischen Eroberung im Jahr 1524 war K'iche' die dominante Sprache der Region und bildet mit weit über einer Million Sprecher bis heute die größte Sprechergemeinschaft unter den dreißig Mayasprachen. Aufgrund seiner guten Dokumentationssituation kann dem K'iche' in der Erforschung kolonialer Sprachen ein besonderer Stellenwert eingeräumt werden. Neben einer Vielzahl missionarslinguistischer Sprachbeschreibungen und doktrinaler Texte liegen uns auch Textquellen autochthoner Geschichtsschreibung sowie notariell-administrative Dokumente aus der Hand indigener Autoren vor. Der überwiegende Teil dieser Sprachdokumente befindet sich heute in Archiven in Europa und den USA (siehe Carmack 1973; Weeks 1990; Niederehe 1994, 1999).

Die koloniale Lexikographie des K'iche' lässt sich nicht von der ihrer Schwestersprachen Kaqchikel und Tz'utujil trennen, da Wörterbücher oft als vergleichende, multilinguale Arbeiten erstellt wurden. Hierzu zählen die beiden Versionen des anonymen *Vocabulario copioso* (BnF Manuscrit Américaine ms. 46 und JCBL) sowie der sogenannte *Tesoro* von Francisco Ximénez, die allesamt Kaqchikel als

Matrixsprache verwenden und abweichende K'iche'- und Tz'utujil-Formen gesondert aufführen. Das erste eigenständige Wörterbuch zum K'iche' liegt mit dem *Vocabulario Quiché* von Domingo de Basseta (1698) vor. Das *Vocabulario de la lengua Quiche Otlatecas* aus dem Ibero-Amerikanischen Institut in Berlin basiert teilweise auf dem *Vocabulario copioso* und ist das einzige koloniale Wörterbuch mit K'iche' als Matrixsprache (Dürr & Sachse im Druck). Als zentrale Quellen mit anonymer Autorschaft sind das Franziskaner-Wörterbuch aus der Tozzer Library in Harvard (1787) und das Ms. Angrand 9 aus der Bibliothèque National de France und zu nennen. Es bleibt bislang unklar, ob die übrigen erhaltenen Wörterbuch-Quellen ganz oder teilweise auf älteren Vorlagen basieren. Verglichen mit dem umfangreichen Dokumentationsstand des Kaqchikel (siehe Acuña 1983; Smailus 1989; Hernández 2009) sind die Wörterbücher des K'iche' bisher kaum erschlossen und nur Basseta und Ximénez als Editionen zugänglich (Acuña 2005; Sáenz de Santa María 1985).

Tabelle 1: Übersicht der wichtigsten erhaltenen kolonialen K'iche'-Wörterbücher

Jahr	Autor	Titel	Umfang	Ort	Ms.-Nr.
17. Jh.	Anonym	* <i>Vocabulario copioso (Vocabulario de la lengua cakchiquel...)</i>	286 fols.	BnF	MA 46
17. Jh.	Anonym	<i>Vocabulario copioso...</i>	705 pp.	JCBL	06396
1698	Basseta, D.	<i>Vocabulario quiché</i>	248 fols.	BnF	MA 59
1722	Ximénez, F.	<i>Primera parte del tesoro de las lenguas Cakchiquel, Quiche y Tzutuhil...</i>	211 fols	BANC	M-M 445
1722	Ximénez, F.	<i>Primera parte del tesoro de las lenguas Cakchiquel, Quiche y Tzutuhil...</i>	132 fols.	BPC	83
1745	Barrera, F.	<i>Abecedario en la lengua que dize qiche</i>	134 fols.	PUL	GGMA-160
1787	Anonym	<i>Vocabulario de la lengua kiché</i>	218 fols.	TOZ	C.A.6V85
18. Jh.	Anonym	<i>Bocabulario en lengua Quiche y Castellana (A-M)</i>	206 fols.	PUL	GGMA-161
18. Jh.	Anonym	<i>Vocabulario en lengua Kiché-castellana</i>	139 fols	BnF	ANG 9
18. Jh.	Anonym	<i>Vocabulario de la lengua castellana y quiché</i>	100 fols.	BnF	MA 64
18. Jh.	Anonym	<i>Vocabulario en lengua 4iche Otlatecas</i>	267 fols.	IAI	Y/2997
19. Jh.	Anonym	<i>Vocabulario de las lenguas quiche y kakchiquel (lettres A, B, C, K, T)</i>	76 fols.	BnF	MA 65

Abkürzungen: BANC = Bancroft Library, University of California at Berkley; BnF = Bibliothèque Nationale de France, Paris (MA: Manuscrits Américaines; ANG: Colección Angrand); BPC = Biblioteca Provincial de Córdoba; IAI = Ibero-Amerikanisches Institut, Berlin; JCBL = John Carter Brown Library, Providence; PUL = Princeton University Library, Princeton (GGMA: Garrett-Gates Collection of Mesoamerican Manuscripts); TOZ = Tozzer Library, Harvard University, Cambridge, Massachusetts

Die kolonialen Wörterbücher geben wertvolle Einblicke in die Entwicklung von Sprache und Wortschatz des K'iche', das in den vergangenen fünfhundert Jahren vor allem auf der Ebene der lexikalischen Semantik einen deutlichen Sprachwandel erfahren hat. Die kolonialen Wörterbücher sind daher in vielen Fällen die einzigen Quellen für Wortbedeutungen und Metaphern, die sich in den modernen Sprachvarianten nicht erhalten haben, und daher von besonderem Wert für die lexikographische und historisch-semantische Forschung.

Den kolonialen Lexika kommt eine zentrale Rolle bei der Übersetzung und inhaltlichen Erschließung k'iche'-sprachiger ethnohistorischer Textquellen zu. In indigenen Gemeinden des Hochlands entstanden in der frühen Kolonialzeit verschiedene Dokumente autochthoner Geschichtsschreibung, die unter anderem dazu eingesetzt wurden, gegenüber der kolonialen Verwaltung territoriale und dynastische Ansprüche geltend zu machen (siehe Carmack 1973). Diese Texte sind heute zentrale Quellen zur Ethnohistorie des vorspanischen Hochlands und umfassen neben dem bekannten und mehrfach herausgegebenen *Popol Vuh* (siehe z.B. Christenson 2003) mehr als ein Dutzend sogenannter *Títulos*, die bisher nur unzureichend erschlossen sind. Da die Texte modernen K'iche'-Sprechern nicht mehr ohne Weiteres verständlich sind, können die Inhalte oft nur über Einträge in den kolonialen Wörterbüchern erschlossen werden. Auch die Analyse k'iche'-sprachiger Verwaltungsdokumente (Urkunden, Testamenten oder Rechtsprotokolle) muss auf die Wörterbücher zurückgreifen, da sich bestimmte Begriffe (z.B. Landeinheiten) im modernen K'iche' nicht erhalten haben.

Die Wörterbücher wurden primär für das Sprachstudium der Missionare kompiliert und enthalten auf die Evangelisierung speziell zugeschnittenes Vokabular. Dieses Vokabular schufen die kolonialen Lexikographen selbst, indem sie k'iche'-sprachige Entsprechungen für christliche Konzepte einführten (siehe Sachse 2015; 2016). Die Wörterbücher enthalten somit eine Vielzahl an doktrinalen Neologismen, aus denen sich Rückschlüsse über die Entstehung christlicher Diskurse im K'iche' ziehen lassen. Der Übersetzungsprozess war nicht standardisiert, so dass die verschiedenen Wörterbücher nicht selten abweichende Übersetzungskorrelationen und multiple semantische Bezüge aufweisen. Somit helfen missionarisch-linguistische Wörterbücher auch bei der inhaltlichen Erschließung von auf K'iche' verfassten doktrinalen Textquellen (z.B. Katechismen oder Predigten) sowie bei der Identifizierung doktrinaler Diskurse in den indigenen Quellen.

Textsynoptische Zusammenhänge zwischen den Wörterbuchquellen geben Aufschluss über den Prozess der kolonialen Wissensentstehung. Wörterbücher, Grammatiken und christliche Lehrtexte wurden zum Teil von denselben Autoren – vornehmlich Mitglieder der franziskanischen und dominikanischen Orden – verfasst bzw. verblieben in den Konventen und wurden dort wieder kopiert. Die Lexikographen arbeiteten im Kollektiv und übernahmen häufig Einträge aus anderen Kompilationen, die sie dann später modifizierten und weiter bearbeiteten (siehe Smith-Stark 2009: 30-31). So entstanden unterschiedliche Abschriften und Versionen von Wörterbüchern. Es lässt sich nachweisen, dass einige K'iche'-Wörterbücher auf der Basis von Kaqchikel-Lexika kompiliert wurden (Sachse 2009: 15-16). Über die Rekonstruktion der Quellengenese von kolonialen Sprachbeschreibungen lassen sich so Rückschlüsse auf die Kommunikationsstrukturen zwischen den Orden ziehen.

Die Zusammenführung der kolonialen Wörterbücher des K'iche' in einem digitalen Korpus würde die bisher weitgehend unveröffentlichten Sprachdaten der Forschung zugänglich machen und eine systematische lexikographische Auswertung und Disambiguierung abweichender Bedeutungskorrelationen ermöglichen. Trotz der Relevanz der Wörterbücher für die sprach- und kulturwissenschaftliche Forschung

bleiben digitale Erschließungsmethoden zur korpusorientierten Erfassung kolonialer Wörterbuchquellen amerindischer Sprachen jedoch bislang ein Desiderat der Forschung.

3 Kriterien und Probleme der Korpuserfassung

Die systematische Überführung kolonialzeitlicher Wörterbücher in digitale Textkorpora geht mit besonderen Problemstellungen einher, die bei der Festlegung von Erfassungsstandards berücksichtigt werden müssen.

Der erste Schritt der Erfassung ist die manuelle Abschrift bzw. Transliteration der Manuskripttexte in ein maschinenlesbares Format. Die Transliteration muss genau sein und auch Fehler des Verfassers bzw. Kopisten des Manuskripts abbilden, da diese für die Analyse von Verschriftungspraktiken und die Rekonstruktion synoptischer und stemmatologischer Beziehungen zwischen Manuskripttexten gegebenenfalls relevant sein können (Dürr 1994). Zu berücksichtigen ist ferner, dass bei der Abschrift kolonial verschrifteter, indigen-sprachiger Texte Ambiguitäten und Uneindeutigkeiten auftreten können und Formen ggf. nicht oder nur teilweise lesbar sind. Voraussetzung für die Kompilation eines systematischen Gesamtkorpus ist die Festlegung einheitlicher Konventionen zur Auflösung standardisierter Kürzel (z.B. \widehat{q} q(ue), S^{to} S(an)to etc.), Verwendung von Groß- und Kleinschreibung, Darstellung von Wortgrenzen, Zeilen- und Seitenumbrüche sowie zum formalen Umgang mit nicht oder nur eingeschränkt lesbaren Textelementen.

Für die Definition von Erfassungsstandards ist die interne Organisation der Wörterbücher besonders zu berücksichtigen. Diese steht in der Tradition europäischer Lexikographien, wie sie vor allem mit dem monolingualen Latein-Wörterbuch von Ambrosio Calepino (1502) und bilingualem Spanisch-Latein-Wörterbuch von Antonio de Nebrija (1492) vorgelegt und mit dem Nahuatl-Wörterbuch von Alonso de Molina (1571) in Neuspanien weiterentwickelt wurden (Smith-Stark 2009: 11-16; 27; Hernández 2009: 130; 142). Sowohl Spanisch (1) als auch K'iche' (2) sind als Matrixsprache für die Organisation der Lexikoneinträge belegt, wobei die Sortierung Spanisch-K'iche' vorherrscht. Nach dem Vorbild des Calepino-Wörterbuchs enthalten die Wörterbucheinträge oft diskursive Satzbeispiele, die meist doktrinalen Texten entstammen (2).

(1) [Antiguamente.]_{spanish} [Oher canoK. May canoK. Nima oher]_{k'iche'}

(Anonym, *Vocabulario de la lengua kiché*, fol. 19v)

(2) [Pixabafj. canu.]_{k'iche'} [mandar, en comendar.]_{spanish 1} [xupixabafj cana K(ahual) J(esu) (Christo) chiquech. App(ostole)s vbixic vtzifj chui ronohel xecafj.]_{k'iche' example 1} [mando N(uestr)o S(eñor) J(esu) (Christo) a sus App(ostole)s que dijese su palabra por todo el mundo.]_{spanish example 1} [o despedir.]_{spanish 2} [cat hupixabafj.]_{k'iche' example 2} [me despido de ti.]_{spanish example 2} [Paçurumal manaxofj apixabafj canok]_{k'iche' example 3} [p(o)r que no te despediste de nosotros.]_{k'iche' example 3}

(Anonym, *Vocabulario de la lengua Quiche Otlatecas*, fol. 150r):

Die Matrixsprache kann Aufschluss über den Prozess der Kompilation geben. Wörterbücher mit spanischer Matrix sind in der Regel auch auf der Basis des Spanischen kompiliert, d.h. Lexikographen suchten systematisch nach K'iche'-sprachigen Entsprechungen für spanische Einträge bzw. Lemmata. Ein eindrucksvolles Beispiel für dieses Verfahren bietet das *Vocabulario de la lengua castellana y quiché* (BNF ms. 64), in dem sämtliche Matrixeinträge auf Spanisch von einem Autor verfasst sind, während die K'iche'-sprachigen Übersetzungen dem Schreibstil nach zu urteilen von der Hand eines indigenen Schreibers ergänzt wurden. In Spanisch-K'iche' sortierten Wörterbüchern sind auch die lexikalischen Unterscheidungen meist auf der Basis des Spanischen definiert. Bei Wörterbuchquellen mit K'iche' als Matrixsprache kann es sich sowohl um Umkehrungen von Wörterbüchern mit spanischer Matrix handeln als auch um Quellen, die auf der Basis des K'iche'-Wortschatzes kompiliert wurden, was im Einzelfall bisher jedoch nicht gut untersucht ist. Die Wörterbücher von Vico, Villacañas und Ximénez sind multilingual als Spanisch-Kaqchikel-Lexika organisiert, die zusätzlich vom Kaqchikel abweichende K'iche'- und Tz'utujil-Formen gesondert aufführen. Das Korpus muss diese unterschiedlichen Sprachen abbilden und den jeweiligen strukturellen Bedingungen Rechnung tragen können.

Einzelne Lemmata können in verschiedenen Wörterbüchern mit unterschiedlichen Übersetzungen angegeben sein. So gibt das „Anonyme Franziskaner-Wörterbuch“ für den spanischen Begriff *convertirse* „konvertieren“ bzw. *enmendarse* „sich bessern“ den Ausdruck <cantzoleomifj nu4oheic>¹ *kan[ujtzoql'omij nuk'oje'ik* „ich wende (= umdrehen) meine Existenz“ an (3), während Basseta mit <xutzir ucoheic ro P(edr)o> *xutzir uk'oje'ik ri Pedro* „es besserte sich Peters Dasein/Existenz“, <xucanah umac> *xukanaj umak* „er/sie ließ seine Sünde zurück“, <xucanah itzel be> *xukanaj itzel b'e* „er/sie ließ den schlechten/bösen Weg zurück“ und <uhachom ranima ruc caxtoε> *ujachom ranima ruk' k'axtok'* „die Trennung seiner/Ihrer Seele vom Teufel“ dem Eintrag insgesamt vier verschiedene Bedeutungen zuordnet (4).

(3) [Combertirse, ô emmendarse.]_{spanish 1} [Cantzoleomifj nu4oheic.]_{k'iche' 1} [emmiendo ô emmendare mis costumbres.]_{spanish 2}

(Anonym, *Vocabulario de la lengua kiché*, fol. 63v)

(4) [Conuertirse o emendarse]_{spanish 1} [xutzir ucoheic ri P.o]_{k'iche' 1} [xucanah umac]_{k'iche' 2} [xucanah itzel be]_{k'iche' 3} [uhachom ranima ruc caxtoε]_{k'iche' 4}

(Basseta fol. 42r)

Solche multiple Zuordnungen betreffen auch Wörterbücher mit K'iche'-Matrix. In Beispiel (5) ist der K'iche'-Eintrag *loq'om* mit den drei spanischen Verben *amar* „lieben“, *apreciar* „schätzen“ und *tener en mucho* „viel von jmd. halten“ korreliert.

(5) [Loεom.]_{k'iche'} [amar,]_{spanish 1} [apreciar,]_{spanish 2} [tener en mucho.]_{spanish 3}

(Anonym, *Vocabulario de la lengua Quiche Otlatecas*, fol. 110v)

Multiple Bedeutungskorrelationen können auf eine mögliche Kompilation des Eintrags aus verschiedenen Quellen hinweisen und erlauben Rückschlüsse auf den Übersetzungsprozess. Da die kolonialen Lexikographen über Mechanismen der Wortbildung, Bedeutungserweiterung und Lehnübersetzung

¹Zur Erläuterung der orthographischen Konventionen kolonialer Wörterbücher sei hier auf Tabelle 2 verwiesen.

Neologismen für bisher nicht bekannte abendländische Konzepte schufen (siehe Smith-Stark 2009: 63-64), können die in diesen Kontexten verwendeten K'iche'-Formen semantisch von autochthonen Wortbedeutungen divergieren.

Im Prozess der Korpusgenerierung müssen multiple Bedeutungskorrelationen entsprechend erfasst werden. Anhand des oben aufgeführten Beispiels (4) stellt sich die Korrelierung wie folgt dar:

(6) [Conuertirse]_{spanish 1} o [emendarse]_{spanish 2} [xutzir ucoheic ri P.o]_{k'iche' 1} [xucanah umac]_{k'iche' 2} [xucanah itzel be]_{k'iche' 3} [uhachom ranima ruc caxtoe]_{k'iche' 4}

spanish 1 = Conuertirse	k'iche' 1 = xutzir ucoheic ri P.o
	k'iche' 2 = xucanah umac
	k'iche' 3 = xucanah itzel be
	k'iche' 4 = uhachom ranima ruc caxtoe
spanish 2 = emendarse	k'iche' 1 = xutzir ucoheic ri P.o
	k'iche' 2 = xucanah umac
	k'iche' 3 = xucanah itzel be
	k'iche' 4 = uhachom ranima ruc caxtoe

Verkompliziert wird die Erfassung der Bedeutungszuordnungen ferner dadurch, dass ggf. nicht alle spanischen Einträge mit sämtlichen angegebenen K'iche'-Formen korreliert werden können. So dient das zweite spanische Verb im Beispiel lediglich der Erläuterung des Haupteintrags und ist eigentlich nur durch die erste K'iche'-Form sinnvoll übersetzt. Für die Erfassung bedeutet dies, dass die beiden spanischen Einträge miteinander kontextualisiert bleiben müssen, um bei späteren Abfragen die Bedeutungsbezüge herstellen zu können.

Bereits im Erfassungsprozess ist zu berücksichtigen, dass multiple Übersetzungen nicht immer syntaktisch linear dargestellt sind. Im folgenden Beispiel (7) ist der Eintrag *abrirse la pared* „die Wand öffnen“ mit zwei unterschiedlichen Verben angegeben, welche jeweils durch die gleiche Nominalphrase komplementiert werden, die allerdings nur einmal angegeben ist und bei der Erfassung entsprechend ergänzt werden muss.

(7) [Abrirse la pared.]_{spanish} [Capax]_{k'iche' 1}, [carrakarox.]_{k'iche' 2} [Vuach xan.]_{k'iche' 1, 2}

(Anonym, *Vocabulario de la lengua kiché*, fol. 6r)

spanish 1 = Abrirse la pared	k'iche' 1 = Capax [Vuach xan]
	k'iche' 2 = carrakarox Vuach xan

Wörterbucheinträge können in verschiedene Untereinträge organisiert sein, wobei sich insbesondere in den Kompilationen mit K'iche'-Matrix auch Homonyme als separate Untereinträge von Haupteinträgen finden. Dabei ist die Abgrenzung von homonymen und polysemen Formen oft auf den ersten Blick nicht eindeutig. Darüber hinaus werden durch den Verschriftungsprozess in die spanisch-basierte Orthographie (s.u.) phonemische Unterschiede oft nicht adäquat abgebildet, so dass unterschiedliche Lexeme unter demselben Eintrag aufgelistet sein können. Die Korpuserfassung soll dazu dienen, diese verschiedenen Bedeutungskorrelationen systematisch zu erkennen und Mehrfachzuordnungen von Homonymen zu differenzieren. Da sich die jeweiligen Korrelationen erst im Zuge der Kompilation ergeben, müssen Bedeutungszuschreibungen im Annotationsverfahren revidierbar sein.

Einige der vorangehenden Beispiele zeigen, dass die K'iche'-Einträge in den Wörterbüchern meist in Form von Phrasen und vollständigen Sätzen angegeben sind, deren Bestandteile einzeln zu erfassen und auszuzeichnen sind. Dabei sind K'iche'-Wörter in der Regel komplex und müssen morphologisch analysiert werden, um das darin enthaltene Lexem bzw. Lemma im Korpus suchbar zu machen. Dieser Prozess der Lemmatisierung ist im K'iche', das wie die meisten amerindischen Sprachen morphologisch agglutinierend ist, in der Regel transparent, da Affigierung die Wortwurzel bzw. den Wortstamm nur in seltenen Ausnahmefällen verändert. Beispiel (8) zeigt die morphologische Analyse des Eintrags *abroquelarse* „sich (mit einem Schild) schützen“ aus dem anonymen Franziskaner-Wörterbuch. Das Lemma <pocoba> *pokob'a* trägt hier die vollständige Verbflexion und ist als transitiviertes Verb, das aus dem Substantiv <pocob> *pokob'* „Schild“ abgeleitet wird, selbst komplex. Sowohl Lemma als auch Wortwurzel und Derivationsmarker sind hier jeweils einzeln zu erfassen.

(8) *abroquelarse*. canupocobafj vib (Anonym, *Vocabulario de la lengua kiché*, fol. 6r)

canu**pocoba**h vib
ka-∅-nu-**pokob'-a**-j w-ib'
INC-3.SG.ABS-1.SG.ERG-N:shield-TRVZ-MOD.VTD 1.SG.POSS-REFL
'I protect myself'

Der Prozess der morphologischen Analyse und Lemmatisierung ist fehleranfällig. Die Anwendung sicherer, automatisierter Verfahren wie Parser und Lemmatisierungstools setzt jedoch immer eine standardisierte Verschriftung voraus. Die nicht standardisierten Orthographien bilden somit das Kernproblem der systematischen Erfassung. Dies betrifft nicht nur die kolonialen Wörterbücher des K'iche', sondern auch die anderer amerindischer Sprachen. Zur Verschriftlichung wurde das spanische Alphabet verwendet, das das Phoneminventar des K'iche' nur unzureichend abbilden kann. Das K'iche' kontrastiert unglottalisierte und glottalisierte Plosiv- und Affrikatlaute, die im Spanischen keine Entsprechung haben und in der Kolonialzeit mit einer eigens dafür entwickelten Notation verschriftet wurden (Campbell 1977; Dürr 1987) (siehe Tabelle 2).

Tabelle 2: Parra-Verschriftungskonventionen

Moderne Orthographie	Phonem	Parra
k'	k'	< 4 >
q	q	< k >
q'	q'	< ε >
tz'	ʧ'	< 4, >
ch'	č'	< 4h >

Diese von Francisco de la Parra eingeführten Sonderzeichen wurde jedoch nicht von allen Verfassern bzw. Schreibern einheitlich verwendet, so dass die Dokumente unterschiedliche orthographische Konventionen aufweisen. Im Zuge einer zunehmenden Hispanisierung ab der zweiten Hälfte des 17. Jhs. gaben einige Autoren das Parra-Alphabet teilweise oder vollständig auf, während in anderen Quellen

eine übergeneralisierte Verwendung der Symbole (z.B. < ε > = k', q, q') festzustellen ist. Dabei weichen die verschiedenen Wörterbücher nicht nur untereinander ab, sondern auch innerhalb einzelner Quellen finden sich zum Teil erhebliche Inkonsistenzen.

Um einzelne Lexeme suchbar machen und multiple semantische Korrelationen erfassen zu können, ist es daher notwendig, die kolonialzeitliche Verschriftung durch Transkription in eine phonemische Umschrift zu vereinheitlichen. Dieser Prozess birgt in Abhängigkeit von Sprache und Quelle jedoch eine hohe Fehleranfälligkeit. Im folgenden Beispiel aus dem *Vocabulario de la lengua kiché*² (Abbildung 1) werden die Konzepte des „Fälschers“ (*falsario, escritor falso* „Fälscher, Falschschreiber“) und des „Lügners“ (*falso decidor* „Falschredner“) mit zwei unterschiedlichen Verschriftungen angegeben (<Tzakol tzih> und <4,akol tzih>). Dass es sich hier in beiden Fällen um den gleichen K'iche'-Ausdruck handelt, zeigt das Komplement <chi 4,ibanic> *chi tz'ib'anik* „im/vom Schreiben“ im ersten Beispiel an, das den „Fälscher“ als „Lügner im Schreiben“ deklariert. Welche der beiden Formen nach der Parra-Konvention fehlerhaft verschriftet ist, bleibt uneindeutig, da beide semantisch plausibel sind und ein Lügner sowohl jemand sein könnte, der Worte wegwirft als auch jemand, der Worte konstruiert.

TzaKol tzih *(chi 4,ibanic)*	*Transliteration*	4,aKol tzih
tzaqol tzij	*Transkription*	tz'aqol tzij
tzaq-ol tzij	*Morphologische Analyse*	tz'aq-ol tzij
wegwerfen-AGT Wort	*Glossierung*	konstruieren-AGT Wort
Wort-Wegwerfer	*wörtliche Übersetzung*	**Wort-Konstrukteur**
Lügner (im Schreiben)	*Übersetzung*	Lügner

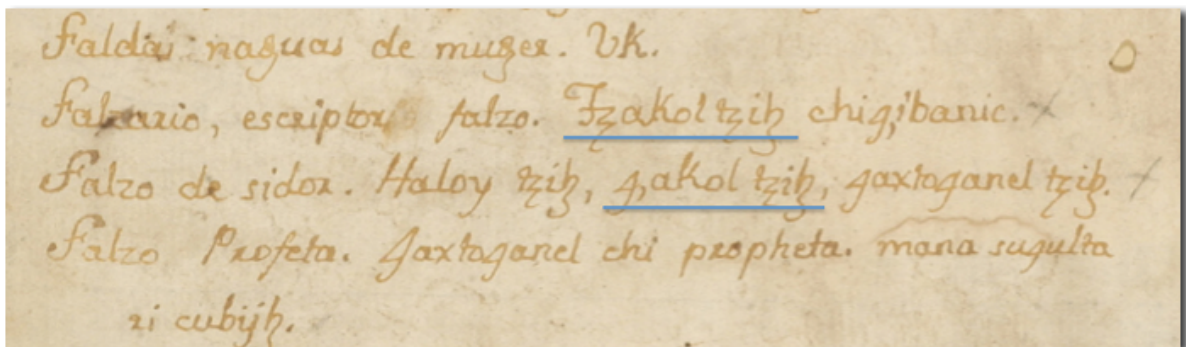


Abbildung 1: Beispiel für multiple Verschriftung aus dem *Vocabulario de la lengua Kiché* (Garrett-Gates Mesoamerican Manuscript Collection, Princeton, ms. 162)

Die genannten Beispiele zeigen, dass sich die Prozesse der phonemischen Transkription, Lemmatisierung bzw. morphologischen Analyse, Glossierung und Übersetzung nicht als einzelne Schritte systematisch voneinander trennen lassen und nur bedingt automatisierbar sind.

²Die Seiten mit dem hier vorgestellten Beispiel gehören zum Bestand der *Garrett-Gates Mesoamerican Manuscripts Collection* der Princeton University Library (ms. 162) und wurden unter bisher ungeklärten Umständen dem Originalmanuskript, das sich heute in der *Special Collection* der Tozzer Library an der Harvard University (ms. C.A.6V85) befindet, entwendet (siehe Sachse 2009).

Für die Schwierigkeiten der Korpuserfassung, die sich aus der Kombination von multiplen Verschriftungskonventionen und komplexen Sprachformen in multilingualen Textdokumenten ergeben, gibt es bisher noch keine Lösungsvorschläge. Die Erfassungskriterien, die am Beispiel der kolonialen K'iche'-Wörterbücher erarbeitet wurden, bilden die Basis für das hier im Folgenden vorgestellte Annotationswerkzeug.

4 Tool for Systematic Annotation of Colonial K'iche'* (TSACK)

Mit dem *Tool for Systematic Annotation of Colonial K'iche'* (TSACK) wurde der Prototyp einer Software-Anwendung entwickelt, welche den Prozess der orthographischen und morphologischen Analyse im Rahmen eines halbautomatisierten Auszeichnungsverfahrens zur XML-basierten Korpuserfassung unterstützt. Das Programm bietet eine graphische Benutzeroberfläche (siehe Abbildung 4), welche die manuelle Eingabe von Auszeichnungen in einem XML-Editor ersetzt und über die Bedienung von Schaltflächen Schritt für Schritt durch den Auszeichnungsprozess leitet.

Hierzu wird der genau transliterierte originale Manuskripttext zunächst in das Programmfenster kopiert. Die Auszeichnung erfolgt durch Anmarkieren von Formen und Bedienung der um das Fenster herum angeordneten Schaltflächen, über die die XML-Tags eingefügt werden. Das Auszeichnungsverfahren, das weiter unten im Detail beschrieben wird, geht von der Originalform aus, die in einzelnen Schritten zunächst in eine phonembasierte, moderne Umschrift transkribiert wird. Die Lemmatisierung und Morphologisierung erfolgt ausschließlich an dieser transkribierten Form. Die Glossen sind in den XML-Tags hinterlegt.

Das Lemmatisierungsverfahren, d.h. die Reduktion flektierter Vollformen auf ihre lexikalischen Grundformen, ist durch die Sprachstruktur vorgegeben. Ein Lemma besteht mindestens aus einer Wortwurzel. Komplexe Lemmata setzen sich aus einer Wortwurzel und mindestens einem Derivationsmorphem zusammen. Das K'iche' weist eine produktive Derivationsmorphologie auf, über die aus lexikalischen Wurzeln andere Wortklassen abgeleitet werden. Derivationsmorpheme im K'iche' sind nahezu ausschließlich suffigierend und abgeleitete Wortstämme können wiederum als Derivationsbasis fungieren. Eine Wortform kann so mehrere Lemmata beinhalten.

Abbildung 2 verdeutlicht das Grundprinzip der Lemmatisierung im K'iche' anhand einer komplexen Form. Die transkribierte Form *ukamisaxik* „seine/ihre Tötung“ lässt sich in das gebundene Possessivpräfix der 3. Person Singular und das Lemma 1 *kamisaxik* „Getötetwerden“ morphologisieren. Lemma 1 ist wiederum eine zusammengesetzte Form aus dem passiven Verbstamm *kamisax* „getötet werden“ (Lemma 2) und der Nominalisierung *-ik*. Das passive Verb ist mit dem Passivmarker *-x* aus dem transitiven Verb *kamisa-* „töten“ (Lemma 3) abgeleitet, das sich wiederum morphologisch in die intransitive Verbwurzel *kam* „sterben“ (Lemma 4) und die Kausativderivation *-isa* unterteilen lässt. In diesem Beispiel ist die Wortwurzel identisch mit dem kleinsten Lemma, was nicht grundsätzlich der Fall ist, da im K'iche' auch lexikalische Wurzelklassen bekannt sind, die allein keine Flexionsmorphologie tragen können und immer mindestens ein Derivationsuffix aufweisen. Das Annotationsverfahren trägt der Komplexität der verschiedenen Lemma-Ebenen Rechnung und bricht Wortformen mit Übersetzung und Glosse bis auf die Wurzel herunter.

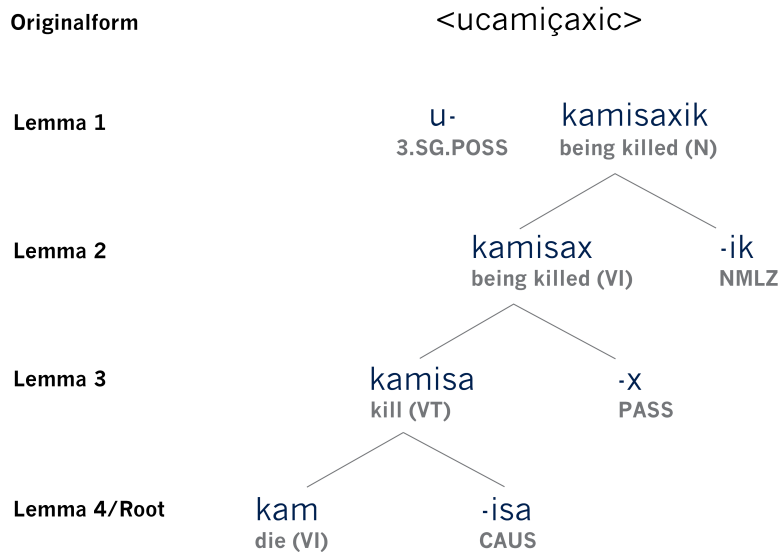


Abbildung 2: Grundprinzip der Lemmatisierung im K'iche'

Die Auszeichnung erfolgt hierarchisch ausgehend vom Haupteintrag (*entry*), der aus mindestens einer spanischen Form (*spanish 1,2, ...*) und mindestens einer korrelierten K'iche'-Form (*k'iche' 1,2, ...*) besteht. Sind mehrere Haupteinträge unter einem Obereintrag (*super entry*) zusammengefasst, kann dies entsprechend markiert werden. Besteht ein K'iche'-Eintrag aus mehr als einem Wort (*word 1, 2, ...*), so werden diese jeweils einzeln ausgezeichnet. Die Lemmatisierung der Wörter erfolgt durch Auszeichnung der grammatischen Affixe (*gram_affix*). Komplexe Lemmata werden durch Auszeichnung der Derivationsmorpheme (*der_affix*) bis auf die Wortwurzel (*root*) heruntergebrochen.

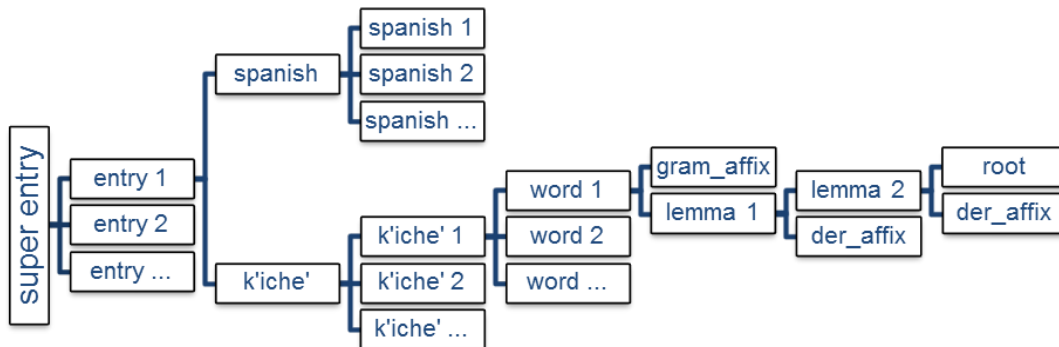


Abbildung 3: Auszeichnungshierarchie

Grammatische und Derivationsmorpheme sind im Programm hinterlegt und werden über Drop-Down-Menüs ausgewählt. Einmal ausgezeichnete Lemmata und Wurzeln werden in eine Liste übernommen und stehen im weiteren Auszeichnungsprozess als Referenz zur Verfügung. Durch diese Vorgaben vereinfacht das Programm den Arbeitsprozess gegenüber herkömmlichen XML-Editoren, reduziert die Fehlerquote bei der Auszeichnung und beschleunigt die Überführung der kolonialen Sprachdaten in ein maschinenlesbares Korpus.

5 Auszeichnungsverfahren

Im Folgenden soll das Auszeichnungsverfahren mit der genauen Funktionsweise des Tools und der einzelnen Schaltflächen anhand verschiedener Beispiele erläutert werden. Die zur Annotation verwendeten XML-Tags sind bisher nur in Teilen TEI-konform; hier müssen in Zukunft bereits bestehende Standards übernommen und nötige Erweiterungen über die Special Interest Group: TEI for Linguists (http://wiki.tei-c.org/index.php/SIG:TEI_for_Linguists³) im Konsortium eingebracht werden. Auch die Definition eines XML-Schemas mit einer genauen Beschreibung der syntaktischen Struktur, welcher die ausgezeichneten Dokumente entsprechen müssen, wurde bislang nicht erstellt.

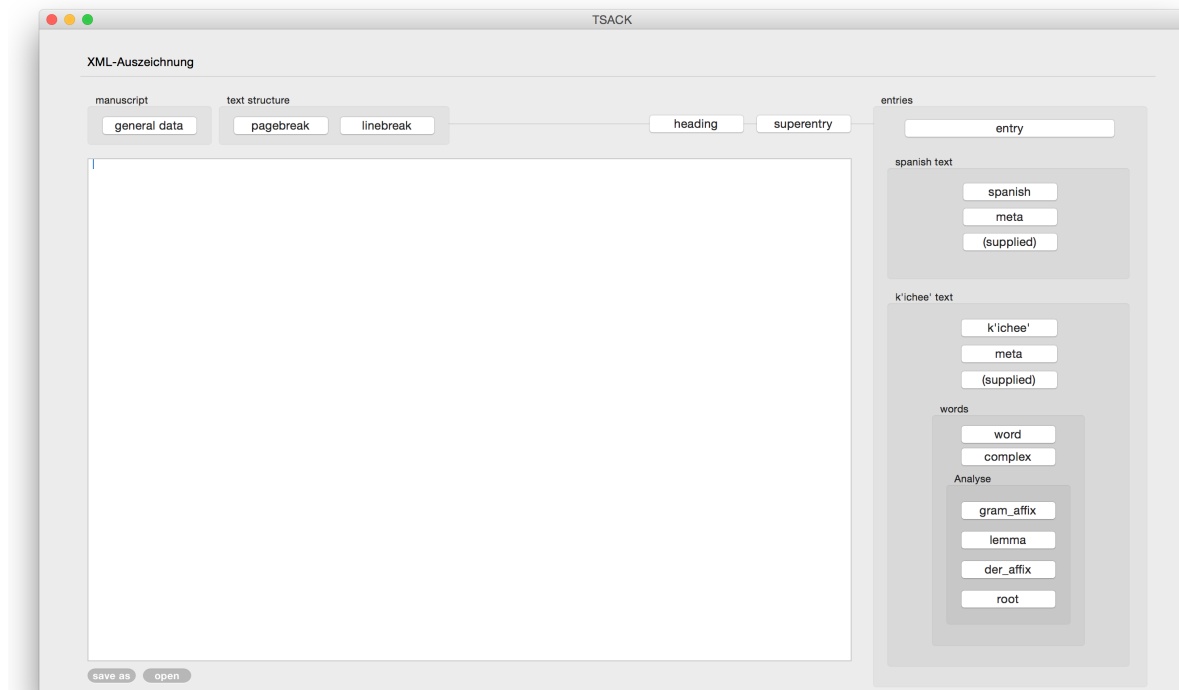


Abbildung 4: Graphische Oberfläche von TSACK

Die graphische Oberfläche des Werkzeugs bietet ein großes Textfenster, in dem der auszuzeichnende Text anmarkiert und durch Bedienen der umliegend angeordneten Schaltflächen ausgezeichnet wird. Allgemeine Metadaten zum Dokument werden über die Schaltfläche *general data* in der oberen Menüleiste eingegeben. Das Menü hält Felder für den Manuskripttitel (*title*), den Autor (*author*), Kopisten (*copyist*), das Jahr des Originals (*year (original)*) und Jahr der vorliegenden Manuskriptkopie (*year (copy)*), die Manuskriptnummer bzw. Signatur des jeweiligen Archivs (*manuscript number*), das Manuskriptformat (*format*), den aktuellen Aufbewahrungsort (*location*) und die Sammlung (*collection*) sowie ein Feld für relevante Zusatzinformationen (*notes*) bereit. Weitere Metadaten können nach Bedarf in dieses Menü integriert werden. Da die aktuellen Aufbewahrungsorte sämtlicher K'ichee'-Wörterbücher bekannt sind,

³http://wiki.tei-c.org/index.php/SIG:TEI_for_Linguists

sind die jeweiligen Bibliotheken und Sammlungen hinterlegt und können über Drop-Down-Menüs ausgewählt werden, um die Fehlerquote bei der Eingabe zu reduzieren.

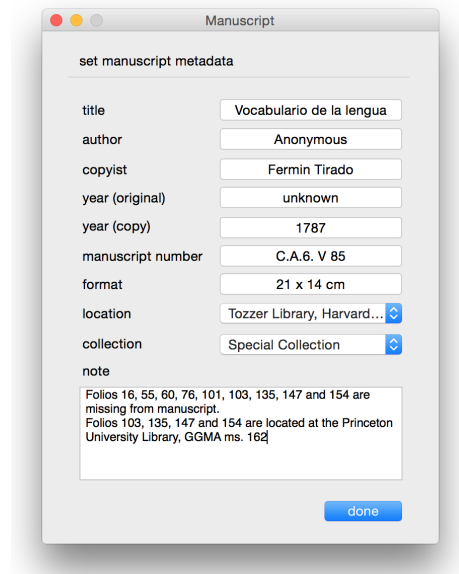


Abbildung 5: Menü zur Eingabe allgemeiner Manuskriptdaten

Die Textstrukturierung erfolgt über Schaltflächen in der oberen Menüleiste, mit denen Tags für Seiten- (<pb>) und Zeilenumbrüche (<lb>) eingefügt werden können. Über eine weitere Schaltfläche lassen sich Überschriften (<head>) auszeichnen.

(9) Annotation von Seiten- und Zeilenumbrüchen sowie Überschriften

```
<pb n="34"/>  
<head>B</head>  
Baçin de Barbero para afeytar. çoca<lb/>  
bal 4hi4h.<lb/>  
Baçin ó servidor. 4izibalboof<lb/>  
Badajo de campana. vbae. campa<lb/>  
na. vel runum. campana.<lb/>  
[...]  
*Balar la oveja. coe chif.<lb/>  
<pb n="35"/>
```

Die nächste Ebene der Textauszeichnung bilden die Wörterbucheinträge, die mittels Annotation über die Schaltflächen im rechten Menüband inhaltlich erschlossen und analysiert werden.

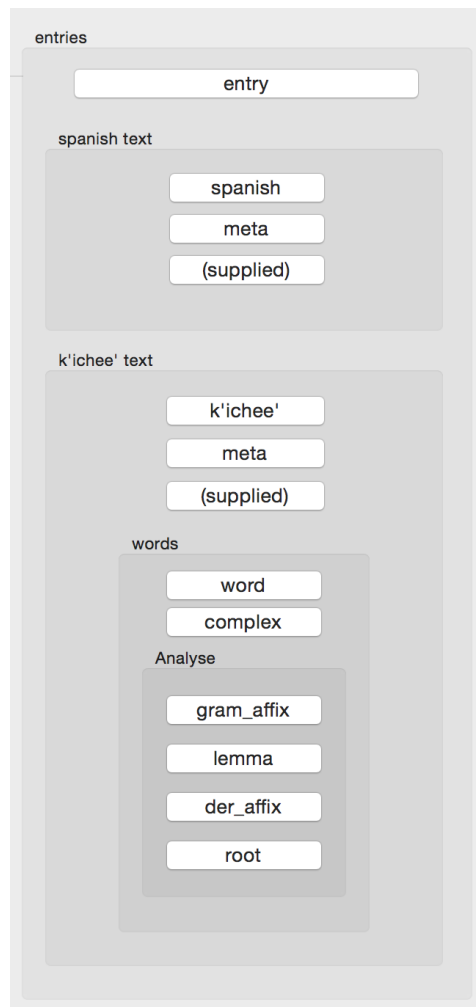


Abbildung 6: Menüband zur Auszeichnung der Wörterbuch-Einträge

Haupteinträge (*entries*) umfassen stets mindestens eine spanische und eine K'ichee'-Form, die miteinander korreliert sind. Die Auszeichnung dieser spanischen und K'ichee'-Untereinträge erfolgt getrennt. Der K'ichee'-Eintrag wird weiter in seine lexikalischen und morphologischen Bestandteile analysiert, wozu jeweils Buttons für die Auszeichnung zur Verfügung stehen. Die Wörter, aus denen ein Eintrag besteht, werden einzeln ausgezeichnet (*word*). Komplexe Wörter werden gesondert gekennzeichnet (*complex*). Hierunter fallen vor allem Komposita, die semantisch von der jeweiligen Bedeutung ihrer Einzelbestandteile abweichen (s.u.). Im nächsten Schritt erfolgt die Auszeichnung grammatischer Marker und die Lemmatisierung entsprechend des oben beschriebenen Verfahrens.

Einfache Haupteinträge bestehen aus einem spanischen Eintrag (*esp_entry*) und einem K'ichee'-Eintrag (*kichee_entry*), wobei der K'ichee'-Eintrag aus einem Wort (10) oder mehreren Wörtern (11) bestehen kann. Die Originalformen von spanischen und K'ichee'-Einträgen erhalten eine ID, über die moderne Verschriftungen und lexikalische Bedeutungen zugeordnet werden; letztere wurden in den nachfolgenden Darstellungen der Übersichtlichkeit halber ausgelassen.

(10) Einfacher Haupteintrag mit K'iche'-Eintrag bestehend aus einem Wort

```
<entry>
  <esp_entry>
    <original_form xml:id="e1">amargura</original_form>
  </esp_entry>
  <kichee_entry>
    <original_form xml:id="w1">4ayil</original_form>
  </kichee_entry>
</entry>
```

(11) Einfacher Haupteintrag mit K'iche'-Eintrag bestehend aus zwei Wörtern

```
<entry>
  <esp_entry>
    <original_form xml:id="e1">Esforsaos</original_form>
  </esp_entry>
  <kichee_entry>
    <word>
      <original_form xml:id="w1">chacourizah</original_form>
    </word>
    <word>
      <original_form xml:id="w2">auib</original_form>
    </word>
  </kichee_entry>
</entry>
```

Bei Wörterbüchern mit K'iche'-Matrix erfolgt die Auszeichnung in umgekehrter Reihenfolge.

(12) Einfacher Haupteintrag mit K'iche'-Matrix

```
<entry>
  <kichee_entry>
    <word>
      <original_form xml:id="w1">Caminak</original_form>
    </word>
  </kichee_entry>
  <esp_entry>
    <original_form xml:id="e1">el Difunto</original_form>
  </esp_entry>
</entry>
```

Komplexere Haupteinträge bestehen aus jeweils mehreren spanischen bzw. K'iche'-Einträgen. Im folgenden Beispiel ist ein spanischer Eintrag mit zwei K'iche'-Einträgen korreliert.

(13) Komplexer Eintrag mit zwei K'iche'-Einträgen

```
<entry>
  <esp_entry>
    <original_form xml:id="e1">Afliccion interior</original_form>
  </esp_entry>
  <kichee_entry>
    <word>
      <original_form xml:id="w1">v4humumic</original_form>
    </word>
  </kichee_entry>
  <kichee_entry>
    <word>
      <original_form xml:id="w2">v&atatic</original_form>
    </word>
    <word>
      <original_form xml:id="w3">4ux</original_form>
    </word>
  </kichee_entry>
</entry>
```

Sind mehrere Untereinträge in einem Eintrag zusammengefasst, kann dieser als Obereintrag (*superEntry*) ausgezeichnet werden. Im folgenden Beispiel (14) sind dem aus den beiden Grundbedeutungen *anochece* „Nacht werden“ und *entrar la noche* „die Nacht tritt ein = dunkel werden“ bestehenden Haupteintrag drei Untereinträge mit den spanischen Übersetzungen *ya va entrando la noche* „die Nacht ist schon dabei einzutreten = es ist dabei dunkel zu werden“, *ya entró la noche* „die Nacht ist schon eingetreten = es ist schon dunkel geworden“ und *entrará la noche* „die Nacht wird eintreten = es wird dunkel werden“, die K'iche'-Spanisch sortiert sind, sowie zwei diskursive Satzbeispiele *primero entrará la noche que vos o tu veías hacer* „es wird Nacht bevor du es wahrnimmst“ sowie *una legua antes del pueblo me anoheció* „eine Legua vor dem Dorf brach über mir die Nacht herein“ zugeordnet.

(14) [Anocheser entrar la noche.]_{spanish_superentry} [Xcoc aεa.]_{k'iche'_1} [ya va entrando la noche.]_{spanish_1} [Xoc aεab.]_{k'iche'---_2} [ya entro la noche.]_{spanish_2} [Xchoc aεab.]_{k'iche'_3} [entrará la noche.]_{spanish_3} [primero entrará la noche q(ue) vos, ó tu veyas acer lo q(ue) te mande.]_{spanish_4} [nabe xhoc niaεab ma4uhi cabeabana ri xinbijfj chaue.]_{k'iche'_4} [una legua antes del Pueblo me anoheció.]_{spanish_5} [hun cuulibal carañ vae tinamit xoc aεab chuech.]_{k'iche'_5} vel. [hun chic thuyulibal maha canriik tinamit xino qui çanaεab]_{k'iche'_5}

(*Vocabulario de la lengua kiché*, fol. 22r-22v)

(15) Auszeichnungsbeispiel eines Obereintrags

```
<superEntry>
  <entry>
    <esp_entry>
      <original_form xml:id="e1">Anocheser</original_form>
    </esp_entry>
    <esp_entry>
      <original_form xml:id="e2">entrar la noche</original_form>
    </esp_entry>
  </entry>
  <entry>
    <kichee_entry>Xcoc aea</kichee_entry>
    <esp_entry>
      <original_form xml:id="e3">ya va entrando la noche</original_form>
    </esp_entry>
  </entry>
  <entry>
    <kichee_entry>Xoc aεab</kichee_entry>
    <meta>.</meta>
    <esp_entry>
      <original_form xml:id="e4">ya entro la noche</original_form>
    </esp_entry>
  </entry>
  <entry>
    <kichee_entry>Xchoc aεab</kichee_entry>
    <esp_entry>
      <original_form xml:id="e5">entrará la noche</original_form>
    </esp_entry>
  </entry>
  <entry>
    <esp_entry>
      <original_form xml:id="e6">primero entrará la noche q(ue) vos, ó tu vevas acer lo q(ue) te mande</ori
    </esp_entry>
    <kichee_entry>nabe xchoc niaεab ma4uhi cabeabana ri xinbijfj chaue</kichee_entry>
  </entry>
  <entry>
    <esp_entry>
      <original_form xml:id="e7">una legua antes del Pueblo me anocheocio</original_form>
    </esp_entry>
    <kichee_entry>hun cuulibal caraƒj vae tinamit xoc aεab chuech</kichee_entry>
    <meta>. vel. </meta>
    <kichee_entry>hun chic thuyulibal maha canriK tinamit xino qui çanaεab</kichee_entry>
  </entry>
</superEntry>
```

Spanische Einträge sind meist uneinheitlich verschriftet. Die Auszeichnung sieht deswegen eine Ebene vor, auf welcher der spanische Originaleintrag in moderne spanische Orthographie nach den Vorgaben der *Real Academia Española* umgesetzt wird. Über das Markieren des spanischen Originaleintrags und Bedienen der Schaltfläche *spanish* wird eine Maske aktiviert, die den Anwender auffordert, unter *transcription* die moderne Verschriftung einzugeben.

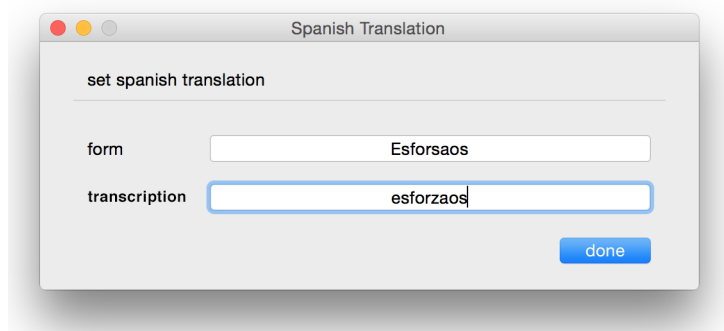


Abbildung 7: Umsetzung spanischer Einträge in moderne Orthographie

Nachfolgend wird Beispiel (11) wieder aufgegriffen, um zu illustrieren wie der Eintrag im Text getaggt wird. Neben der Annotation der Originalform „Esforsaos“ (*original_form*) wird eine modernschriftliche Transliterationsebene „esforzaos“ (*ref target ... „transcription“*) für den spanischen Eintrag eingefügt:

(16) Transkription spanischer Einträge in moderne Orthographie

```
<esp_entry>
  <original_form xml:id="e1">Esforsaos</original_form>
  <ref target="e1" type="transcription">esforzaos</ref>
</esp_entry>
```

Wie das Beispiel *esforzaos* „strengt Euch an“ zeigt, weisen spanische Einträge, insbesondere solche mit diskursiven Satzbeispielen, nicht selten flektierte Wortformen auf, die sinnvollerweise einer Lemmatisierung (17) unterzogen werden müssten, welche das Programm zum gegenwärtigen Zeitpunkt allerdings noch nicht leistet.

(17) Flektierte Wortformen in spanischen Einträgen

```
<esp_entry>
  <original_form xml:id="e1">Esforsaos</original_form>
  <ref target="e1" type="transcription">esforzaos</ref>
  [<ref target="e1" type="root">esforzar</ref>]
</esp_entry>
```

Um den Text in seiner Gesamtheit zu erhalten, werden Satz- und Leerzeichen in spanischen wie K'iche'-Einträgen separat getaggt (*meta*) und Kürzungen im Manuskripttext aufgelöst bzw. ergänzt (*supplied*).

Beispiel 18 illustriert die Auszeichnungsmodalitäten anhand des Eintrags „Grada p(ar)a subir. Cumuc.“ aus dem *Vocabulario de la lengua Quiché* (fol. 55r).

(18) Auszeichnung von Satzzeichen, Leerzeichen und Kürzungen

```
<entry>
  <esp_entry>
    <original_form xml:id="e1">Grada p<supplied>ar</supplied>a subir</original_form>
    <ref target="e1" type="transcription">grada para subir</ref>
  </esp_entry>
  <meta>.</meta>
  <kichee_entry>Cumuc</kichee_entry>
  <meta>.</meta>
</entry>
```

In K'iche'-Einträgen, die aus mehreren Wörtern bzw. ganzen Sätzen bestehen, werden sämtliche Wörter einzeln ausgezeichnet und über den Annotationsprozess lemmatisiert, analysiert und glossiert. Auf der Ebene der Auszeichnung der einzelnen Wörter findet die Umsetzung der kolonialen Verschriftung in eine phonembasierte Orthographie statt. Verwendet wird hier die moderne Standardorthographie der Mayasprachen (ALMG 1988). Durch Markieren des einzelnen K'iche'-Wortes und Betätigen der Schaltfläche *word* wird eine Maske aktiviert, die den Anwender auffordert, die moderne Umschrift einzugeben.

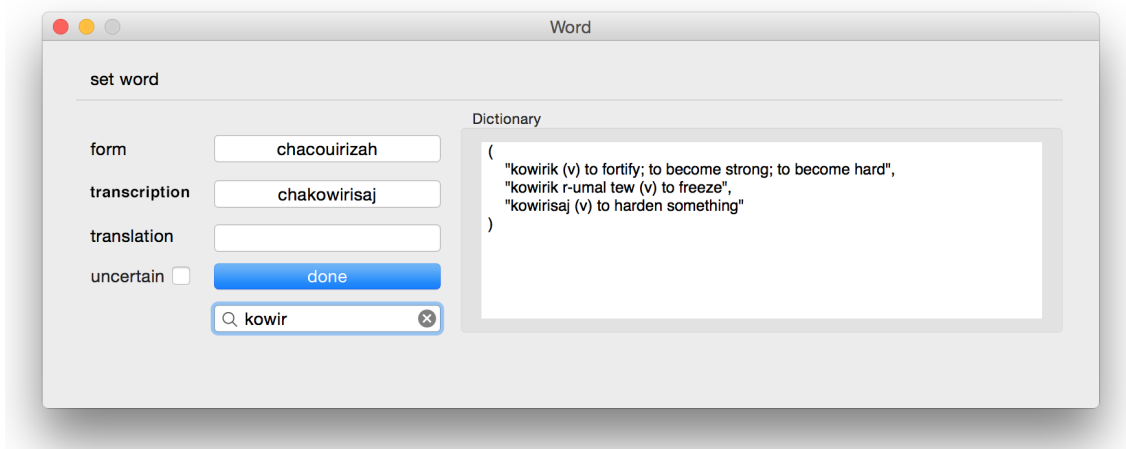


Abbildung 8: Transkription des K'iche'-Originaleintrags in moderne Standardorthographie

Als Referenz sind moderne K'iche'-Wörterbücher eingepflegt, die über ein Suchfenster in der Maske zugänglich und einfach nach Verschriftungsmöglichkeiten durchsucht werden können. Die modernen Formen können allerdings nur eine Tendenz vorgeben (siehe das oben genannte Beispiel *tzaqol tzij* vs. *tz'aqol tzij*). Da die Umsetzung nicht in allen Fällen eindeutig ist, können unsichere Auszeichnungen an dieser Stelle mit Hilfe der Auswahl *uncertain* qualifiziert werden. Die Transkription in die moderne Orthographie

muss revidierbar sein und sinnvollerweise sollten alternative Eingaben möglich sein, was das Programm zum gegebenen Zeitpunkt aber noch nicht leistet.

Die moderne Verschriftung wird in der Annotation in einer Referenzzeile, die sich auf die ID der Originalform bezieht, hinzugefügt.

(19) Transkription des Originaleintrags in moderne Orthographie

```
<word>
  <original_form xml:id="w1">chacourizah</original_form>
  <ref target="w1" type="transcription" status="certain">chakowirisaj</ref>
</word>
```

Lemmatisierung und morphologische Analyse erfolgen ausschließlich an der transkribierten, d.h. modernisierten Form. Im ersten Auszeichnungsschritt werden sämtliche grammatische Affixe markiert und getaggt. Der anmarkierte Teil wird als *form* in einer Maske angezeigt und kann über ein Drop-Down-Menü, in dem sämtliche grammatischen Marker des K'iche' hinterlegt sind, in seiner grammatischen Funktion (*gram_affix function*) annotiert werden. Diese Auswahl muss derzeit noch manuell erfolgen. Die Auszeichnung der grammatischen Affixe erfolgt von links nach rechts. So werden im folgenden Beispiel 20 zuerst der Aspektmarker *ka-*, dann das Absolutivpräfix *ø-*, das Ergativpräfix *nu-* und zum Schluss das Modalsuffix *-o* ausgezeichnet.

(20) Auszeichnungsreihenfolge

ka- ø- nu- koj -o
1 → 2 → 3 → 4

Über die Maske wird eine Auswahl zur Kategorisierung des Affixes als Präfix oder Suffix vorgenommen.

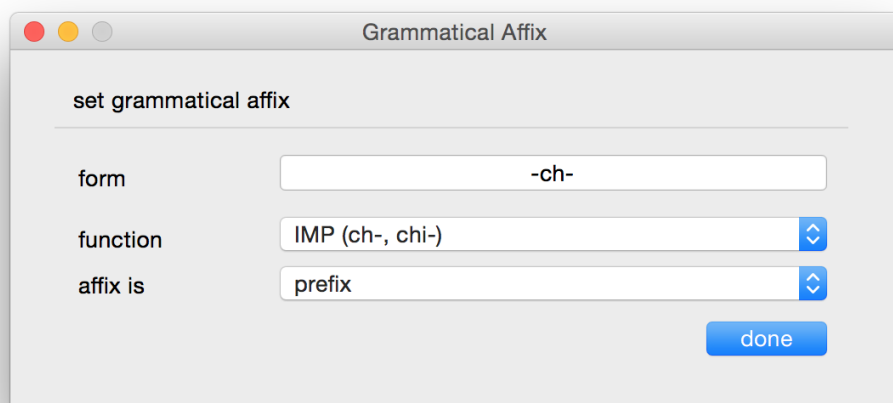


Abbildung 9: Markierung der grammatischen Marker und der Affix-Kategorie

Die vollständige Auszeichnung enthält die grammatische Funktion des Affixes und die Kategorie:

```
<gram_affix function="IMP" affix_is="prefix">ch</gram_affix>
```

Die transkribierte Form wird mittels dieser Tags in ihre morphologischen Bestandteile untergliedert. Am bereits genannten Beispiel <chacourizah> *ch-ø-a-kowirisa-j* „strenge (dich) an“ erfolgt die Auszeichnung wie nachfolgend dargestellt.

(21) Annotation der grammatischen Morpheme

```
<kichee_entry>
  <word>
    <original_form xml:id="w1">chacourizah</original_form>
    <ref target="w1" type="transcription" status="certain">
      <gram_affix function="IMP" affix_is="prefix">ch</gram_affix>
      <gram_affix function="3.SG.ABS" affix_is="prefix">ø</gram_affix>
      <gram_affix function="2.SG.ERG" affix_is="prefix">a</gram_affix>
      <lemma>kowirisa</lemma>
      <gram_affix function="MOD.VTD" affix_is="suffix">j</gram_affix>
    </ref>
  </word>
</kichee_entry>
```

Zur Auszeichnung des Lemmas wird über die entsprechende Menü-Schaltfläche (*lemma*) eine Maske aktiviert, welche die Zuordnung zu einer lexikalischen Klasse, semantischen Domäne und Übersetzung ermöglicht.

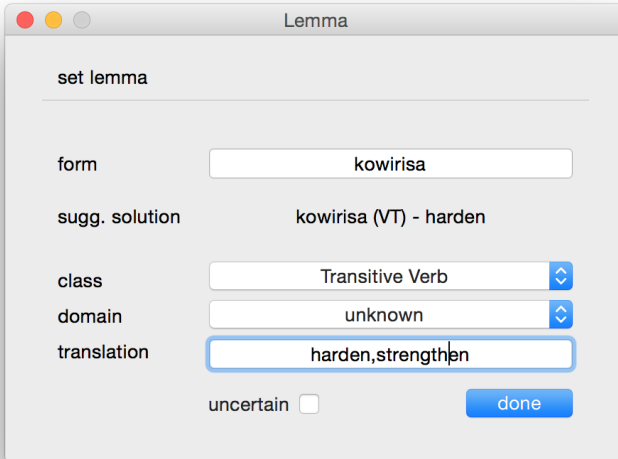


Abbildung 10: Auszeichnung des Lemmas

Das Tagging der Wortklasse (*class*) und Domäne (*domain*) wird aus einem Drop-Down-Menü ausgewählt. In Fällen, in denen die Klasse nicht eindeutig ist, besteht die Möglichkeit der Auswahl *unknown*. Die Übersetzung des Lemmas (*translation*) wird auf Englisch annotiert, wobei mehrere Bedeutungen abgetrennt durch Kommata eingegeben werden können. Lemmata, die einmal eingegeben wurden, werden unter *suggested solution* als Referenzen angezeigt. Unsichere Auszeichnungen können derzeit mithilfe des Auswahlbuttons *uncertain* als solche markiert werden; alle übrigen Auszeichnungen werden automatisch als *certain* getaggt. Bedeutungszuweisungen und *suggested solutions* müssen revidierbar sein, was das Programm derzeit noch nicht unterstützt.

Die vollständige Auszeichnung des Lemmas umfasst ein Tag mit der Lemma-ID (*lemma xml:id*), lexikalischen Klasse und semantischen Domäne. Bedeutungsreferenz und Übersetzungsstatus (*certain/uncertain*) folgen in einer separaten Zeile der morphologischen Annotation; bei mehreren Bedeutungen werden wie im vorliegenden Beispiel mehrere Referenzzeilen angefügt:

```
<lemma xml:id="l1" class="VT" domain="unknown">kowirisa</lemma>
```

(...)

```
<ref target="l1" type="translation" status="certain">harden</ref>
```

```
<ref target="l1" type="translation" status="certain">strengthen</ref>
```

Komplexe Lemmata werden im Auszeichnungsprozess weiter analysiert. Sämtliche Derivationsmorpheme des K'iche' (*der_affix*) sind in einem Drop-Down-Menü hinterlegt und können über eine Maske selektiert werden. Auch hier wird ausgewählt, ob es sich um ein Präfix oder ein Suffix handelt, wobei die Derivationsmorphologie im K'iche' nahezu ausschließlich suffigierend ist und nur zwei Derivationspräfixe existieren.

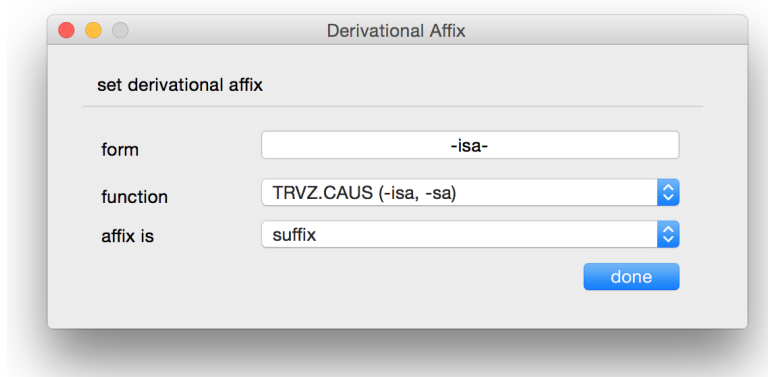


Abbildung 11: Auszeichnung der Derivationsmarker

Das XML-Tag enthält die Funktion des Derivationsmarkers und die Affixklasse und folgt in der Zeile unmittelbar nach dem Lemma.

```
<der_affix function="TRVZ.CAUS" affix_is="suffix">isa</der_affix>
```

Handelt es sich bei dem Bestandteil, an den das Derivationsmorphem affigiert, um ein eigenständiges Lexem, wird auch dieser wiederum als Lemma definiert und über die Schaltfläche *lemma* in der Menüzeile ausgezeichnet. Ist die resultierende Form ebenfalls komplex, wird wiederum das Derivationsmorphem analysiert u.s.f., bis alles auf die nicht weiter segmentierbare Wurzel (*root*) heruntergebrochen ist.

Epenthetische Laute werden dem Lemma bzw. der Wurzel zugeordnet und als *supplied information* gekennzeichnet. Im hier verwendeten Beispiel ist das zugrundeliegende intransitive Verb *kowir* aus der Adjektivwurzel *ko* und dem Inchoativmarker *-ir* gebildet; *w* fungiert als Bindekonsonant.

```
<lemma xml:id="l3" class="ADJ" domain="unknown">
  <supplied>w</supplied>
  <root xml:id="r1" class="ADJ">ko</root>
</lemma>
<der_affix function="INTRVZ.INCH" affix_is="suffix">ir</der_affix>
```

Die Auszeichnung der Wurzel erfolgt über eine eigene Schaltfläche und Maske, über die Bedeutung und lexikalische Klasse eingegeben werden können. Auch die Wortklasse wird über ein Drop-Down-Menü ausgewählt; einmal eingegebene Wurzeln werden als *suggested solution* zur Verfügung gestellt und die Verlässlichkeit der Auszeichnung kann als *uncertain* markiert werden.

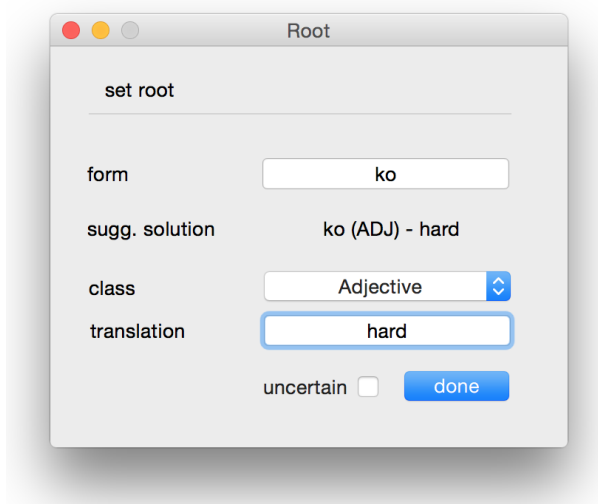


Abbildung 12: Auszeichnung der Wurzel

Die Auszeichnung beinhaltet die Root-ID (*root xml:id*) und Wortklasse der Wurzel; die Grundbedeutung der Wurzel folgt in einer separaten Referenzzeile.

```
<root xml:id="r1" class="ADJ">ko</root>
```

(...)

```
<ref target="r1" type="translation" status="certain">hard</ref>
```

Die vollständige Auszeichnung der Lemmata und der Wurzel des verwendeten Beispiels lässt sich wie folgt darstellen.

(22) Auszeichnung der Lemmata und der Wurzel

```
<kichee_entry>
  <word>
    <original_form xml:id="w1">chacourizah</original_form>
    <ref target="w1" type="transcription" status="certain">
      <gram_affix function="IMP" affix_is="prefix">ch</gram_affix>
      <gram_affix function="3.SG.ABS" affix_is="prefix">ø</gram_affix>
      <gram_affix function="2.SG.ERG" affix_is="prefix">a</gram_affix>
      <lemma xml:id="l1" class="VT" domain="unknown">
        <lemma xml:id="l2" class="VI" domain="unknown">
          <lemma xml:id="l3" class="ADJ" domain="unknown">
            <supplied>w</supplied>
            <root xml:id="r1" class="ADJ">ko</root>
          </lemma>
          <der_affix function="INTRVZ.INCH" affix_is="suffix">ir</der_affix>
        </lemma>
        <der_affix function="TRVZ.CAUS" affix_is="suffix">isa</der_affix>
      </lemma>
      <gram_affix function="MOD.VTD" affix_is="suffix">j</gram_affix>
    </ref>
  </word>
</kichee_entry>
```

Eine vollständige exemplarische Auszeichnung des hier verwendeten Lexikoneintrags ist im Anhang aufgeführt.

6 Ergebnis und Ausblick

Mit dem *Tool for Systematic Annotation of Colonial K'iche'* wird eine speziell für die Mayasprache K'iche' zugeschnittene Lösung zur XML-basierten Korpuserfassung unstandardisiert verschrifteter Wörterbuchtexte vorgelegt. Das Werkzeug integriert die Prozesse der Transkription, Lemmatisierung, morphologischen Analyse, Glossierung und Bedeutungszuschreibung. Es vereinfacht im Vergleich zu gängigen XML-Editoren den Arbeitsprozess, minimiert die Fehlerquote und ermöglicht dadurch die schnellere

Auszeichnung größerer Datenmengen. Das Ergebnis der Auszeichnung ist eine vollständige morphologische Analyse und Glossierung der jeweiligen Lexikoneinträge, deren einzelne Bestandteile somit suchbar sind und für verschiedene Datenbankabfragen verwendet werden können.

Es wird angestrebt, das Tool in der virtuellen Forschungsumgebung TextGrid zu implementieren, um es so für die dezentrale Korpusfassung durch mehrere Anwender nutzbar zu machen. Hierzu muss der derzeit proprietär unter MacOS laufende Prototyp innerhalb der modularen TextGrid-Architektur neu programmiert und weiterentwickelt werden. Dies schließt die Entwicklung eines validen XML-Schemas und die Verwendung TEI-konformer Tags bzw. die Einbringung neuer Tags im TEI-Konsortium ein.

Notwendige Erweiterungen betreffen die Eingabe von alternativen Auszeichnungen und die Revidierbarkeit sämtlicher Annotationsschritte auf allen Ebenen. Darüber hinaus sollen verschiedene Automatisierungsverfahren implementiert werden, die den Auszeichnungsprozess beschleunigen. Das betrifft vor allem eine Zugriffsmöglichkeit auf bereits annotierte Formen, die eine vollständige Auszeichnung einzelner Wörter über einfache Menü-Auswahl erlauben würde. Um die Fehlerquote bei der Auszeichnung zu minimieren, wäre ferner eine Festlegung der Auszeichnungsreihenfolge bzw. Beschränkung der Funktionalität von Auswahl Schaltflächen auf bestimmte Auszeichnungsstufen sinnvoll. Neben der Implementierung und Erweiterung des Werkzeugs sind in einem zweiten Schritt Abfragestrukturen und eine Nutzeroberfläche zu programmieren, über die sich das angelegte Korpus auswerten lässt. Sinnvoll ist ferner die Erweiterung des Verfahrens zur Erfassung kolonialer Textdaten, deren Integration die Disambiguierung lexikalischer Bedeutungen in Abhängigkeit vom syntaktischen und pragmatischen Kontext ermöglichen würde.

Erfassungs- und Abfragestandards sollen so weiterentwickelt werden, dass das Tool auch für die Auszeichnung anderer amerindischer Sprachen nutzbar wird. Für das yukatekische Maya wurde das Auszeichnungsverfahren bereits erfolgreich getestet. Es ist geplant, ein Interface zu schaffen, über das die im Programm hinterlegten sprachspezifischen Annotationen der Flexions- und Derivationsmorphologie durch den Nutzer flexibel angepasst und somit auf die jeweilige Sprache zugeschnitten werden können. Dabei ist zu berücksichtigen, dass die Erfassung von Sprachen mit abweichender Wortbildung, komplexer Morphologie und stammverändernden phonologischen Prozessen weitere Adaptionen erforderlich macht. Ziel ist es, durch Schaffung eines spezialisierten Tools das Angebot digitaler Werkzeuge in der TextGrid-Umgebung zu erweitern und damit TextGrid als Infrastruktur für die nachhaltige digitale Erfassung und systematische Auswertung kolonialer Wörterbücher und unstandardisiert verschrifteter fremdsprachiger Texte zu erschließen.

Besonderes Potential hätte ein solches Tool in der TextGrid-Forschungsinfrastruktur insbesondere für den Aufbau komparativer Korpora zu den verschiedenen Mayasprachen. Konkrete Synergie-Effekte ergeben sich mit dem an der Abteilung für Altamerikanistik durch die Nordrhein-Westfälische Akademie der Wissenschaften geförderten Projekt „Textdatenbank und Wörterbuch des Klassischen Maya“ (www.mayawoerterbuch.de⁴), das TextGrid bereits nutzt. Durch Angleichen der Auszeichnungsstandards könnte innerhalb der TextGrid-Umgebung die Grundlage für die langfristige Kompilation eines komparativen Datenkorpus zu Mayasprachen geschaffen werden. Erste Grundlagen zur Festlegung von Glossierungsstandards wurden bereits in einem gemeinsamen Workshop, der vom 4.-6. September 2014 in Bonn stattfand, erarbeitet (siehe Sachse & Dürr 2016).

⁴<http://www.mayawoerterbuch.de>

7 Literaturverzeichnis

Acuña, René

1983 → Coto

2005 → Basseta

Academia de las Lenguas Mayas de Guatemala (ALMG) (1988): *Lenguas Mayas de Guatemala: Documento de referencia para la pronunciación de los nuevos alfabetos oficiales*. Documento; 1. Guatemala: Instituto Indigenista Nacional.

Anonym (1787): *Vocabulario de lengua kiché compuesto por el apostólico zelo de los m.r.p. Franciscanos de esta Santa Provincia del Dulcísimo Nombre de Jesús del Arzobispado de Guatemala*. Copiado por d. Fermín Joseph Tirado. 218 folios. Tozzer Library, Harvard, Special Collections, ms. C.A.6 V 85.

Anonym (18. Jh.): *Vocabulario de la lengua Quiche Otlatecas*. Ibero-Amerikanisches Institut, Y/2997

Barrera, Francisco (1745): *Abecedario en la lengua que dize qiche hecho por Mr. Francisco Barrera...* 134 folios. Princeton University Library; Garrett-Gates Collection of Mesoamerican Manuscripts, ms. 160.

Basseta, Domingo [1698] (2005): *Vocabulario de la lengua Quiché*. Fuentes para el estudio de la cultura maya; René Acuña (ed.). México: Universidad Nacional Autónoma de México.

Carmack, Robert M. (1973): *Quichean Civilization: The ethnohistoric, ethnographic, and archaeological sources*. Berkeley: University of California Press.

Christenson, Allen J. (2003): *Popol Vuh: The Sacred Book of the Maya*. Winchester/New York: O Books.

Campbell, Lyle (1977): *Quichean Linguistic Prehistory*. University of California Publications in Linguistics; 81. Berkeley, Los Angeles: University of California Press.

Dürr, Michael (1987): *Morphologie, Syntax und Textstrukturen des (Maya-)Quiche des Popol Vuh. Linguistische Beschreibung eines kolonialzeitlichen Dokuments aus dem Hochland von Guatemala*. Mundus Reihe Alt-Amerikanistik; 2. Bonn: Holos.

Dürr, Michael (1994): El Popol Vuh, la obra de Francisco Ximénez y el Título de Totonicapán: aspectos comparativos de grafías y gramática. In *De orbis Hispani linguis litteris historia moribus: Festschrift für Dietrich Briesemeister zum 60. Geburtstag*, Bd. 2; Axel Schönberger und Klaus Zimmermann (eds.): 1153-1165. Frankfurt am Main: Domus Editoria Europaea

Dürr, Michael und Frauke Sachse (Hg.) im Druck *Diccionario k'iche' de Berlín: El Vocabulario en lengua 4iche otlatecas. Edición crítica*. Estudios Indiana, 10. Berlin: Ibero-Amerikanisches Institut / Gebr. Mann Verlag.

Hernández, Esther (2009): Los vocabularios hispano-mayas del siglo XVI. In *Missionary Linguistics IV / Lingüística Misionera IV. Lexicography*. Selected Papers from the Fifth International Conference on Missionary Linguistics, Mérida, Yucatán 14-17 March 2007; Otto Zwartjes, Ramón Arzapalo & Thomas C. Smith-Stark (eds.): 129-150. Amsterdam, Philadelphia: John Benjamins.

Niederehe, Hans-Josef (1994): *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español (BICRES): Desde los comienzos hasta el año 1600*. Amsterdam, Philadelphia: John Benjamins.

Niederehe, Hans-Josef (1999): *Bibliografía cronológica de la lingüística, la gramática y la lexicografía del español (BICRES): Desde el año 1601 hasta el año 1700*. Amsterdam, Philadelphia: John Benjamins.

Sachse, Frauke (2007): *Documentation of Colonial K'ichee' Dictionaries and Grammars*. Report submitted to the Foundation for the Advancement of Mesoamerican Studies, Inc. (FAMSI). <http://www.famsi.org/reports/06009/index.html>

Sachse, Frauke (2009): *Reconstructing the Anonymous Franciscan K'ichee' Dictionary*. *Mexicon* 31(1):10-18

Sachse, Frauke (2015): *Und Gott sprach K'iche': Ein Überblick über die Quellen und Forschungsansätze zur sprachlichen Mission im Hochland von Guatemala*. In *Mesoamerikanistik: Archäologie, Ethnohistorie, Ethnographie und Linguistik. Eine Festschrift der Mesoamerika-Gesellschaft, Hamburg e.V.*; Lars Frühsorge et al. (eds.): 432-467. Aachen: Shaker.

Sachse, Frauke (2016): *The Expression of Christian Concepts in Colonial K'iche' Missionary Texts*. In *La transmisión de conceptos cristianos a las lenguas amerindias: Estudios sobre textos y contextos en la época colonial*; Sabine Dedenbach-Salazar Sáenz (ed.). *Collectanea Instituti Anthropos*. Sankt Augustin: Anthropos.

Sachse, Frauke und Michael Dürr (2016): *Morphological Glossing of Mayan Languages under XML: Preliminary Results*. *Textdatenbank und Wörterbuch des Klassischen Maya, Working Paper*; 4. (www.mayawoerterbuch.de)

Sáenz de Santa María, Carmelo (1985): *Primera parte del Tesoro de las lenguas Cakchiquel, Quiché y Zutuhil, en que las dichas lenguas se traducen a la nuestra, española [1722]*; Francisco de Ximénez. Guatemala: Academia de Geografía e Historia de Guatemala.

Smailus, Ortwin (1989): *Vocabulario en lengua castellana y guatemalteca que se llama Cakchiquel Chi. Análisis gramatical y Lexicológico del Cakchiquel colonial según un antiguo diccionario anónimo Wayasbah*; 14. (3 vols.). Hamburg: Wayasbah.

Smith-Stark, Thomas C. (2009): *Lexicography in New Spain (1492-1611)*. In *Missionary Linguistics IV / Lingüística Misionera IV. Lexicography*. Selected Papers from the Fifth International Conference on Missionary Linguistics, Mérida, Yucatán 14-17 March 2007; Otto Zwartjes, Ramón Arzapalo & Thomas C. Smith-Stark (eds.): 3-82. Amsterdam, Philadelphia: John Benjamins.

Weeks, John M. (1990): *Mesoamerican ethnohistory in United States libraries: reconstruction of the William E. Gates Collection of historical and linguistic manuscripts*. Culver City: Labyrinthos.

8 Abkürzungsverzeichnis linguistischer Glossierungen

1	1. Person
2	2. Person
3	3. Person
ABS	Absolutiv
AGT	Agentiv
CAUS	Kausativ
ERG	Ergativ
IMP	Imperativ
INC	Inkompletiv
MOD	Modal
N	Nomen, Substantiv
NMLZ	Nominalisierung
PASS	Passiv
PL	Plural
POSS	Possessiv
REFL	Reflexiv
SG	Singular
TRVZ	Transitivierung
VT	Transitives Verb
VI	Intransitives Verb
VTD	Deriviertes Transitives Verb

9 Anhang

Vollständiges Annotationsbeispiel des Originaleintrags:

Esforsaos *chacourizah auib*

(Barrera 1745, Seite 77)

Auszeichnung:

```
<entry>
  <esp_entry>
    <original_form xml:id="e1">Esforsaos</original_form>
    <ref target="e1" type="transcription">esforzaos</ref>
  </esp_entry>
  <kichee_entry>
    <word>
      <original_form xml:id="w1">chacourizah</original_form>
      <ref target="w1" type="transliteration" status="certain">
        <gram_affix function="IMP" affix_is="prefix">ch</gram_affix>
        <gram_affix function="3.SG.ABS" affix_is="prefix">ø</gram_affix>
        <gram_affix function="2.SG.ERG" affix_is="prefix">a</gram_affix>
        <lemma xml:id="l1" class="VT" domain="unknown">
          <lemma xml:id="l2" class="VI" domain="unknown">
            <lemma xml:id="l3" class="ADJ" domain="unknown">
              <supplied>w</supplied>
              <root xml:id="r1" class="ADJ">ko</root>
            </lemma>
            <der_affix function="INTRVZ.INCH" affix_is="suffix">ir</der_affix>
          </lemma>
          <der_affix function="TRVZ.CAUS" affix_is="suffix">isa</der_affix>
        </lemma>
        <gram_affix function="(null)" affix_is="suffix">j</gram_affix>
      </ref>
      <ref target="l1" type="translation" status="certain">harden</ref>
      <ref target="l1" type="translation" status="certain">strengthen</ref>
      <ref target="l2" type="translation" status="certain">become hard</ref>
      <ref target="l2" type="translation" status="certain">become strong</ref>
      <ref target="l3" type="translation" status="certain">hard</ref>
      <ref target="r1" type="translation" status="certain">hard</ref>
    </word>
    <word>
      <original_form xml:id="w2">auib</original_form>
      <ref target="w2" type="transcription" status="certain">
        <gram_affix function="2.SG.POSS" affix_is="prefix">aw</gram_affix>
        <lemma xml:id="l3" class="RN" domain="unknown">
          <root xml:id="r2" class="RN">ib'</root>
        </lemma>
      </ref>
      <ref target="l3" type="translation" status="certain">self</ref>
      <ref target="r2" type="translation" status="certain">self</ref>
    </word>
  </kichee_entry>
</entry>
```