Das Swahili-Projekt

In den bisherigen Anleitungen wurden einerseits vorkonfigurierte Projekte zur Verfügung gestellt, andererseits sehr detaillierte schrittweise Anweisungen vorgegeben für die Erstellung der für ein Projekt erforderlichen Komponenten (Sprachkodierung, Datenbank-Typen, Wörterbücher, Texte). Es kommt jetzt die Zeit, da Sie lernen sollen, diese Dinge selbständig und ohne ausführliche Anleitung zu bewerkstelligen. Sie sollen lernen, aus welchen Bestandteilen ein Projekt sinnvollerweise bestehen sollte, welche Eigenschaften diese Komponenten haben können oder müssen, welche logischen und programmtechnischen Abhängigkeiten zwischen ihnen bestehen. Dies soll in Verbindung mit Sprachdaten aus dem Swahili geschehen, einer wichtigen Bantu-Sprache, die in Ostafrika, vor allem in Tansania, aber auch in Kenia, Somalia und Uganda gesprochen wird.[1]

© National Maritime Museum, London

Wir erstellen zunächst ein neues Projekt und beginnen dann mit der Sprachkodierung für das Swahili, weil diese einerseits unabhängig von den anderen Komponenten ist, andererseits aber von diesen vorausgesetzt wird.

Das phonologische System des Swahili

Die Erstellung einer Sprachkodierung setzt natürlich voraus, dass bereits eine phonologische Analyse und Beschreibung der zur Diskussion stehenden Sprache vorliegt.

Das Phonemsystem des Swahili ist einigermaßen komplex und sieht folgendermaßen aus (HINNEBUSCH 1992:101):[2]

Konsonanten:

	Labial	Interdental	Alveolar	Palatal	Velar	Glottal
Okklusive:						
Stimmlos	p		t	tʃ	k	
Aspiriert	p^h		t^h	$tʃ^h$	k^h	
Implosiv	ɓ		ɗ	ʄ	ɠ	
Pränasaliert	^{m}b		^{n}d	^{ɲ}ʄ[dʒ]	^{ŋ}g	
Frikative:						
stimmlos	f	θ	s	ʃ	x	h
stimmhaft	v	ð	z		ɣ	
Nasale:	m		n	ɲ	ŋ	
Lateral:			l			
Vibrant:			r			
Gleitlaute:	w			j		

Die Frikative /θ, ð, x, ɣ/kommen in Wörtern arabischen Ursprungs vor.
Es gibt augenscheinlich einen phonemischen Kontrast zwischen stimmlosen aspirierten und nicht-aspirierten Okklusiven: *paa* [paː] 'Dach' vs. *paa* [p^haː] 'Gazelle'. Dies gilt allerdings nicht für alle Dialekte und wird in der Orthographie nicht zum Ausdruck gebracht. In Hinnebusch (1992:101) findet sich auch der Hinweis, dass die aspirierten Okklusive durch einen synchronischen morphophonemischen Prozess aus einer zugrunde

[1] Nähere Angaben zur Verbreitung des Swahili finden Sie unter folgender Adresse: http://www.ethnologue.com/show_language.asp?code=swh

[2] Thomas J. Hinnebusch, Swahili. In: William Bright (ed.), *International Encyclopedia of Linguistics*, vol. 4, 99-106, Oxford University Press: London – New York, 1992

liegenden Folge von Nasal + stimmlosem Okklusiv abgeleitet werden können: NC̥ → NC̥^{h} → C̥^{h}.
Die pränasalierten stimmhaften Okklusive können als phonetische Realisierung von Phonemsequenzen aus Nasal und stimmhaftem Okklusiv aufgefasst werden: /mb/, /nd/, /ɲdʒ/ und /ŋg/. Unter dieser Annahme sind die Implosive und die stimmhaften pulmonischen Okklusive komplementär verteilt, wobei die Implosive die phonetische Norm darstellen: /b/ [ɓ, b] – /d/ [ɗ, d] – /dʒ/ [ʄ,dʒ] – /g/ [ɠ, g]. Dadurch vereinfacht sich das Konsonantensystem wie folgt:

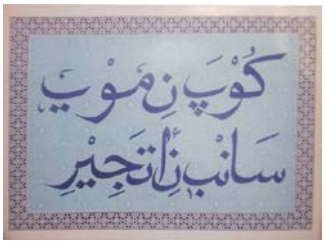
	Labial	Interdental	Alveolar	Palatal	Velar	Glottal
Okklusive:						
Stimmlos	p		t	tʃ ⟨ch⟩	k	
Stimmhaft	b		d	dʒ ⟨j⟩	g	
Frikative:						
stimmlos	f	θ ⟨th⟩	s	ʃ ⟨sh⟩	x ⟨kh⟩	h
stimmhaft	v	ð ⟨dh⟩	z		ɣ	
Nasale:	m		n	ɲ ⟨ny⟩	ŋ ⟨ng'⟩	
Lateral:			l			
Vibrant:			r			
Gleitlaute:	w			j ⟨y⟩		

Wo die orthographische Repräsentation von den phonetischen Symbolen abweicht, wurde sie in ⟨ ⟩ angegeben. Der Apostroph in ⟨ng'⟩ dient dazu, die Wiedergabe des Phonems /ŋ/ von der Phonemfolge /ŋg/ zu unterscheiden: /ŋombe/ ⟨ng'ombe⟩ 'Vieh' vs. /ŋgoma/ ⟨ngoma⟩ 'Tanz'.

Vokale (alle kurz):

/a/, /e/ [ɛ], /i/, /o/ [ɔ], /u/

Der Wortakzent liegt immer auf der vorletzten Silbe. Swahili hat keine Töne, was für eine Bantusprache ungewöhnlich ist.



Swahili wurde ursprünglich in modifizierter arabischer Schrift geschrieben, und zwar schon seit dem 13. Jhd. Mehr als fünf Jahrhunderte lang war diese Schrift in den schriftlichen Dokumenten des Swahili – hauptsächlich Lieder und Gedichte – dominierend. Mit Beginn des 19. Jhd. wurde von europäischen Reisenden und Missionaren, die nach Ostafrika kamen und Swahili lernten, das lateinische Alphabet eingeführt (Romanisierung). Sie erstellten auf dieser Grundlage Swahili-Wortlisten, Grammatiken, Lehrbücher und Lexika.

Es gab im Laufe der Zeit verschiedene Varianten der Romanisierung. Die jetzige Form verwendet alle Buchstaben des Alphabets mit Ausnahme von *q* und *x*. Der Buchstabe *c* kommt nur im Digraphen *ch* vor.

Da das Phonemsystem des Swahili mehr Phoneme aufweist, als das Alphabet Buchstaben hat, mussten Konventionen eingeführt werden, um die zusätzlichen Phoneme orthographisch zu repräsentieren. Dazu werden Digraphe (Kombinationen aus zwei Buchstaben) eingesetzt, wie sie aus der englischen Orthographie bekannt sind, z.B. *ch* für die Affrikate /tʃ/ oder *sh* für den Palatoalveolar /ʃ/.

Swahili verwendet folgende Vokalbuchstaben (jeweils mit Groß- und Kleinschreibung), wobei die Aussprache jener der IPA-Symbole entspricht:

A, a – E, e – I, i – O, o – U, u

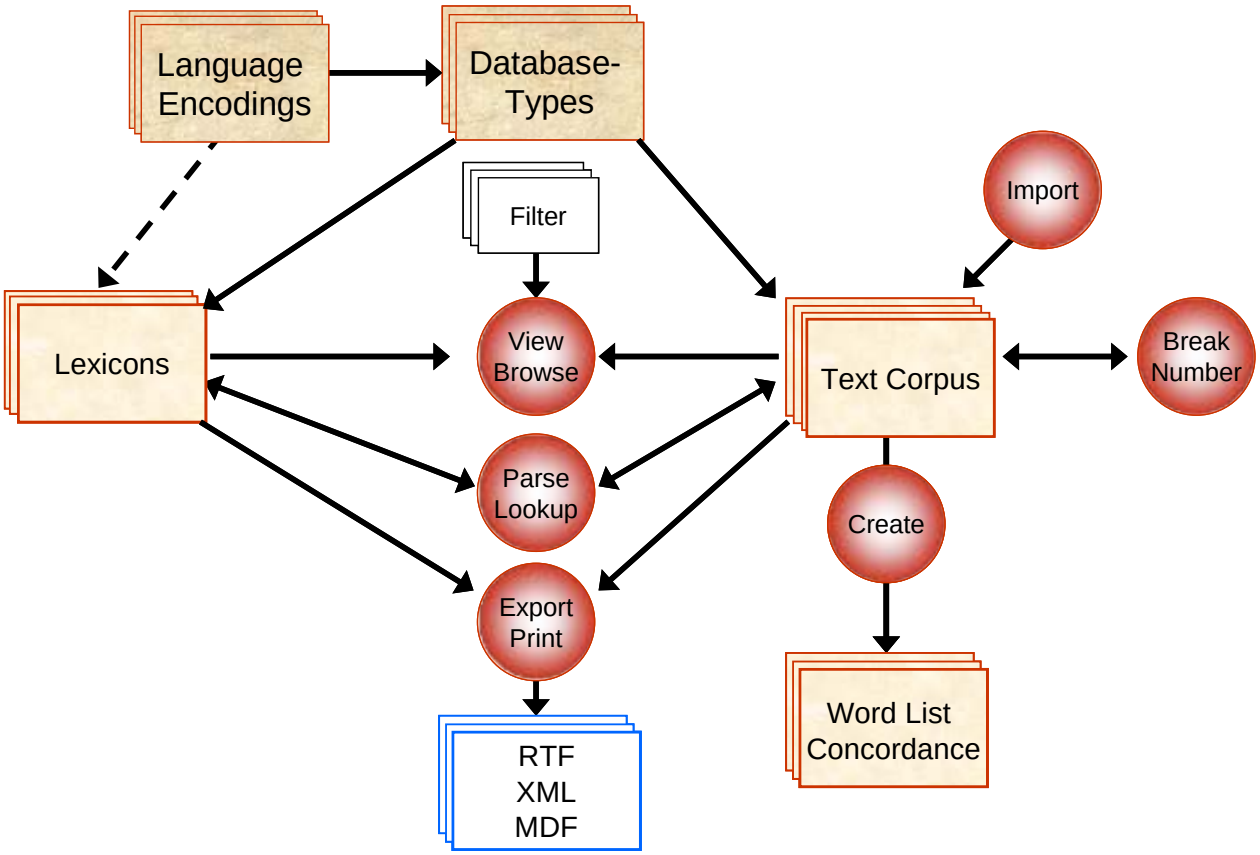
Swahili hat 26 Konsonanten (genauer 24 Konsonanten und zwei Gleitlaute oder Halbvokale: j, w), wobei einige nur in arabischen Lehnwörtern vorkommen. Diese werden wie folgt kodiert:

B, b –Ch, ch /tʃ/ – D, d – Dh, dh /ð/ – F, f – G, g – Gh, gh /ɣ/ – H, h – J, j /dʒ/ – K, k – Kh, kh /x/, L, l – M, m – N, n– Ng', ng' /ŋ/ – Ny, ny /ɲ/ – P, p – R, r – S, s – Sh, sh /ʃ/ – T, t – Th, th /θ/ – W, w /w/ – V, v – Y, y /j/ – Z, z /z/.

Die Buchstaben Q, q, und X, x kommen im Swahili **nicht** vor.

Erstellung eines neuen Projekts

Ein Toolbox-Projekt besteht aus einer Reihe von Komponenten, die jeweils ihren spezifischen Beitrag für die Lösung der zu bearbeitenden Aufgabe leisten. Das folgende Diagramm zeigt den typischen Aufbau eines solchen Projektes und die Abhängigkeiten und Wechselbeziehungen, die zwischen den Komponenten bestehen. Die relevanten Einstellungen (engl. *settings*) eines Projektes werden in eine Datei mit der Erweiterung **.prj** (Projektdatei)



vermerkt.

Wir wollen bei der Anlage des Swahili-Projekt wirklich von Null anfangen und schrittweise die für dieses Projekt erforderlichen Komponenten definieren und damit die entsprechenden Dateien erzeugen, und zwar so, dass diese Herangehensweise auch auf andere Projekte übertragbar ist. Alle zu einem Projekt gehörenden Dateien sollten in einem gemeinsamen Verzeichnis liegen, es sei denn, es handelt sich um Dateien, die auch in anderen Projekten ohne Änderung verwendbar sind.

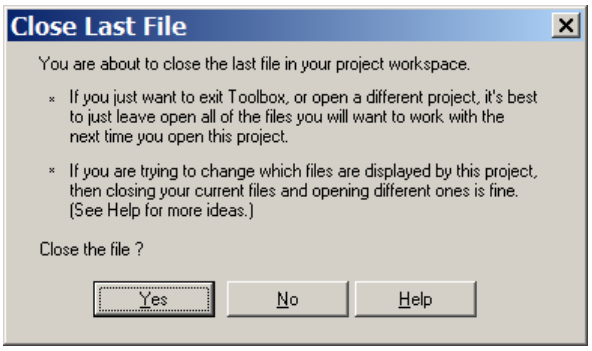
Wir beginnen also mit der Anlage der Projektdatei. Dabei sind zwei Entscheidungen zu treffen:

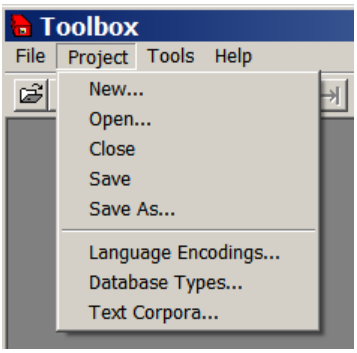
1. Wo soll die Datei auf dem Speichermedium (z.B. Festplatte) liegen?

2. Wie soll die Datei heißen?

Bisher haben wir zunächst mit dem Dateimanager *Windows Explorer* die Verzeichnisstruktur festgelegt. Man kann diese jedoch auch vom Programm *Toolbox* aus vornehmen, und diesen Weg wollen wir im Folgenden einschlagen. Gehen Sie dazu folgendermaßen vor:

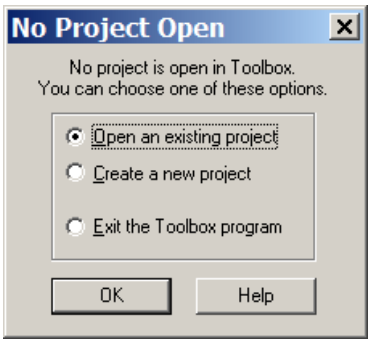
☐ Starten Sie das Programm *Toolbox*. Es meldet sich mit den Fenstern, die bei der letzten Bearbeitung geöffnet waren. Schließen Sie zunächst alle Fenster. Beim Versuch, das letzte Fenster zu schließen, erscheint die folgende Meldung:



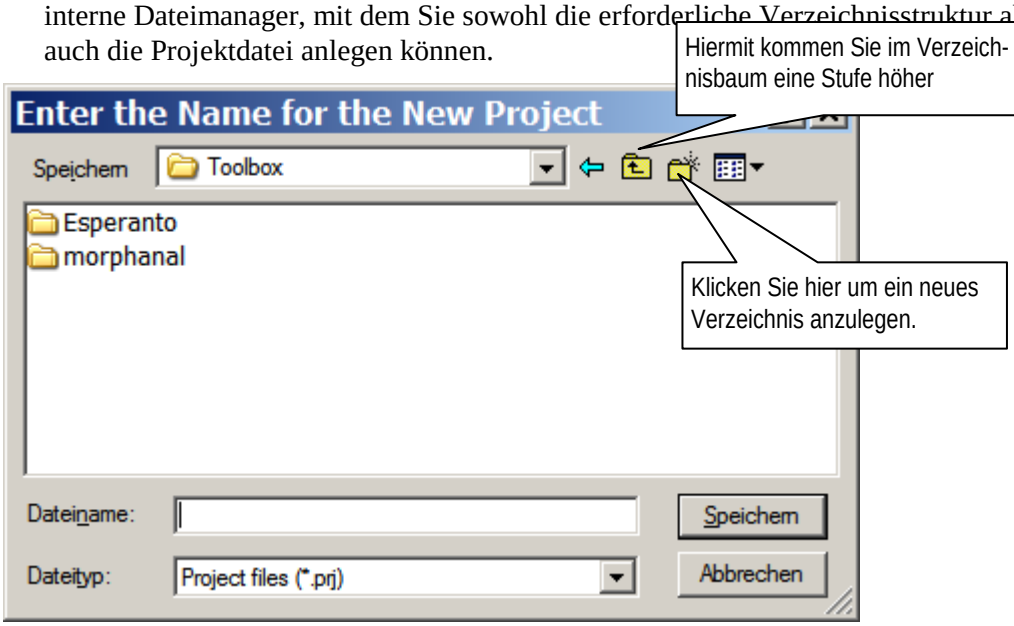


Wählen Sie Yes, so dass nur noch die leere Arbeitsfläche zu sehen ist. Sehen Sie sich an, welche Menüpunkte im Programm jetzt eigentlich zur Verfügung stehen. Im Menü Project sind es einerseits die Punkte New ... Open ... Close ... Save ... Save As ..., andererseits die Optionen Language Encodings ... Database Types ... Text Corpora ... Sie können mit der ersten Gruppe ein neues Projekt anlegen, ein bereits existierendes Projekt öffnen, ein geöffnetes Projekt schließen, sichern oder unter einem anderen Namen speichern.

☐ Wenn Sie alle Fenster geschlossen haben, ist immer noch das Projekt als solches geöffnet. Wählen Sie daher Project >> Close um auch dieses zu schließen.[3] Sie kommen damit in den absoluten Anfangszustand, der zu der nebenstehenden Meldung führt. Sie können in diesem Zustand nur ein bereits existierendes Projekt öffnen, ein neues Projekt anlegen oder das Programm ganz verlassen.

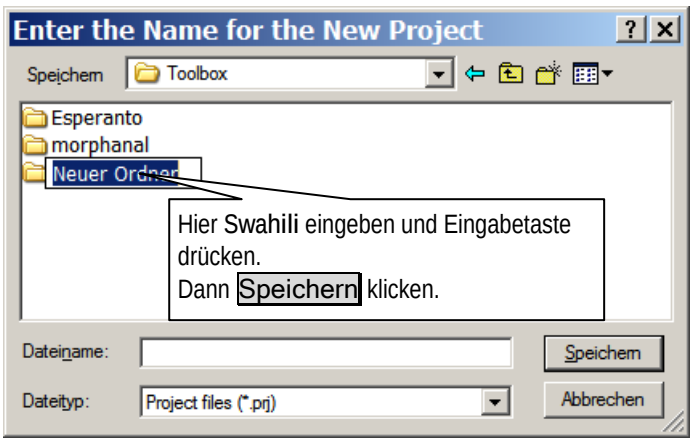


☐ Wählen Sie hier die Option Create a new project. Es meldet sich jetzt der programminterne Dateimanager, mit dem Sie sowohl die erforderliche Verzeichnisstruktur als auch die Projektdatei anlegen können.

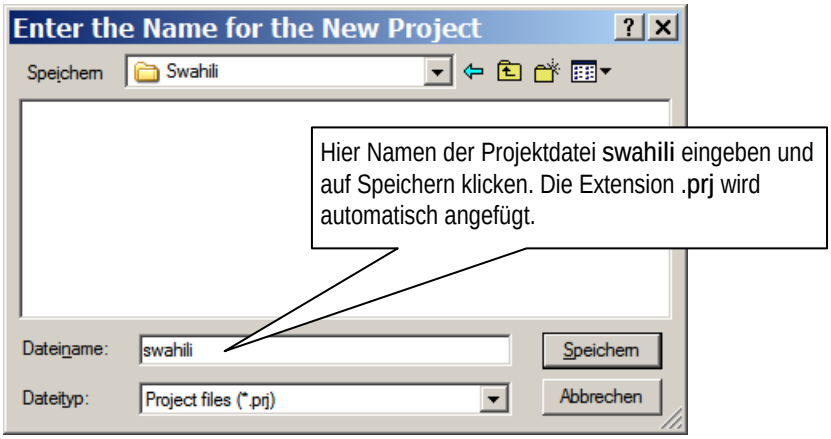


Navigieren Sie zum Verzeichnis Z:\Toolbox – falls Sie sich nicht ohnehin schon dort befinden, indem Sie Verzeichnisse öffnen oder im Verzeichnisbaum eine Stufe nach oben wandern.

☐ Legen Sie hier ein neues Unterverzeichnis Swahili an, indem Sie auf das Symbol klicken und den entsprechenden Namen eingeben.

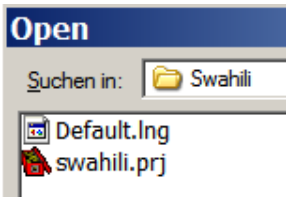


[3] Sie müssen ein geöffnetes Projekt nicht schließen, bevor Sie neues Projekt anlegen. Sobald Sie ein anderes Projekt öffnen oder ein neues anlegen, wird das geöffnete Projekt geschlossen. Es soll hier nur gezeigt werden, wie der absolute Anfangszustand aussieht.



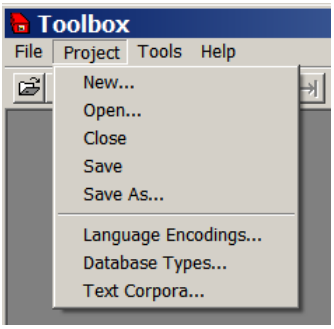
Bereitstellung des Swahili Zeichensatzes

Wenn Sie jetzt den Menüpunkt File >> Open wählen, werden Sie feststellen, dass neben der Projektdatei swahili.prj eine weitere Datei mit dem Namen Default.lng angelegt worden ist. In Dateien mit der Extension .lng werden so genannte **Sprachkodierungen** (*language encodings*) definiert.

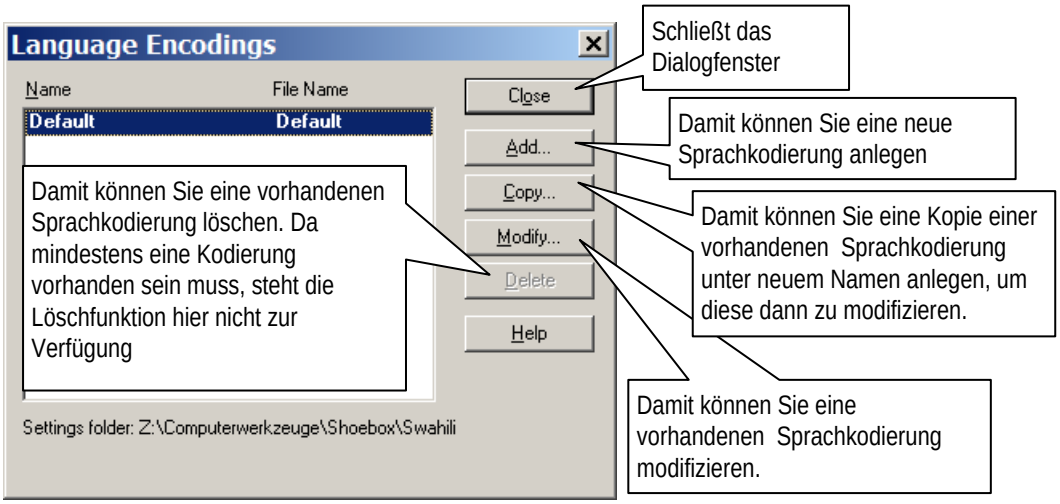


Jedes Projekt benötigt mindestens eine solche Sprachkodierung, die vom Programm automatisch erzeugt wird und auf dem englischen Alphabet und seinen Eigenschaften basiert.

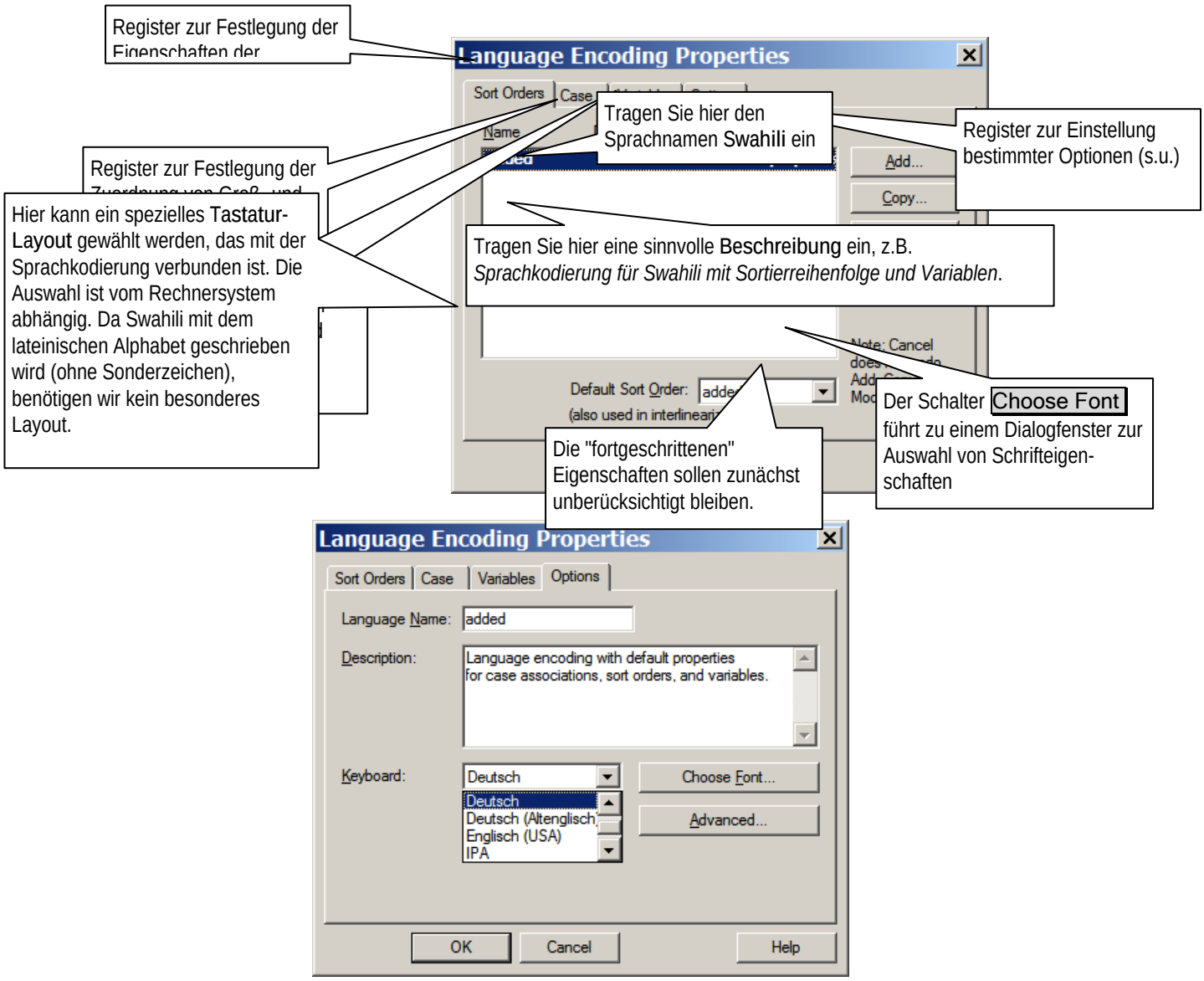
Im folgenden wird es darum gehen, eine Sprachkodierung für das Swahili anzulegen und dabei die wesentlichen Eigenschaften einer solchen Sprachkodierung kennen zu lernen. In einem sonst leeren Projekt stehen im Projektmenü neben den Projektbefehlen im engeren Sinn (New ... – Open ... – Close ... – Save – Save As …) die Menüpunkte Language Encodings ... Database Types ... und Text Corpora ... zur Verfügung.



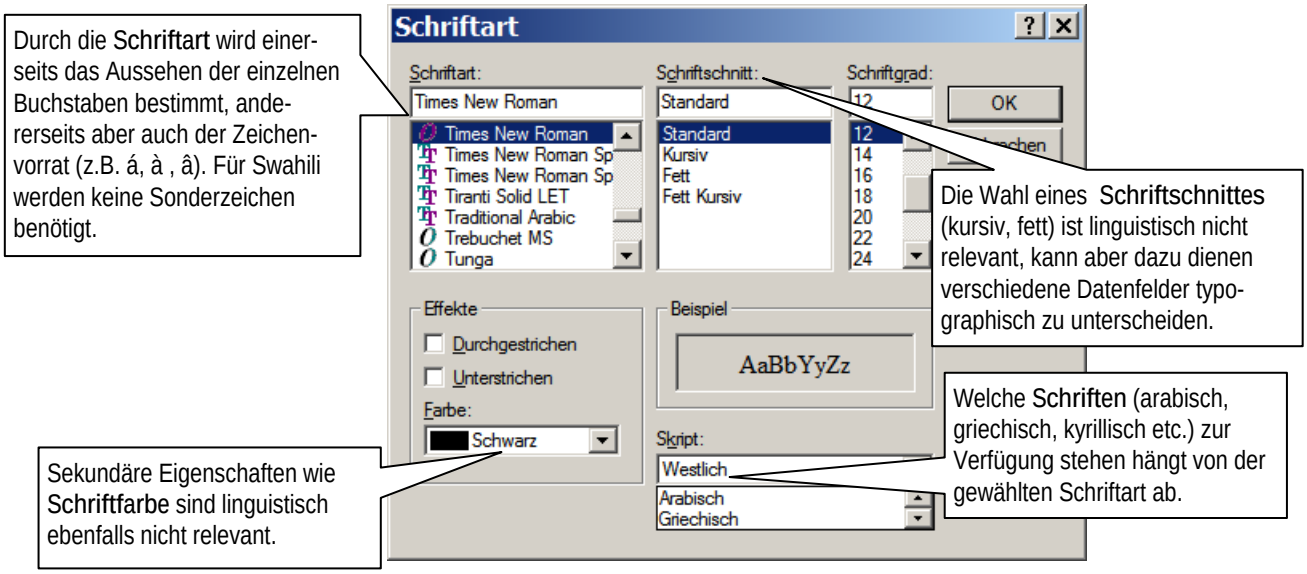
☐ Wählen Sie den Menü-Befehl Project >> Language Encodings ... Es öffnet sich ein sog. Dialogfenster, das die im Projekt verfügbaren Sprachkodierungen auflistet, wobei in der linken Spalte der – frei wählbare – programminterne Name für die Sprachkodierung steht und in der rechten der Dateiname. Es gibt eine Reihe von Operationen, die an einer gewählten Sprachkodierung durchgeführt werden können.



- Da wir eine neue Sprachkodierung für Swahili anlegen wollen, müssen Sie [Add] wählen. Sie sehen dann das Dialogfenster für die Festlegung der Eigenschaften der jeweiligen Sprachkodierung. Ein solches Dialogfenster ist wie eine Kartei aus mehreren Registern angelegt.
- Wählen Sie zunächst das Register Options.



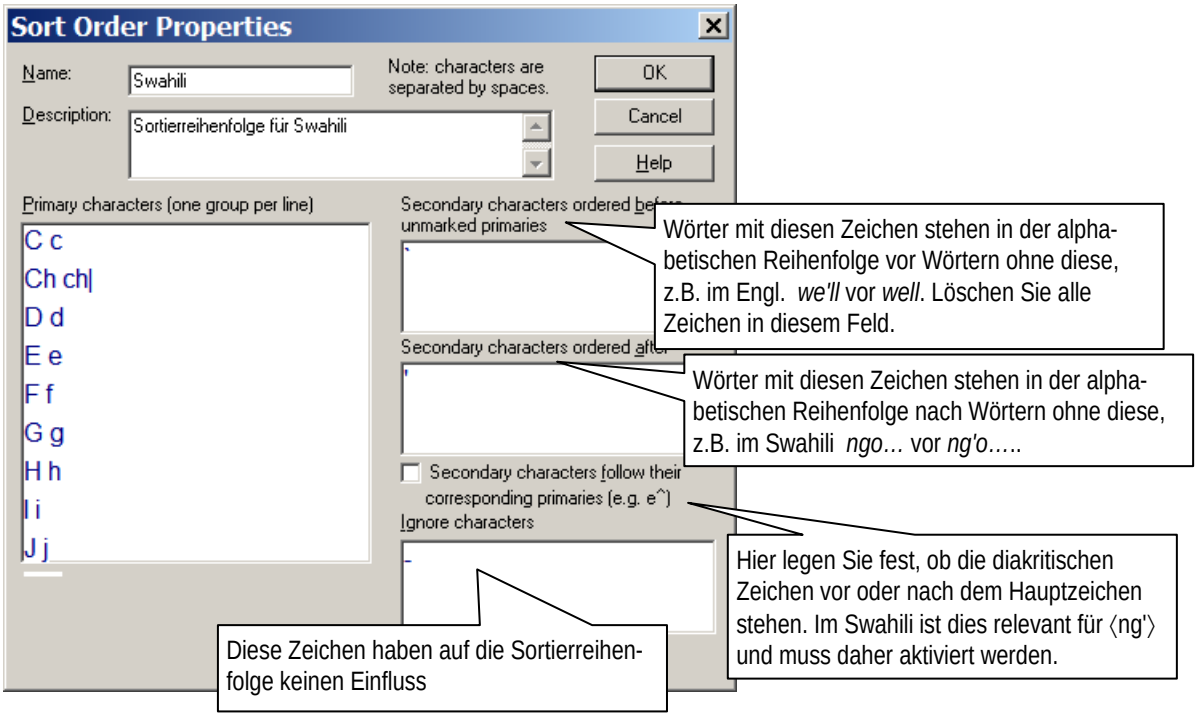
☐ Wählen Sie jetzt den Schalter Choose Font ... . Hier können Sie die Eigenschaften der Schrift festlegen, mit der Texte in der fraglichen Sprache geschrieben werden.



☐ Wählen Sie die Schriftart Arial und die Schriftfarbe Marineblau um Swahili-Text von anderen Texten zu unterscheiden. Beide sind linguistisch ohne Belang. Schließen Sie die Festlegung der Schrifteigenschaften mit OK ab.

Sortierreihenfolge

☐ Wählen Sie als nächstes das Register Sort Orders. Beim Anlegen einer neuen Sprachkodierung wird automatisch eine Kopie der *Default language encoding* zugrunde gelegt. Das gilt auch für die Sortierreihenfolge. Um diese der jeweiligen Sprache anzupassen, wählen Sie den Schalter Modify. Das führt Sie zu einem Dialogfenster Sort Order Properties. Sie werden feststellen, dass dort die Zeichensatz-Eigenschaften (Schriftart und Schriftfarbe) bereits angezeigt werden. Tragen Sie dort den Namen Swahili und eine geeignete Beschreibung ein. Modifizieren Sie dann die Liste der Primärzeichen (*primary characters*) nach den Informationen auf den Seiten 1–3 dieses Textes, indem Sie nicht verwendete Buchstaben entfernen (q, x) und die Digraphe (⟨ch, th, dh, gh⟩) etc. hinzufügen. (Leider arbeitet das Programm hier nicht ganz fehlerfrei. Um die Probleme zu umgehen, empfiehlt es, sich mit der Tast Pos1 an den Anfang der Liste zu gehen und dann mit den Pfeiltasten zu navigieren.



Verlassen Sie nach Abschluss der Bearbeitung das Dialogfenster mit OK.

Primary characters sind die normalen Buchstaben (⟨a, b, c⟩ etc.), *secondary characters* sind diakritische Zeichen wie Akzentzeichen oder der Apostroph. Diese spielen eine Rolle, wenn für bestimmte Sonderzeichen Zeichen im Zeichensatz nicht zur Verfügung stehen. So könnte man z.B. das Zeichen â durch die Zeichenfolge a^ oder ^a darstellen.

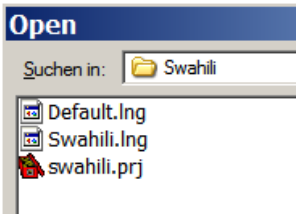
Groß- und Kleinschreibung (case)

Legen Sie als nächstes im Register Case die Zuordnung von Groß- und Kleinbuchstaben fest. Diese ist für das Swahili praktisch identisch mit der Sortierreihenfolge (ohne Ziffern).

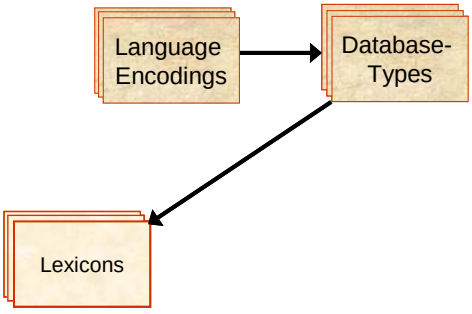
Zeichenklassen

Den Abschluss bilden die Zeichenklassen im Register Variables. Auch hier müssen die fehlenden Digraphe (⟨Ch, ch, Dh, dh⟩ etc.) ergänzt und die Zeichen ⟨q, x⟩ entfernt werden.

Nach Fertigstellung der Sprachkodierung ist zu unserem Projekt eine weitere Datei hinzugekommen. Es besteht jetzt aus zwei Sprachkodierungen (Default und Swahili) und der Projektdatei.



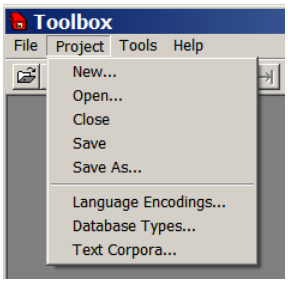
Einrichten einer Lexikon-Datenbank



Für unser Swahili-Projekt benötigen wir als weitere wesentliche Komponente mindestens eine Lexikondatenbank. Die Eigenschaften einer solchen Datenbank werden in einem Datenbanktyp (*Database Type*) definiert, der ebenfalls als Datei – mit der Extension .typ abgespeichert wird. Für bestimmte Felder kommen dabei die Sprachkodierungen zum Tragen.

Der Datenbank-Typ

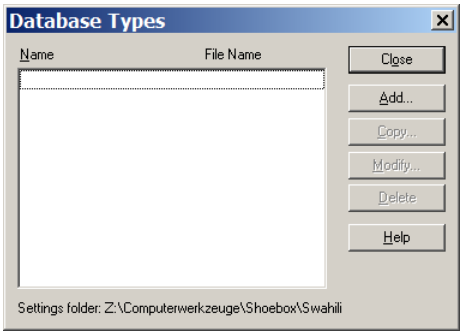
Bisher haben wir Datenbanktypen meist parallel zur Erstellung einer Datenbank definiert. Diese sind aber – als Typ – eigentlich unabhängig von einer spezifischen Ausprägung (Instanz). In der Tat kann man einen Datenbanktyp auch unabhängig von einer Instanz definieren.



Die Befehlsoptionen im Projekt-Menü sind nach wie vor wie nebenstehend angezeigt.

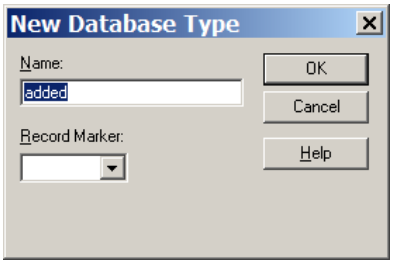
☐ Wählen Sie den Menüpunkt Database Types ...

Das sich öffnende Dialogfenster mit dem Titel Database Types ist analog zu dem für die Sprachkodierungen aufgebaut. Es listet normalerweise die verfügbaren Datenbanktypen auf und stellt entsprechende Operationen zur Verfügung: Add, Copy, Modify, Delete … Da wir für das Swahili-Projekt bisher aber keine Datenbanken bzw. Datenbanktypen erstellt haben, ist dieses Fenster leer und die einzigen verfügbaren Operationen sind das Schließen des Fensters (Close) und das Hinzufügen eines neuen Datenbanktyps (Add).

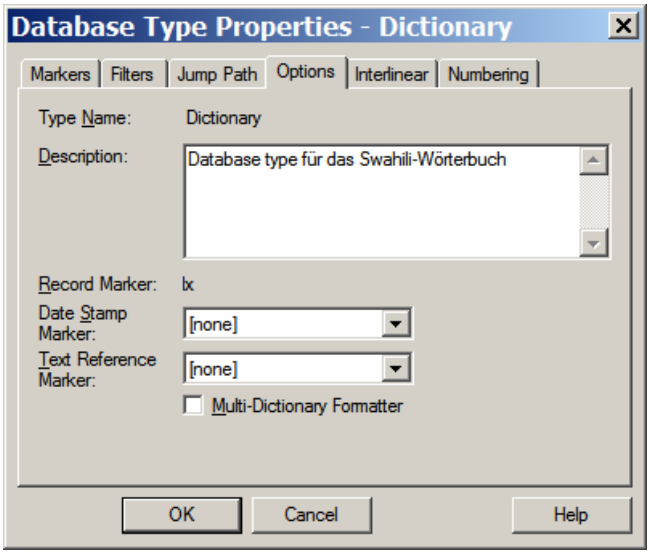


☐ Wählen Sie Add...
☐ Tragen Sie als Namen *Dictionary* ein
☐ Spezifizieren Sie *lx* (= Lemma) als Datensatz-Markierung (*Record Marker*).

☐ Wählen Sie OK.



Es erscheint ein großes Dialogfenster mit sechs Registern und der Aufschrift *Database Type Properties - Dictionary* (Datenbank-Typ-Eigenschaften). Hier können alle für eine Datenbank eines bestimmten Typs relevanten Eigenschaften festgelegt werden. Wir befinden uns im Register *Options*. Das Programm fordert uns damit auf, eine Beschreibung (*Description*) des Datenbank-Typs zu erstellen.[4]



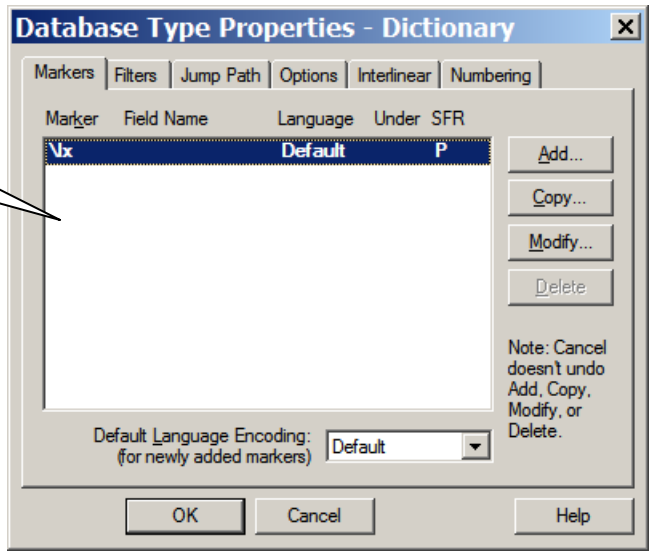
☐ Geben Sie den in der Graphik angezeigten Text als Beschreibung ein.

[4] Diese Information kann nützlich sein, wenn man nach einer längeren Unterbrechung seine Arbeit wieder aufnehmen will, oder wenn andere mit den Daten arbeiten wollen.

(Der *date stamp marker* kann jetzt noch nicht spezifiziert werden, weil bis jetzt nur die Datensatz-Markierung (*record marker*) definiert worden ist. Wir werden später hierher zurückkommen, nachdem einige weitere Markierungen definiert worden sind.)

☐ Wählen Sie das Register Markers.

Liste der Marker mit fünf Spalten: **Marker** (hier: \lx), **Field Name** (noch zu bestimmen), **Language** (Sprachkodierung), **Under** (Position in der Markerhierarchie), **SFR** (Style, Font, Rangeset)



☐ Wählen Sie Modify...

Hier wird die Kurzform des Markers eingetragen

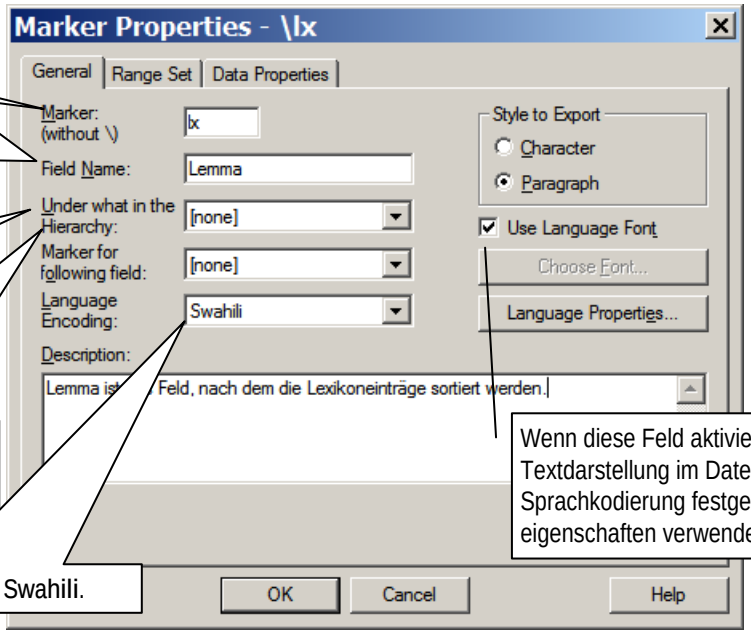
Feldbezeichnung. Tragen Sie hier **Lemma** ein. Ein Lemma ist das Stichwort in einem Wörterbuch, dem alle anderen Informationen untergeordnet sind.

Hier kann angegeben werden, welchem anderen Marker der neue Marker untergeordnet werden soll (Markerhierarchie).

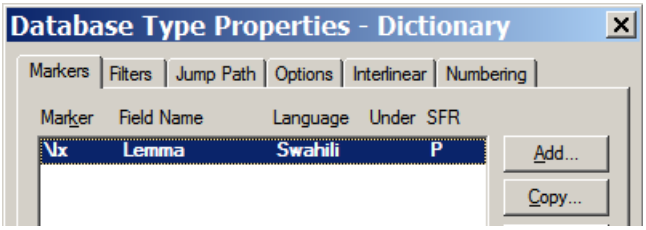
Da noch keine weiteren Marker definiert sind, kann hier noch nichts ausgewählt werden.

Wählen Sie als Sprachkodierung **Swahili**.

Wenn diese Feld aktiviert ist, werden für die Textdarstellung im Datensatzfeld die in der Sprachkodierung festgelegten Zeichen-eigenschaften verwendet.



Zur Festlegung der Marker Properties stehen drei Register zur Verfügung: General, Range Set und Data Properties. Gewählt ist zunächst das Register General. Nehmen Sie die im Bild gezeigten Einträge vor und klicken Sie auf OK. Hier sehen Sie das Ergebnis:

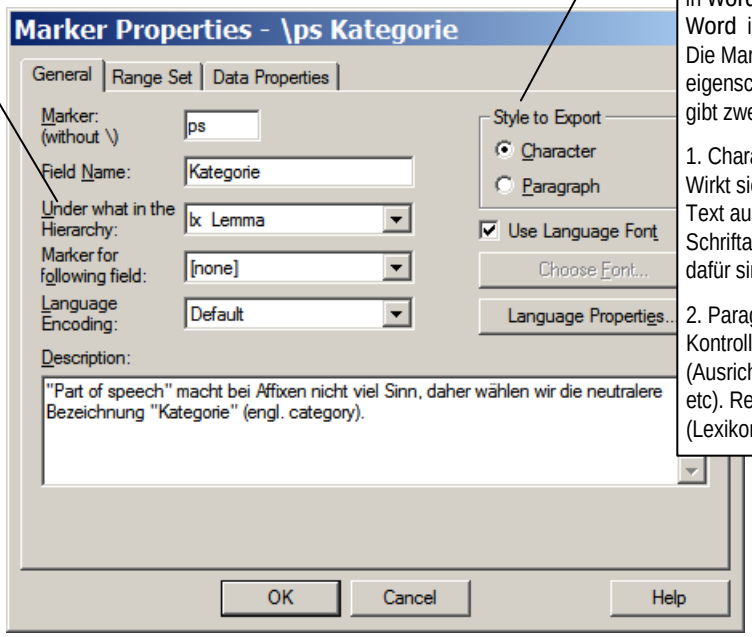


Als nächstes werden die weiteren für das Wörterbuch erforderlichen Markierungen definiert. Dies sollen neben den bereits bekannten Markierungen \ps (*Part of Speech* – "Redeteil") und \gl (Glosse) die folgenden Markierungen sein: \de (Definition), \xv (Beispiel in der Landessprache, hier Swahili), \xe (Übersetzung des Beispiels) und \dt (Datum).

Marker	Field Name	Language	Under	SFR
\lx	Lemma	Swahili		P
\ps	Kategorie	Default	\lx	C

☐ Wählen Sie jetzt Add... und tragen Sie die Eigenschaften für den Marker \ps ein.

Der Marker \ps wurde automatisch dem Marker \lx untergeordnet. Andere Marker stehen derzeit auch nicht zur Verfügung.



Style to Export: Style entspricht einer Formatvorlage in Word. Eine Datenbank kann zur Bearbeitung in Word im Rich Text Format (RTF) exportiert werden. Die Marker werden dabei mit den definierten Zeicheneigenschaften einer Formatvorlage zugewiesen. Es gibt zwei Möglichkeiten:

1. Character style (Zeichenformat)
Wirkt sich nur auf den in diesem Feld enthaltenen Text aus und bezieht sich auf Eigenschaften wie Schriftart, -grad, -schnitt, -farbe. Relevante Marker dafür sind \ps und \gl (Glosse).

2. Paragraph style (Absatzformat)
Kontrolliert alle visuellen Aspekte eines Absatzes (Ausrichtung, Tabulatoren, Zeilenabstand, Ränder, etc). Relevante Markers dafür sind \lx (Lexikoneintrag, Lemma) und \se (*subentry*).

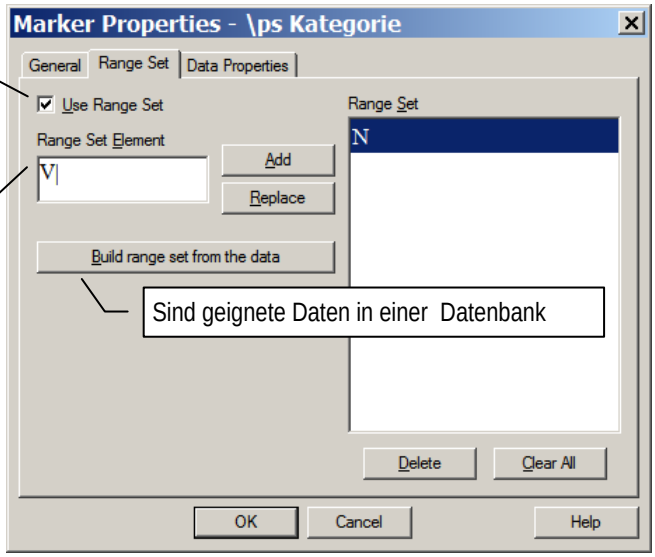
Bei dem Marker für die Kategorie (\ps) bietet es sich an, den Wertebereich, aus dem die Einträge gewählt werden können, zu kontrollieren und zu beschränken. Dies ist besonders dann von Vorteil, wenn mehrere Personen an einem gemeinsamen Projekt arbeiten. Für diesen Zweck ist das Register Range Set vorgesehen.

☐ Wählen Sie das Register Range Set und aktivieren Sie darin das Feld Use Range Set.

☐ Tragen Sie in dem Feld Range Set Element den Buchstaben N ein und klicken Sie Add.

Hiermit legen Sie fest, ob ein Wertebereich für den Marker definiert werden soll.

Hier wird das Symbol eingetragen, das dem Wertebereich hinzugefügt werden (Add) oder einen bereits vorhandenen Wert ersetzen soll (Replace).



Verfahren Sie analog mit den Kategorien V (Verb) und A (Adjektiv). Sie werden feststellen, dass die Einträge in dem Range Set automatisch alphabetisch angeordnet werden. Da wir noch gar kein Lexikon erstellt und mit der Analyse noch gar nicht begonnen haben, werden wir weitere Einträge parallel zur Erstellung und Erweiterung des Lexikons ergänzen.

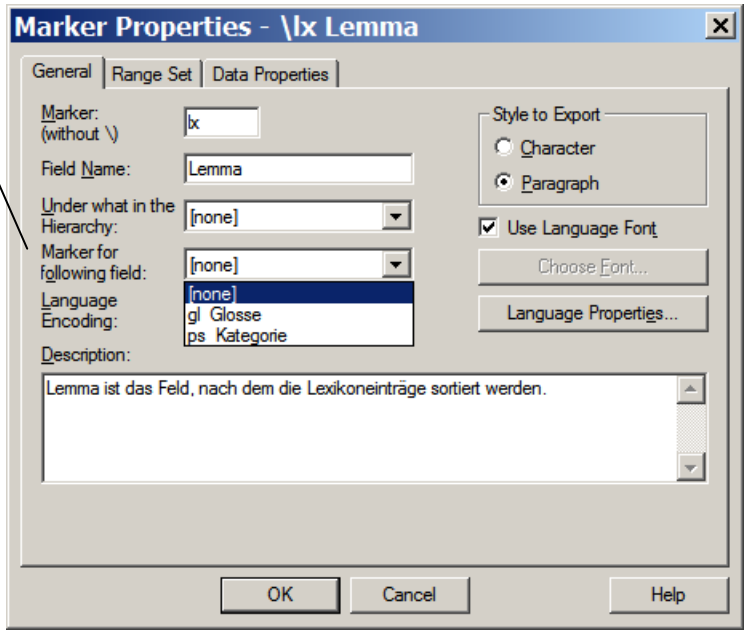
- ☐ Wählen Sie jetzt wieder das Register General.
- ☐ Deaktivieren Sie das Feld Use Language Font. Dadurch wird der Schalter Choose Font aktiviert, so dass Sie jetzt zusätzliche Schriftarteigenschaften festlegen können.
- ☐ Wählen Sie den Schalter Choose Font und legen Sie als Schriftschnitt *kursiv* fest. Mit OK kehren Sie wieder zurück.
- ☐ Klicken Sie erneut auf OK um die Kartei zu verlassen. Sie werden feststellen, dass in dem Eigenschaftsfenster in der Zeile für \ps in der Spalte SFR die Buchstaben F und R hinzugekommen sind. Der Buchstabe F zeigt an, dass für für den Marker Schriftarteigenschaften (*Font properties*) definiert worden sind, der Buchstabe R zeigt an, dass für diesen Marker ein Range Set existiert.

Marker	Field Name	Language	Under	SFR
\lx	Lemma	Swahili		P
\ps	Kategorie	Default	\lx	CFR
\gl	Glosse	Default	\lx	CF

- ☐ Machen als nächstes einen Eintrag für den Marker \gl. Dabei sollen für diesen Marker spezifische Schriftarteigenschaften festgelegt werden: Schriftart – Arial, Schriftschnitt – Fett, Schriftfarbe – grün, Schriftgrad – 11pt.

Die Festlegung von spezifischen Schriftarteigenschaften für Marker wie \ps und \gl ist unter linguistischen Gesichtspunkten natürlich irrelevant, sie verschafft andererseits einen besseren Überblick.

Hiermit kann festgelegt werden, welches andere Feld auf den zur Diskussion stehenden Marker (hier: Lemma) folgen soll. Sinnvollerweise sollte dies hier \ps sein. Wenn man ein Lemma eingegeben hat und dann die Eingabetaste drückt, wird automatisch ein Feld für \ps (Kategorie) eingefügt.



Es gibt noch eine weitere interessante Marker-Eigenschaft: Marker for following field. Damit können Sie festlegen, welches andere Datenfeld bei der Eingabe eines Feldes als nächstes folgen soll. So können Sie beispielsweise bestimmen, dass bei der Eingabe eines Lemmas automatisch das Feld für die Kategorie (*Part of Speech* \ps) folgen soll und auf jenes das Feld für die Glosse (\gl). Gehen Sie folgendermaßen vor:

- ☐ Wählen Sie den Marker \lx und Modify.
- ☐ Wählen Sie aus der Markerliste unter Marker for following field den Eintrag \ps. Kehren Sie mit OK zu den Database Type Properties zurück.
- ☐ Verfahren Sie entsprechend mit dem Marker \ps und ordnen sie diesem den Maker \gl als Nachfolger zu.

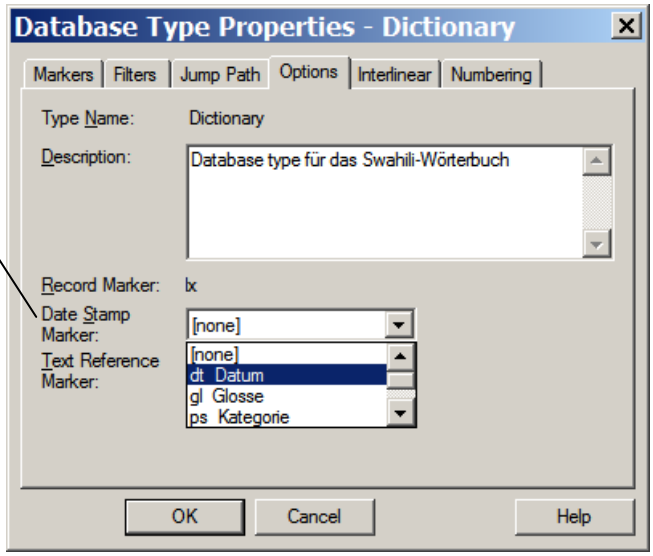
Marker	Field Name	Language	Under	SFR
\lx	Lemma	Swahili		P
\ps	Kategorie	Default	\lx	CFR
\gl	Glosse	Default	\lx	CF
\de	Definition	Default	\lx	
\xv	Beispiel	Swahili	\lx	P
\xe	Übersetzung	Default	\lx	P
\dt	Datum	Default	\lx	C

Die Marker Definition (\de), Beispiel (\xv), Übersetzung (\xe) und Datum (\dt) sind für die Interlinearisierung ohne Bedeutung, sind aber wichtig für das *Wörterbuch* qua Wörterbuch. Die Glosse liefert nur eine grobe Bedeutungsangabe, während die Definition als eine genauere Charakterisierung der Bedeutung eines Lexikoneintrags gedacht ist. Der Markername \xv ist abgeleitet aus dem engl. ***ex**ample* ***v**ernacular* und meint ein Belegbeispiel in der Landessprache für den Lexikoneintrag, hier also Swahili. Daher muss hier die Sprachkodierung Swahili zugeordnet werden. Das **e** in \xe steht für Englisch und soll hier allgemein für die Übersetzung des Belegbeispiels verwendet werden (für uns auf Deutsch).

- ☐ Ergänzen Sie die Marker entsprechend der obigen Tabelle.

Der Marker \dt für das Datumsfeld dient Zwecken der Dokumentation.

Hier kann ein Marker aus der Liste der verfügbaren Marker bestimmt werden, der das Feld identifiziert, in dem beim Anlegen und Ändern eines Datensatzes automatisch das aktuelle Datum eingetragen wird. Wir haben dafür den Marker \dt als "Datumsstempel" definiert.

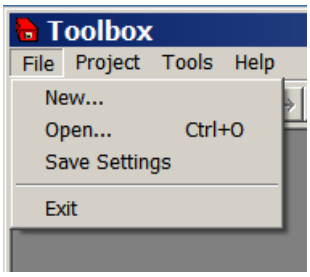


- ☐ Aktivieren Sie im Dialogfenster Database Type Properties das Register Options und wählen als Datumsstempel (*Date Stamp*) den Marker \dt.

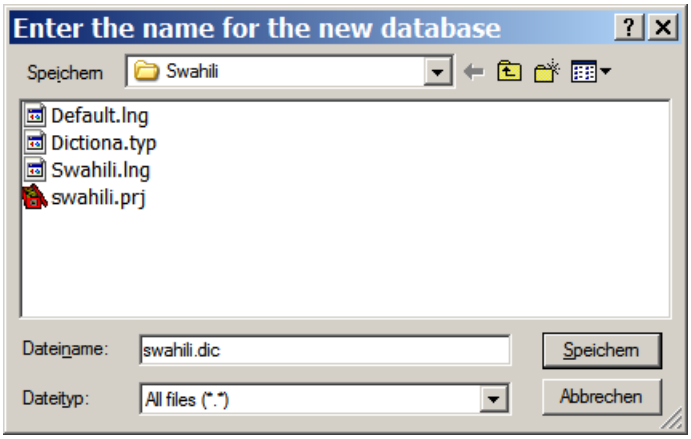
Damit ist die Einrichtung des Datenbanktyps für Wörterbücher abgeschlossen. Schließen Sie alle Fenster jeweils durch OK bzw. Close.

Erstellen eines Wörterbuches für Swahili

Alle Dialogfenster sollten inzwischen geschlossen sein. Alle Datenbanken – gleich welcher Art – werden über das Dateimenü File angelegt.

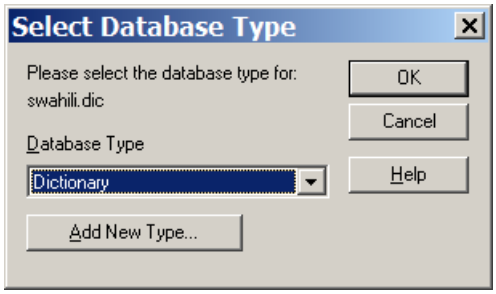


- ☐ Wählen Sie den Menübefehl File New ... und legen Sie eine Datei mit dem Namen swahili.dic an. Falls Sie sich – aus welchen Gründen auch immer – nicht im Verzeichnis Swahili befinden sollten, müssen Sie dorthin navigieren.



- ☐ Klicken Sie auf den Schalter Speichern, um die Datei anzulegen.

Jede Datenbank, die auf diese Weise angelegt wird, muss einem passenden Datenbanktyp zugeordnet werden. Daher wird man als nächstes dazu aufgefordert, einen Datenbanktyp auszuwählen:
Please select the database type for: swahili.dic.

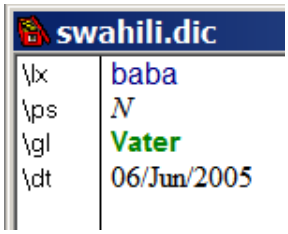


Da wir bisher nur einen Datenbanktyp definiert haben, steht auch nur dieser zur Verfügung. Im allgemeinen Fall müsste man einen Typ aus der Liste auswählen. Es besteht an dieser Stelle auch die Option, einen neuen Datenbanktyp hinzuzufügen.

- ☐ Wählen Sie OK um den Datenbanktyp *Dictionary* auszuwählen.
- ☐ Beenden Sie mit OK.

Es öffnet sich ein Fenster mit dem ersten (noch leeren) Datensatz für das Swahili-Lexikon. Achten Sie darauf, dass gleichzeitig mit dem Öffnen einer Datenbank weitere Menüpunkte zur Verfügung stehen, beispielsweise Edit, Database, View, und Window.

Machen Sie Einträge für *baba* 'Vater' und *mama* 'Mutter'. Sie werden feststellen, dass nach Eingabe des Lemmas sofort ein Feld für die Kategorie eingefügt wird, und anschließend ein Feld für die Glosse. Die Texte in den Feldern weisen die für die jeweiligen Marker definierten Schriftart-Eigenschaften auf. Wenn Sie zum ersten Datensatz (*baba*) zurückkehren, sehen Sie, dass auch das Datumsfeld ausgefüllt worden ist.

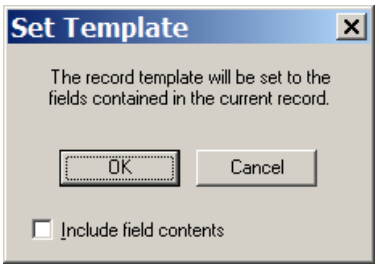


Die Felder für eine Definition der Bedeutung (\de), ein Beispiel in der Landessprache (\xv) sowie seine Übersetzung (\xe) werden wir nur bei Bedarf einsetzen.

Erstellen eines neuen Datensatzmusters

Obwohl wir für die Marker \lx und \ps festegelegt haben, dass bei der Eingabe des Feldinhaltes automatisch das nächste Feld (\ps bzw. \gl) eingefügt wird, wollen wir für das Swahili-Lexikon ein Muster (*Template*) für einen Datensatz erstellen. Es handelt sich dabei um eine Zusammenstellung der Markierungen, die automatisch eingefügt werden sollen, wann immer ein neuer Datensatz erstellt werden soll.

☐ Wählen Sie Database, Template.

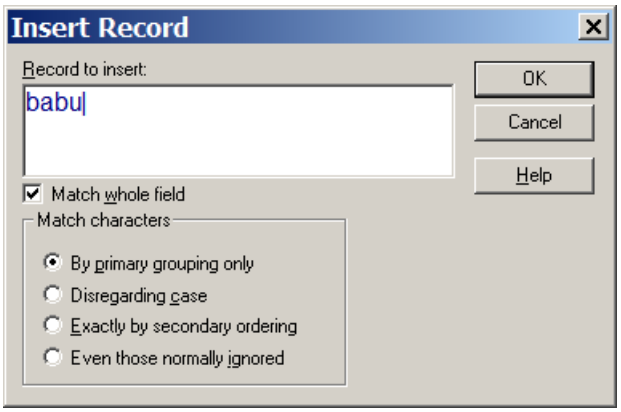


☐ Wählen Sie OK und das Dialogfenster verschwindet wieder.

Einfügen eines weiteren Datensatzes

Führen Sie folgende Schritte aus, um einen weiteren Datensatz zu generieren und zu sehen, wie das *Template* funktioniert:

☐ Wählen Sie Database, Insert Record oder benutzen Sie das Tastaturkürzel (Strg+N). Es öffnet sich ein Dialogfenster mit dem Titel *Insert Record*.



☐ Geben Sie *babu* 'Großvater' als einzufügenden Datensatz (*Record to insert*) an.

☐ Wählen Sie OK.

Der neue Datensatz wird eingetragen und das Programm kehrt zum Lexikonfenster (*swahili.dic*) zurück, das diesen neuen Lexikoneintrag mit dem Lemma *babu* anzeigt. Die Schreibmarke befindet sich auf der **\ps**-Zeile. Beachten Sie auch, dass das Muster die Markierungen **\ps**, **\gl**, und **\dt** automatisch eingefügt hat.

☐ Geben Sie *N* als Kategorie und *Großvater* als Glosse ein.

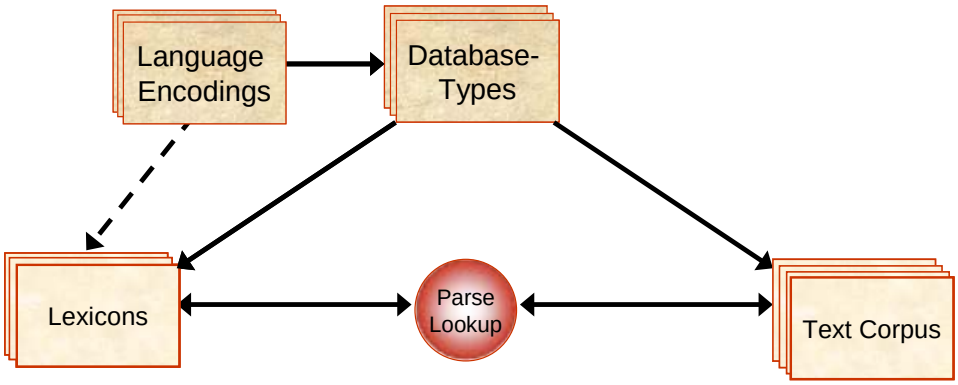
Der Datumsstempel ist noch nicht ausgefüllt. Dies geschieht erst wenn man den neuen Datensatz verlassen hat:

☐ Klicken Sie auf den Previous Record Schalter ← um zum Eintrag *baba* zu gelangen.

☐ Klicken Sie dann auf den Next Record Schalter → um zu *babu* zurückzukehren. Das Datumsfeld sollte jetzt mit dem heutigen Datum ausgefüllt sein.

Erstellen eines Textkorpus für Swahili

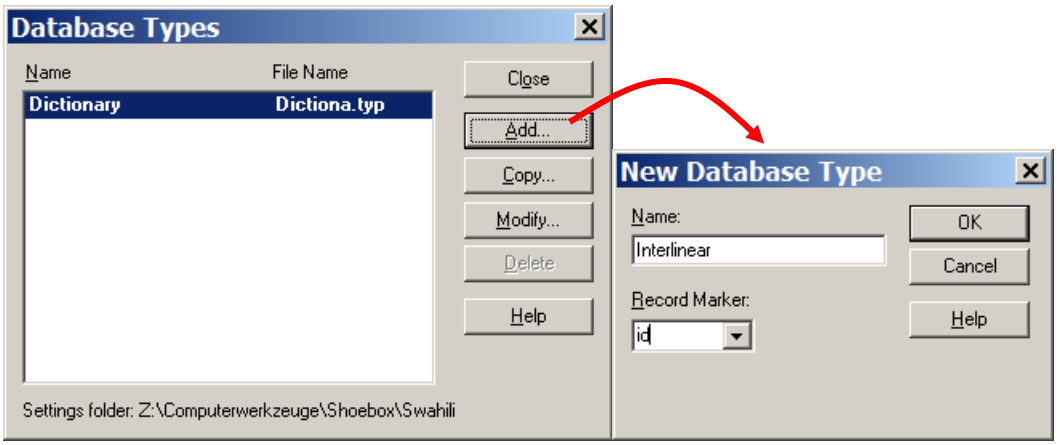
Die Datengrundlage für die morphologische Analyse mit Toolbox ist ein Korpus von Texten mit spezifischen, auf die Interlinearisierung ausgerichteten Eigenschaften. Diese Eigenschaften werden, wie beim Lexikon auch, in einem Datenbanktyp definiert.



Erstellen eines Datenbanktyps für Swahili-Texte

Zunächst muss eine Entscheidung darüber getroffen werden, was in einem Interlineartext als Datensatz gelten soll. Zwei Möglichkeiten stehen zu Wahl:

1. Jeder Satz im Text ist ein Datensatz, identifiziert durch den Referenzmarker \ref. Diese Struktur hatten die ersten Texte, mit denen wir gearbeitet haben, und wird auch beim Import mit TextPrep.cct erzeugt.
2. Der Datensatz ist der ganze Text mit der Identifikation \id. Diese Option hat eine Reihe von Vorteilen und wird auch von den Programmautoren empfohlen. Da alle Sätze zum selben Datensatz gehören, steht ein sehr viel größerer Kontext für Analyseentscheidungen zur Verfügung, als wenn nur der einzelnen Satz im Blick ist.

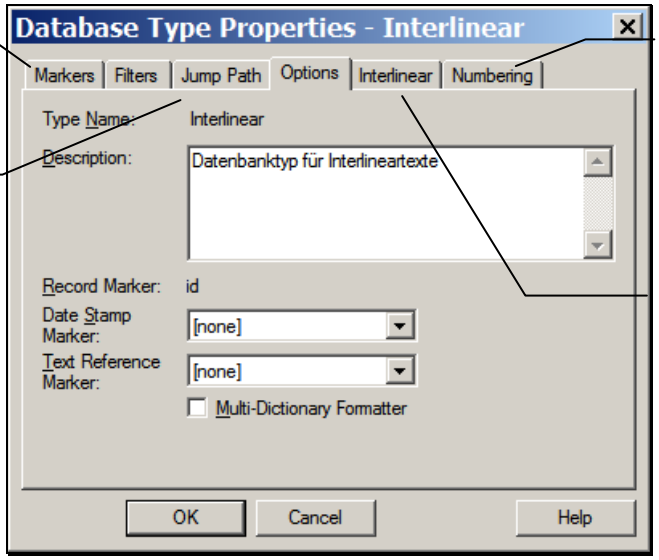


- ☐ Wählen Sie Project >> Database Types ... und den Schalter [Add ...], um einen neuen Datenbanktyp für Interlineartexte hinzuzufügen.
- ☐ Tragen Sie als Namen *Interlinear* und als *Record Marker* id ein.
- ☐ Wählen Sie [OK].

Es öffnet sich das gleiche Dialogfenster für die Database Type Properties wie beim Lexikon. Es werden jetzt aber andere Eigenschaften relevant. Neben dem Marker-Register werden wir uns insbesondere mit den Registern Jump Path, Interlinear und Numbering zu beschäftigen haben.

Wie beim Lexikon werden in diesem Register die Eigenschaften der Marker festgelegt.

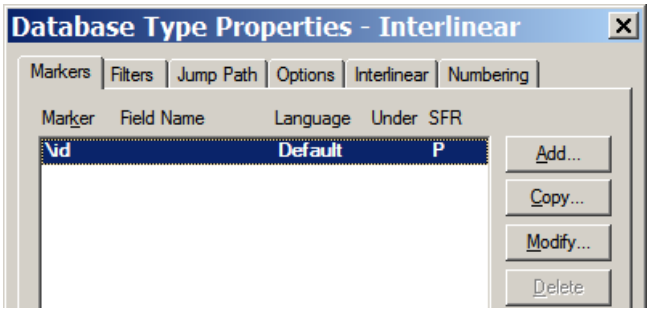
Mit dem Jump Path wird eine Verbindung beispielsweise zwischen Text und Wörterbuch



Über dieses Register werden die grundlegenden Eigenschaften für die Numerierung der Textabschnitte festgelegt.

Dies ist für die Interlinearisierung das wichtigste Register. Hier werden die Eigenschaften der Prozesse Parse und Lookup definiert.

Nach dem Anlegen des Datenbanktyps mit dem Datensatzmarker id haben wir zunächst nur einen Marker und damit nur ein Datenfeld festgelegt.

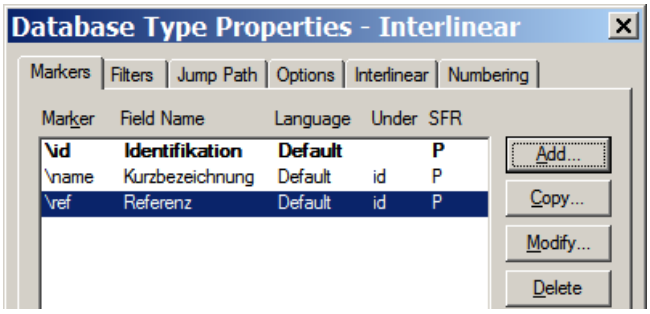


Definition von Markern und ihren Eigenschaften

Wie wollen daher als erstes die weiteren für diesen Datenbanktyp erforderlichen Marker mit ihren Eigenschaften definieren.

Marker	Field Name	Language	Under	SFR
\id	Identifikation	Default		P
\name	Kurzbezeichnung	Default	id	P
\ref	Referenz	Default	id	P

- ☐ Modifizieren Sie zunächst den Marker id (Schalter Modify), indem Sie den Feldnamen *Identifikation* und eine passende Beschreibung ergänzen.
- ☐ Fügen Sie dann die Marker name und ref mit den o.a. Eigenschaften hinzu (Schalter Add). Der Marker name führt ein Feld für eine Kurzbezeichnung ein, die wir später für die Erzeugung einer Referenznummer (wie z.B. Text1.001) verwenden wollen. Diese Referenznummer wird dann im Referenzfeld ref stehen.



Markern für die Interlinearisierung und deren Eigenschaften

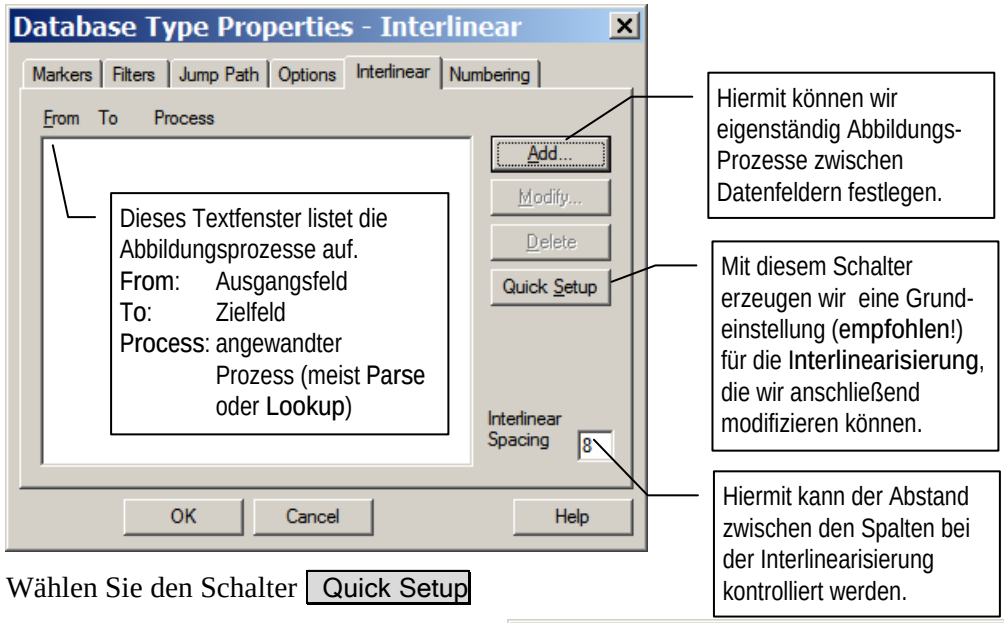
Die folgenden Marker mit den damit verbundenen Datensatzfeldern benötigen wir für den Interlinearisierungsprozess:

Marker	Field Name	Language	Under	SFR
…	…	…	…	…
\tx	Text	Swahili	ref	P
\mb	Morpheme	Swahili	tx	PF
\ps	Kategorie	Default	tx	PF
\gl	Glosse	Default	tx	PF

Das Feld mit dem Marker tx (Text) soll den zu interlinearisierenden Ausgangstext enthalten, das Feld mit dem Marker mb (Morpheme) die Segmentierung dieses Textes in Morphe(me). Da diese beiden Felder Material der Ausgangssprache enthalten, muss ihnen die Sprachkodierung (Spalte: Language) dieser Sprache (hier: Swahili) zugeordnet werden. Im Feld mit dem Marker ps (Kategorie) stehen die lexikalischen und grammatischen Kategorien, denen die Morphe(me) im mb-Feld angehören, und in gl die entsprechenden Glossen. Das "F" in der Spalte SFR zeigt an, dass bestimmte Schriftart-Eigenschaften festgelegt werden sollen.

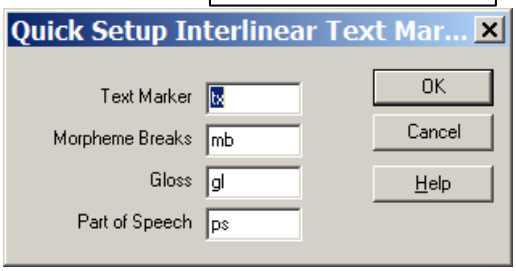
Bei der Definition dieser Marker müssen gleichzeitig die Prozesse festgelegt werden, die den Übergang vom Textfeld tx zu den anderen Feldern steuern. Wir können nun sowohl die Marker und ihre Eigenschaften als auch die erforderlichen Prozesse Schritt für Schritt manuell festlegen. Für die Grundeinstellung eines Interlineartextes stellt uns das Programm aber eine Funktion zur Verfügung, die einen großen Teil dieser Aufgaben automatisch erledigt.

☐ Öffnen Sie dazu das Register Interlinear.



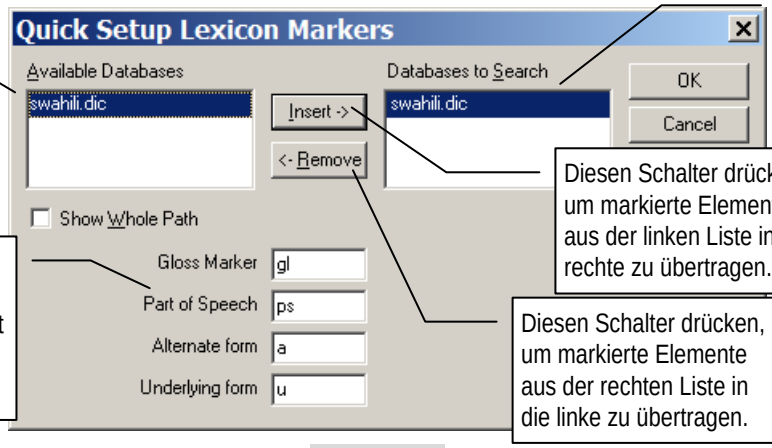
☐ Wählen Sie den Schalter Quick Setup

Die Einstellungen für die relevanten Marker (tx, mb, gl, ps) können so übernommen werden.



☐ Bestätigen Sie die vorgeschlagenen Einstellungen mit OK.

Dies führt uns zum Dialogfenster für die Einstellungen der erforderlichen Beziehungen zwischen dem Interlineartext und dem zugeordneten Lexikon.



Liste der verfügbaren **Datenbanken**. Bis jetzt steht nur **swahili.dic** zur Verfügung

Liste der **Datenbanken**, die bei der **Interlinearisierung** durchsucht werden sollen.

Diesen Schalter drücken, um markierte Elemente aus der linken Liste in die rechte zu übertragen.

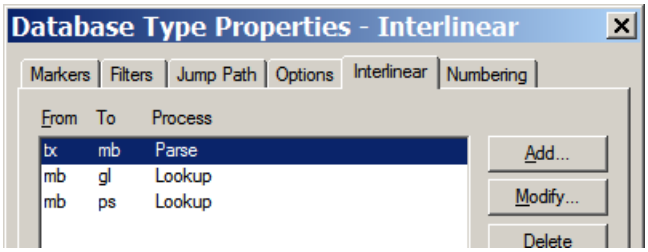
Hier können die für die Interlinearisierung relevanten Marker im zugeordneten Lexikon festgelegt werden. Die Einstellung kann so übernommen werden.

Diesen Schalter drücken, um markierte Elemente aus der rechten Liste in die linke zu übertragen.

☐ Wählen Sie den Schalter [Insert →] um die einzige verfügbare Datenbank in die Liste der zu durchsuchenden Datenbanken zu übertragen.

☐ Bestätigen Sie die Einstellungen mit [OK].

Das Ergebnis sehen Sie im folgenden Bild.

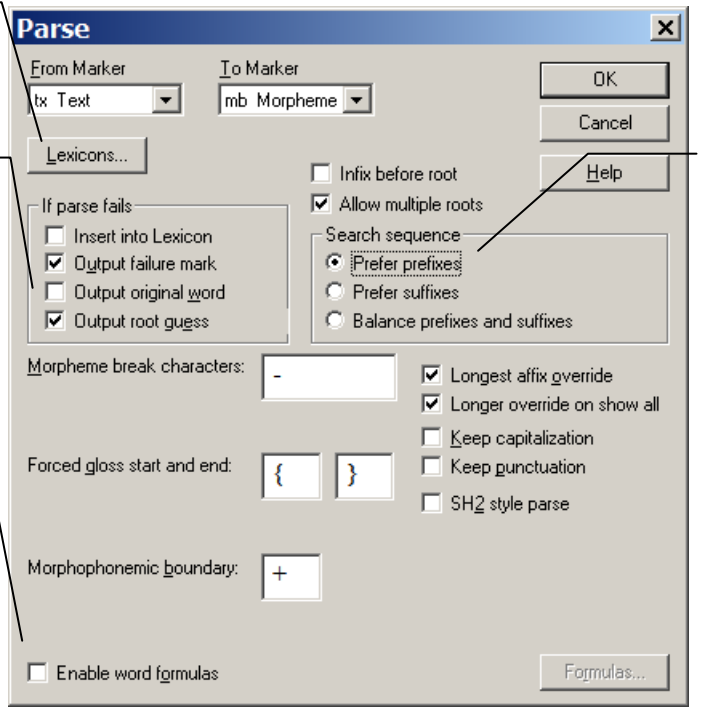


Modifikation der Parse-Eigenschaften

Der Prozess, der den Text im Textfeld **tx** in Morphe(me) zerlegt, wird **Parse** genannt. Dieser Ausdruck stammt ursprünglich aus der englischen Schulgrammatik und ist vom lat. *pars orationis* (Pl. *partes*) 'Redeteil' (engl. *part of speech*) abgeleitet. Das Verb *parse* bezeichnet dabei das Zerlegen eines Satzes in seine grammatischen Bestandteile.

☐ Modifizieren Sie den Eintrag für den Parse-Prozess (Schalter: [Modify...])

Hiermit können die Einstellungen der zugeordneten **Lexika** modifiziert werden.



Hier können Aktionen eingestellt werden für den Fall, dass die Analyse **scheitert**. Voreingestellt ist die Ausgabe eines *failure mark*. **Aktivieren** Sie unbedingt noch das Feld *Output root guess*.

Hiermit kann das Verhalten des Parsers beeinflusst werden:
- o Prefer prefixes—Präfixe werden vor Suffixen probiert.
- o Prefer suffixes—Suffixe werden vor Präfixen probiert.
- o Balance prefixes and suffixes—Präfixe und Suffixe werden gleichrangig behandelt.

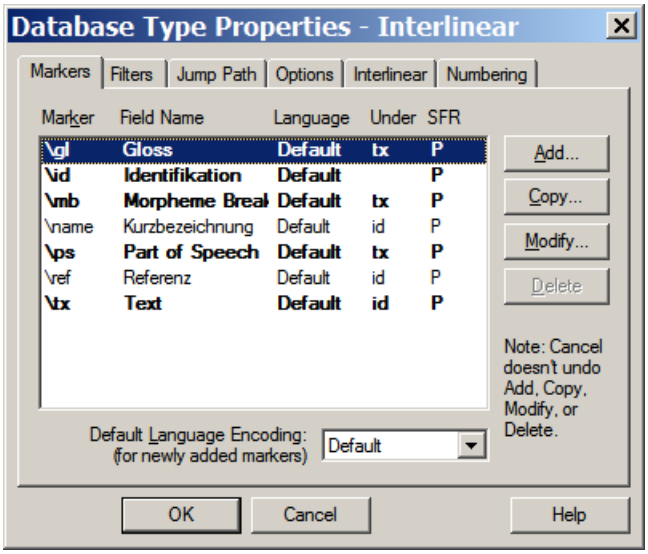
Da **Swahili** Präfixe bevorzugt, sollte die erste Option eingestellt werden.

Wortformeln sind eine Art Wortbildungsregeln, die den Parseprozess steuern können. Wir werden später darauf zurückkommen.

o Aktivieren Sie das Feld Output root guess und die Option Prefer prefixes. Alle weiteren Einstellungen können übernommen werden. Bestätigen Sie mit [OK].

Die Einstellungen für die beiden Lookup-Prozesse brauchen nicht modifiziert zu werden.

- Schließen Sie die Interlinearisierungseinstellungen mit OK ab.
- Da wir die Marker-Einstellungen noch modifizieren und ergänzen wollen, wählen Sie noch einmal den Schalter Modify ...



Anpassung der Eigenschaften der Marker von Interlineartexten

Die Marker tx, mb, gl und ps sind vom Programm angelegt und mit Default-Eigenschaften versehen worden. Hier müssen wir einige Änderungen vornehmen.

Marker	**Field Name**	**Language**	**Under**	**SFR**
…	**…**	**…**	**…**	**…**
\tx	Text	Swahili	ref	P
\mb	Morpheme	Swahili	tx	PF
\ps	Kategorie	Default	tx	PF
\gl	Glosse	Default	tx	PF
\ft	Freie Übersetzung	Default	tx	PF

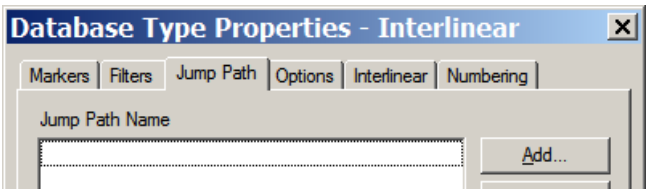
- Der Textmarker tx muss die Sprachkodierung für *Swahili* erhalten und soll dem Marker ref untergeordnet werden.
- Ändern Sie den Feldnamen des Morphemmarkers mb in *Morpheme* und die Sprache in *Swahili* und wählen Sie eine andere Schriftfarbe (z.B. dunkelrot).
- Ändern Sie den Feldnamen *Part of Speech* in *Kategorie* um und wählen Sie als Schriftschnitt *kursiv*.
- Ändern Sie den Feldnamen *Gloss* in *Glosse* um und wählen Sie als Schriftart Arial, als Schriftgrad 11pt und als Schriftfarbe grün.
- Fügen Sie den Marker ft (Freie Übersetzung) unter tx mit dem Schriftschnitt *kursiv* hinzu.

Festlegung eines "Jump Path"

Ein *Jump Path* ist die Festlegung einer Zieldatenbank, in der, ausgehend von einem Datenfeld in der Ausgangsdatenbank, nach Informationen gesucht werden soll. Im Folgenden wollen wir festlegen, dass nach Morphemen im Morphemfeld (mb) in der Lexikondatenbank gesucht bzw. dort Einträge gemacht werden sollen.

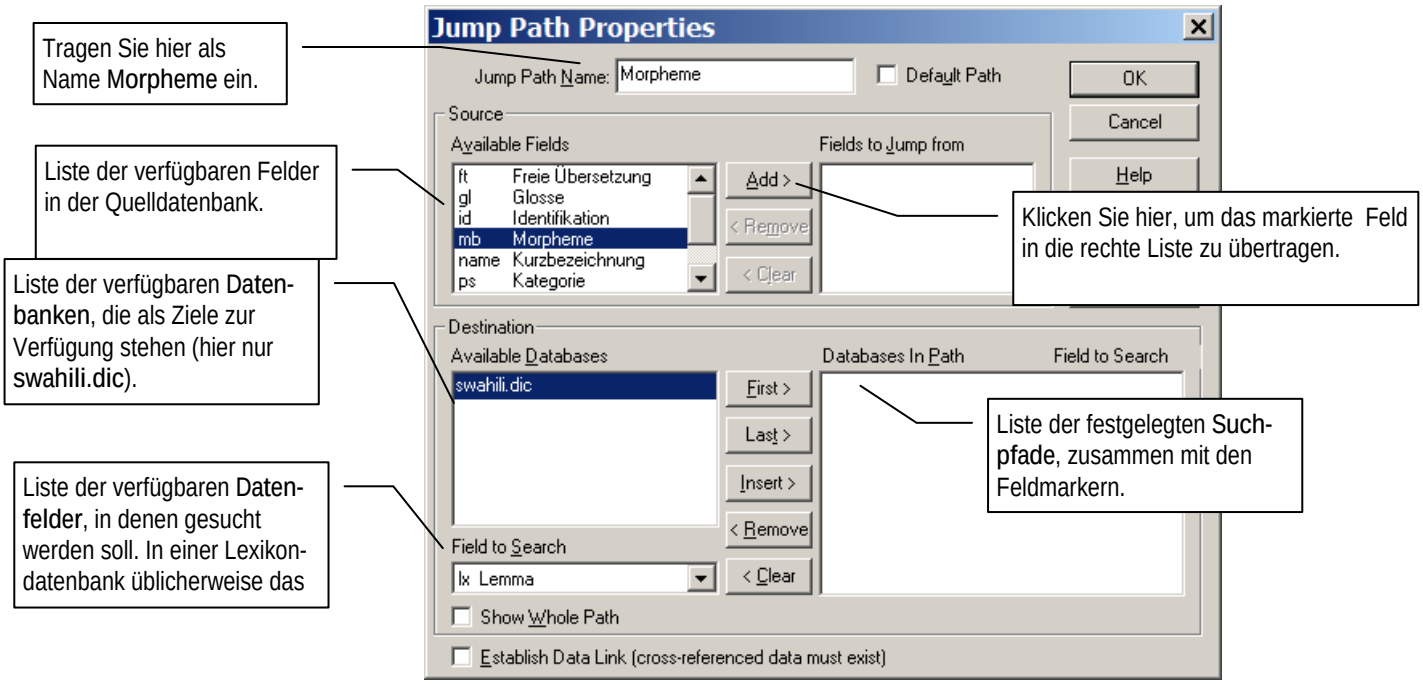
- Öffnen Sie das Jump Path-Register.

Da noch keine Jump Path definiert worden ist, steht nur der Schalter Add ... zur Verfügung.



- Wählen Sie Add ...

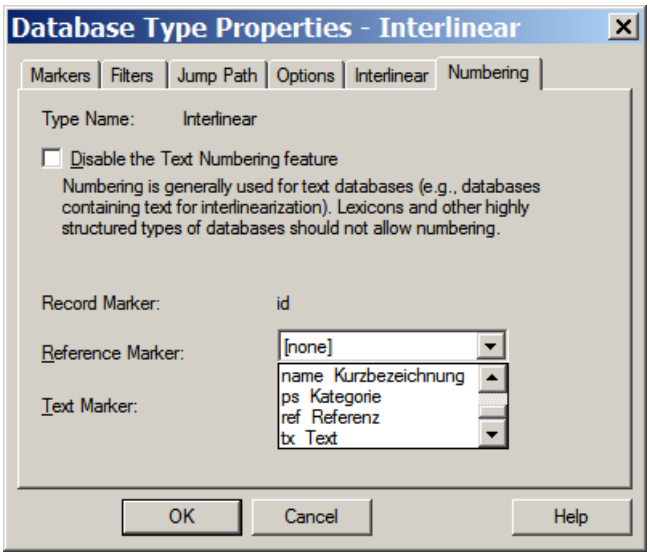
Ein Jump Path hat einen Namen.
Für jeden Jump Path muss eine Quelle (*Source*) und ein Ziel (*Destination*) angegeben werden.



- Geben Sie als Jump Path Name *Morpheme* ein.
- Markieren Sie in der Liste Available Fields den Eintrag mb Morpheme und übertragen Sie ihn mit dem Schalter Add > in die Liste Fields to Jump from.
- Wählen Sie die Zieldatei swahili.dic aus (Suchfeld: lx Lemma).
- Alle übrigen Einstellungen können übernommen werden. Bestätigen Sie das mit OK.

Einstellung der Numerierungseigenschaften

Eine Numerierung macht keinen Sinn für Lexikondatenbanken, wohl aber für Textdatenbanken, und wir werden im weiteren Verlauf dieses Projektes auch davon Gebrauch machen.

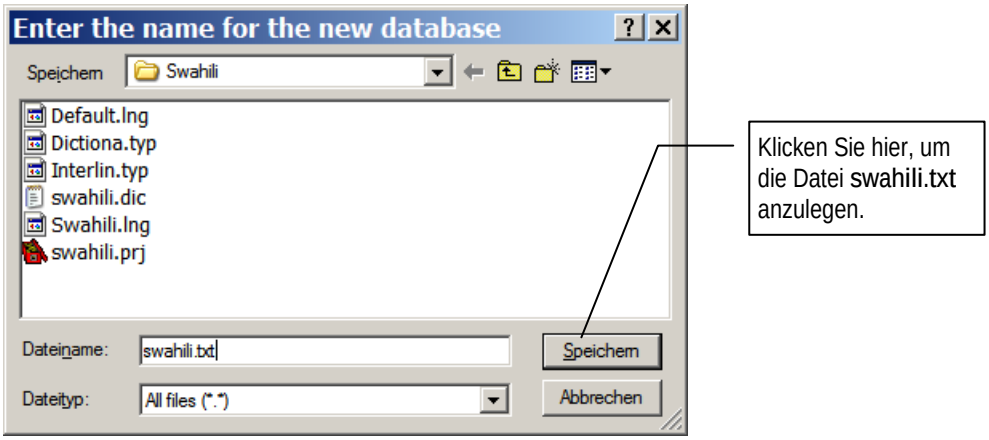


- Öffnen Sie das Numbering-Register.
- Wählen Sie aus der Liste der verfügbaren Marker ref als Referenz-Marker
- Wählen Sie tx als Text-Marker.
- Bestätigen Sie die Auswahl mit OK.

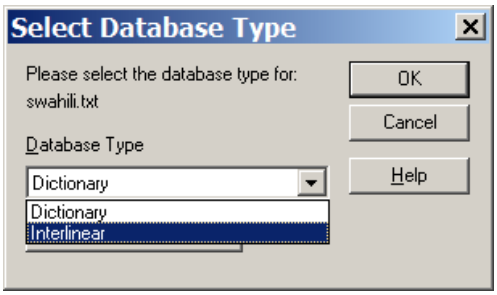
Erstellen des Text-Korpus für Swahili

Damit haben wir die Festlegung der Eigenschaften des Datenbanktyps für Swahili-Texte abgeschlossen und können jetzt auf dieser Grundlage eine Textdatenbank und einen ersten Datensatz in dieser Datenbank anlegen.

- Wählen Sie den Menübefehl File >> New und legen Sie unter dem Namen swahili.txt eine neue Textdatei an.

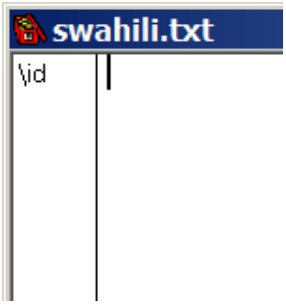


Sie werden aufgefordert dieser Datei einen Datentyp zuzuordnen: *Please select the database type for: swahili.txt*.



- Wählen Sie Interlinear aus der Liste der zur Verfügung stehenden Datenbanktypen aus.
- Bestätigen Sie die Auswahl mit OK.

Es öffnet sich dann ein Fenster für die Text-Datenbank mit dem Namen *swahili.txt*, das den ersten – noch leeren – Datensatz mit dem Feld id (Identifikation) anzeigt (wohlgemerkt: der Datensatz ist der gesamte Text).



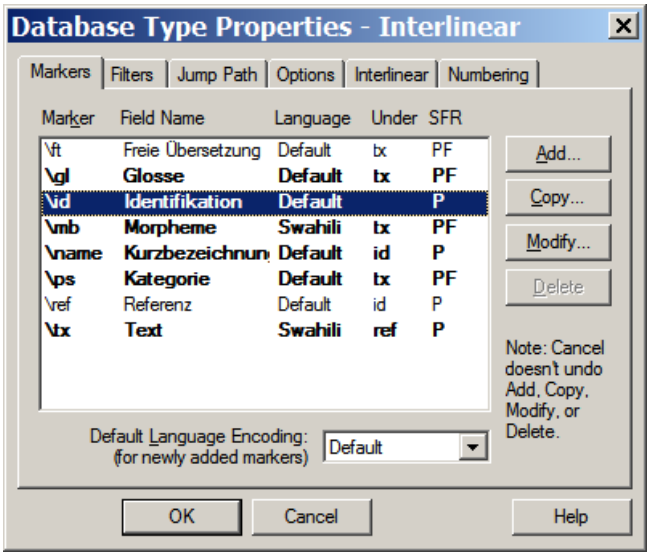
Wir wollen dem Text – es handelt sich um die Daten für die Swahili-Hausaufgabe – zunächst eine Identität geben.

- Schreiben Sie in das id-Feld den Text *Morphologische Analyse – Swahili*.
- Fügen Sie als nächstes eine name-Feld mit dem Text *Analyse* hinzu.[5]

Der Text selbst soll dann Satz für Satz zusammen mit den Übersetzungen manuell eingegeben werden. Bevor wir damit beginnen, wollen wir uns die Arbeit etwas erleichtern, indem wir einige Marker-Eigenschaften modifizieren. Was wir erreichen wollen ist, dass bei Eingabe eines Referenzfeldes (ref) automatisch ein Textfeld (tx) eingegeben wird, bei Eingabe eines Textfeldes ein Übersetzungsfeld (ft), und bei Eingabe des letzteren wiederum ein Referenzfeld.

- Wählen Sie den Menü-Befehl Database >> Properties …

[5] Zur Erinnerung: gehen Sie in eine neue Zeile, geben Sie den 'Backslash' \ ein und dann die Zeichenfolge name.



- Modifizieren Sie \ref, indem Sie als *Marker for following field* \tx auswählen.
- Modifizieren Sie \tx, indem Sie als *Marker for following field* \ft auswählen.
- Modifizieren Sie \ft, indem Sie als *Marker for following field* \ref auswählen.
- Bestätigen Sie die Änderungen mit OK.
- Tragen Sie im Text unterhalb von \name ein Referenzfeld \ref ein, das aber leer bleiben soll.

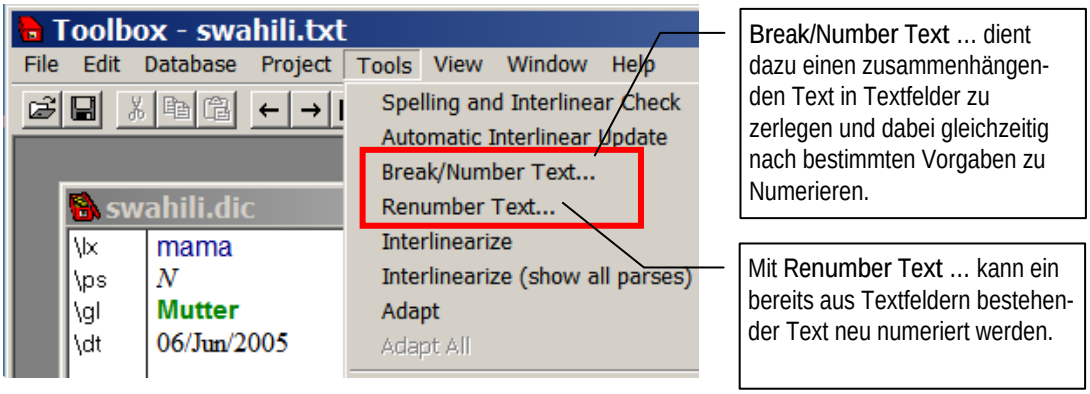
Sie werden feststellen, dass nach dem Drücken der Eingabetaste automatisch ein Textfeld eingefügt wird.

- Geben Sie im Textfeld den ersten Satz aus der Aufgabe ein: *Mtafika*. Nach Betätigen der Eingabetaste wird ein Übersetzungsfeld eingefügt, in das Sie die zugeordnete Übersetzung eintragen: *Ihr werdet ankommen*.

Danach kommt wieder ein Referenzfeld. Tragen Sie nun auf diese Weise nacheinander sämtliche 25 verbleibenden Sätze in die Datenbank ein.

Referenzfelder neu numerieren

Referenzfelder, die sich nicht voneinander unterscheiden, machen nicht viel Sinn.



Wir hätten natürlich bei der Eingabe der Referenzfelder jeweils unterschiedliche Referenzausdrücke eintragen können. Wir haben dies nicht getan, weil diese Aufgabe vom Programm übernommen werden soll.

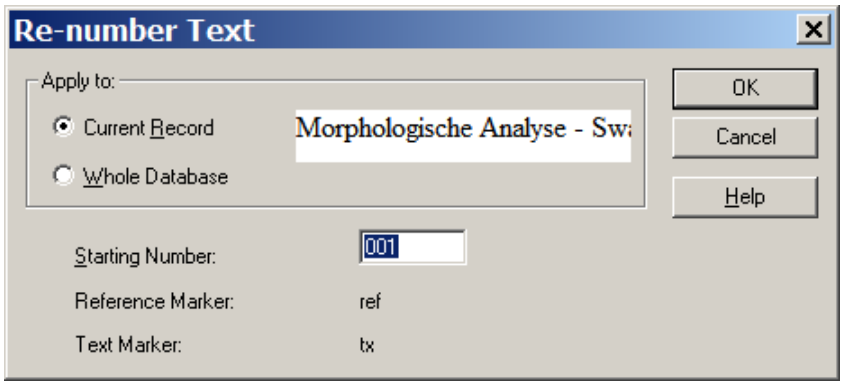
Im Tools-Menü sind dafür zwei Befehle vorgesehen: Break/Number Text ... und Renumber Text...

1. Break/Number Text … dient dazu einen zusammenhängenden Text in Textfelder zu zerlegen und dabei gleichzeitig nach bestimmten Vorgaben zu numerieren.

2. Mit Renumber Text ... kann ein bereits aus Textfeldern bestehender Text neu numeriert werden.

Da unser Text bereits aus Textfeldern besteht und auch Referenzfelder enthält, scheint für uns nur die zweite Möglichkeit in Betracht zu kommen.

- Wählen Sie den Menü-Befehl Tools >> Renumber Text.



Wir haben bei der Definition des Datenbanktyps für Interlineartexte bereits *Reference Marker* und *Text Marker* festgelegt. Die Einstellung für den Anfangswert des Zählers kann beibehalten werden.

- Bestätigen Sie die Einstellungen mit OK.

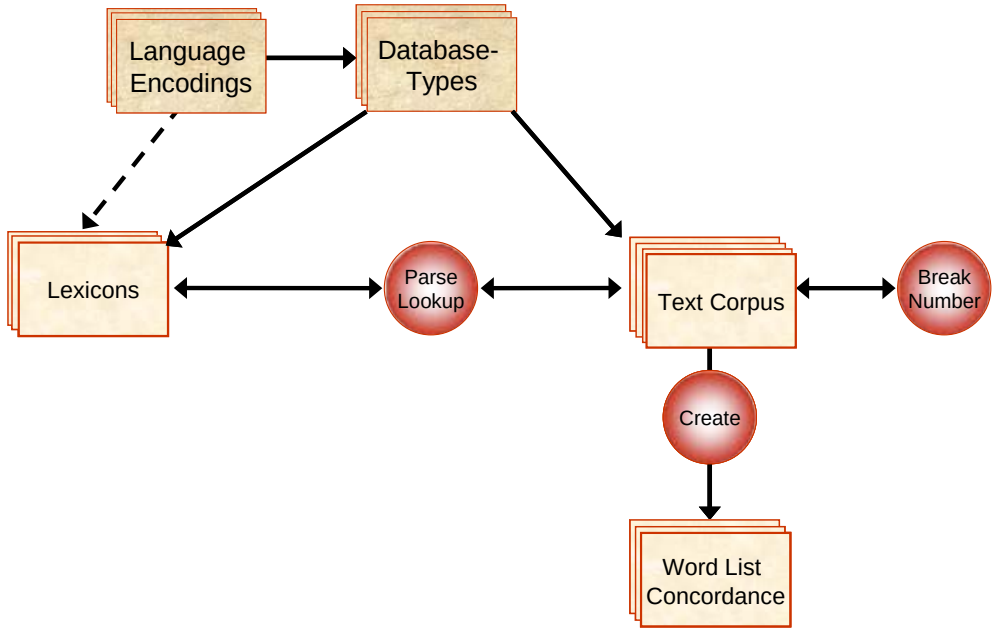
Das Ergebnis ist etwas enttäuschend, weil die Referenzfelder nur die Nummern ohne Bezug zur Textidentität enthalten. Um dies zu reparieren, müssen wir im Referenzfeld zum ersten Textfeld ein Muster vorgeben.

- Modifizieren Sie die Referenznummer des ersten Textfeldes so: Analyse.001
- Nehmen Sie dann die Numerierung erneut vor.

Jetzt erhalten wir das gewünschte Ergebnis.

Die Aufbereitung unseres Textkorpus – zur Zeit nur aus einem Datensatz bestehend – ist damit abgeschlossen und wir könnten mit der Interlinearisierung und dem parallelen Aufbau des zugeordneten Lexikons beginnen. Wir wollen uns jedoch zunächst noch eine weitere Daten-Grundlage für die morphologisch Analyse beschaffen.

Erstellen von Wortlisten aus dem Textkorpus

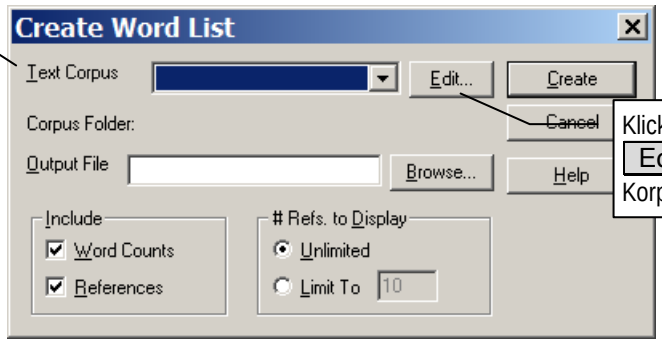


Erstellen einer Wortliste für Swahili

Die erste Aufgabe in der morphologischen Analyse besteht in der Segmentierung der Daten in **Morphe**, wobei ein Morph die kleinste rekurrente (d.h. in anderen Zusammenhängen wiederkehrende) bedeutungstragende Einheit (Phonem- oder Graphem-Sequenz) einer Sprache ist, die nicht weiter in kleinere bedeutungstragende Einheiten zerlegt werden kann, ohne dass die Bedeutung dieser Einheit zerstört wird. Dabei ist zu unterscheiden zwischen Wurzeln und Affixen, die jeweils nach ihrer Funktion (Bedeutung) und ihrer Distribution zu beschreiben sind. Dafür müssen Wortformen verglichen werden. Für diesen Vergleich ist es von Vorteil, wenn man über Listen der in den Daten vorkommenden Formen und ihren Kontexten verfügt. Toolbox stellt dafür ein geeignetes Werkzeug zur Verfügung.

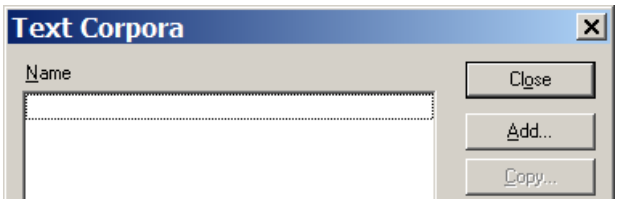
- Wählen Sie Tools >> Wordlist ... Es öffnet sich die folgende Eingabemaske zur Erstellung einer Wortliste (*Create Word List*):

Hier kann man aus einer Liste verfügbarer Korpora ein Text-Korpus auswählen. Da wir noch kein Text-Korpus definiert haben, ist die Liste leer.

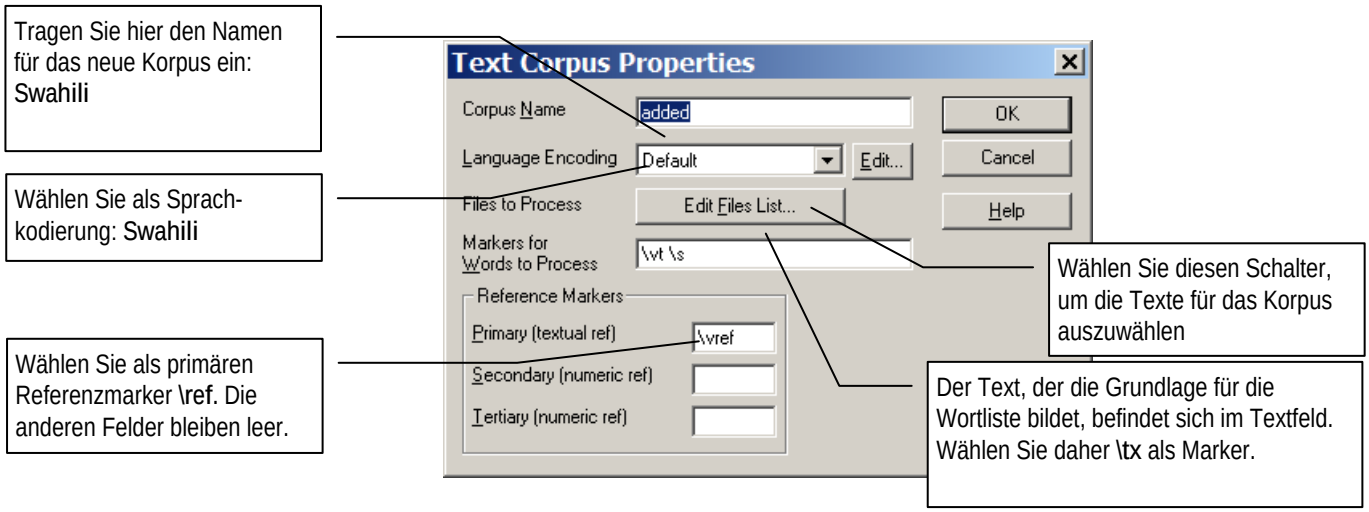


Klicken Sie auf den Schalter Edit..., um der Liste ein Text-Korpus hinzuzufügen.

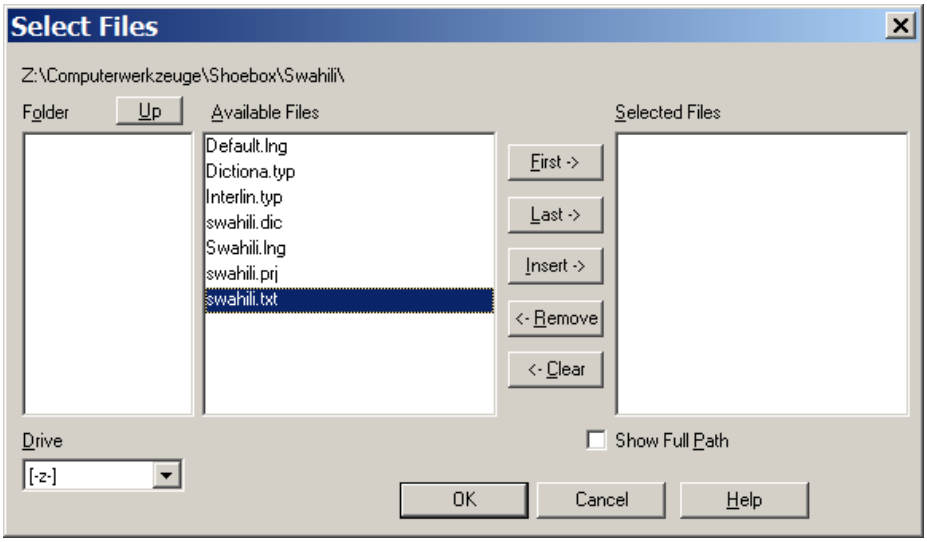
- Wählen Sie den Schalter Edit..., um der Liste der verfügbaren Korpora (derzeit leer) für die Verwaltung der Text-Korpora zu öffnen.



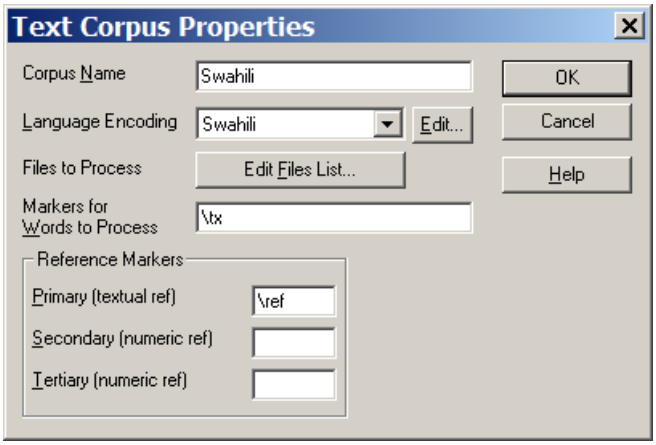
- Wählen Sie den Schalter Add..., um ein neues Text-Korpus zu definieren.



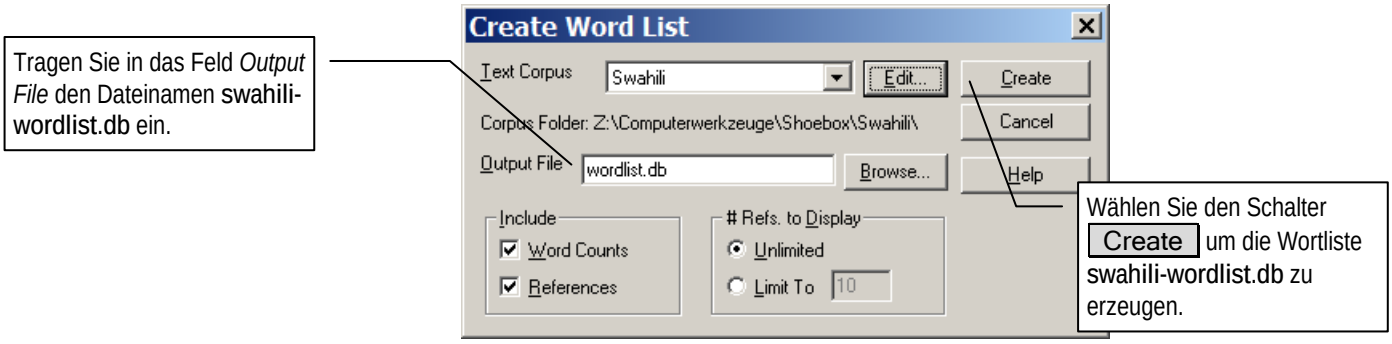
- Tragen Sie als Korpusnamen (*Corpus Name*) **Swahili** ein.
- Wählen Sie als Sprachkodierung (*Language Encoding*) ebenfalls **Swahili**.
- Ersetzen Sie im Feld *Markers for Words to Process* den Eintrag durch \tx.
- Ersetzen Sie im Feld *Primary (textual ref)* den Eintrag durch \ref.
- Wählen Sie den Schalter Edit Files List ...
- Übertragen Sie aus der Liste der *Available Files* die Datei swahili.txt mit einem der Schalter First →, Last →, oder Insert → in das Feld *Selected Files*.



- Wählen Sie OK.



- Wählen Sie erneut OK und anschließend Close.



- Modifizieren Sie den Namen der Ausgabedatei (*Output File*) zu swahili-wordlist.db.

Jetzt haben wir alle Einstellungen vorgenommen, um die Wortliste zu erzeugen:

- Wählen Sie den Schalter Create.

Das folgende Bild zeigt den Anfang der Wortliste in drei Spalten. Die erste Spalte enthält die Wörter in alphabetischer Reihenfolge. Dabei wurden die großen Anfangsbuchstaben in Kleinbuchstaben umgewandelt. In der zweiten Spalte (*Count*) steht, wie oft das Wort im Korpus vorkommt, und in der dritten Spalte sind die Referenzen aufgelistet. Das Wort *kimoja* kommt zweimal vor, und zwar in den Datensätzen Analyse.018 und Analyse.023.

swahili-wordlist.db

Word	Count	References
aliniona	1	Analyse.004
alisoma	1	Analyse.017
anafika	1	Analyse.020
atafaa	1	Analyse.024
atasoma	1	Analyse.010
jana	5	Analyse.013; Analyse.014; Analyse.015; Analyse.016; Analyse.017
kilifaa	1	Analyse.023
kimoja	2	Analyse.018; Analyse.023
kisu	2	Analyse.015; Analyse.023
kitabu	2	Analyse.013; Analyse.018
kitafaa	1	Analyse.018
mliona	1	Analyse.015
mmoja	2	Analyse.020; Analyse.024
mpishi	2	Analyse.017; Analyse.024
mtafika	1	Analyse.001
mtoto	1	Analyse.020
ninafika	1	Analyse.002
nitakisoma	1	Analyse.011
nitawaona	1	Analyse.005
tulifika	1	Analyse.014
tutakuona	1	Analyse.007
tutawaona	1	Analyse.006

Im nächsten Abschnitt soll gezeigt werden, wie diese Listen für die morphologische Analyse eingesetzt werden können.

Aufgabe: Erstellen einer Wortliste für Deutsch

Erstellen Sie nach dem oben für Swahili beschriebenen Muster aus der freien Übersetzung (Markierung \ft) eine deutsche Wortliste. Die zugrunde liegende Datei (in *Files to Process*) ist dieselbe, nämlich swahili.txt. **Wichtig:** Der Name des Text-Corpus muss jedoch anders lauten als für die Swahili-Wortliste. Wählen Sie daher als Namen für dieses Text-Corpus Swahili-Übersetzung.

Geben Sie der Wortliste den Dateinamen (*Output File*) german-wordlist.db. **Wichtig:** Damit deutsche Wörter richtig erkannt werden, muss eine Sprachkodierung für Deutsch erstellt und zugeordnet werden, in der an allen relevanten Stellen die Umlaute und 'ß' aufgeführt werden.

Wortformeln

Einige Affixe des Swahili sind *kategorial* (Marker: \ps) und/oder *semantisch* mehrdeutig.[6] Das Präfix *ki-* z.B. fungiert einerseits als Klassen-Marker (KM) mit der Bedeutung KL4.Sg, andererseits als Subjekt-Marker (SM) und Objekt-Marker (OM). Das führt dazu, dass beim Ausführen des *Parse*-Prozesses während der Interlinearisierung sehr häufig ein Dialog mit dem Benutzer geführt werden muss, um diese Ambiguitäten aufzulösen. Das ist auf die Dauer ziemlich lästig.

Der *Parse*-Prozess lässt sich jedoch steuern durch sog. Wortformeln (engl. *word formulas*). Wortformeln sind Wortstruktur-Regeln, die wie Konstituentenregeln in der Syntax hierarchisch aufgebaut sein können. Für unsere Swahili-Texte gelten folgende Regeln:

(1) Word → { Nominal | Verbal }

(2) Nominal → { KM N | KM A }

(3) Verbal → SM TAM (OM) V

In der Standardeinstellung ist das Startsymbol (*primary symbol*) Word.[7] Die erste Regel besagt, dass ein (komplexes) Swahili-Wort entweder eine Nominal-Form ist (z.B. *kitabu* 'Buch' oder *kimoja* 'eins') oder eine Verbal-Form (z.B. *nitafaa*). Zu den Nominal-Formen zählen auch die Adjektive.

Die zweite Regel drückt aus, dass eine Nominal-Form entweder aus einer Folge von Klassen-Marker (KM) und Nomen (N), oder von Klassen-Marker und Adjektiv (A) besteht.

Eine Verbal-Form (Regel 3) setzt sich aus Subjektmarker (SM), Tempus-Aspekt-Marker (TAM), optional einem Objekt-Marker (OM) und dem Verbstamm (V) zusammen. Fakultative Konstituenten werden eingeklammert.

In Toolbox werden Wortformeln in Verbindung mit dem *Parse*-Prozess – die Zerlegung der Wortformen in Morph(em)-Folgen – definiert. Wir müssen daher den Datenbanktyp für die Interlinearisierung modifzieren.

Eine Formel besteht aus einem der Symbole, die in den obigen Regeln auf der linken Seite des Pfeils stehen, wobei das Startsymbol (*primary*) standardmäßg Word ist, und einer Menge von *patterns*, die den Ausdrücken auf der rechten Seite des Pfeils entsprechen. Dem Symbol Word sind also die Muster (patterns) Nominal und Verbal zuzuordnen.

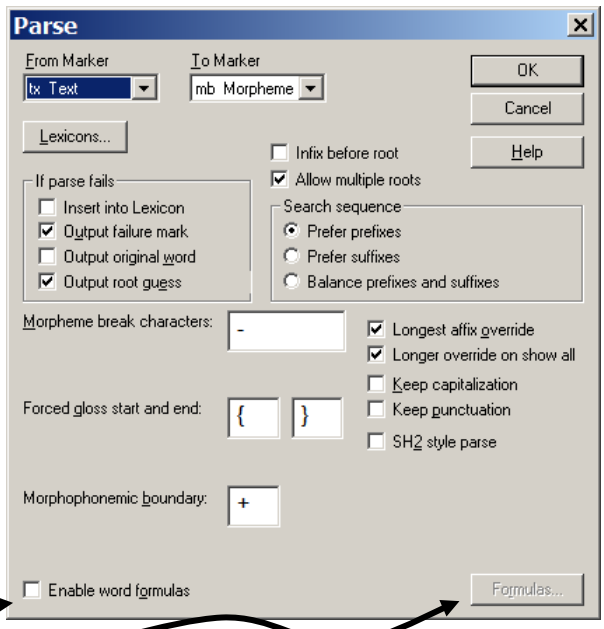
---

[6] *Kategorial* bezieht sich hier auf die Angaben im \ps-Feld, *semantisch* auf die im \gl-Feld. Die Ausführungen zu den Wortformeln setzen die Kenntnis des Textes *Swahili-Aufgabe-Kommentar* voraus.

[7] Das kann geändert werden; wir wollen es jedoch bei der Standardeinstallung belassen.

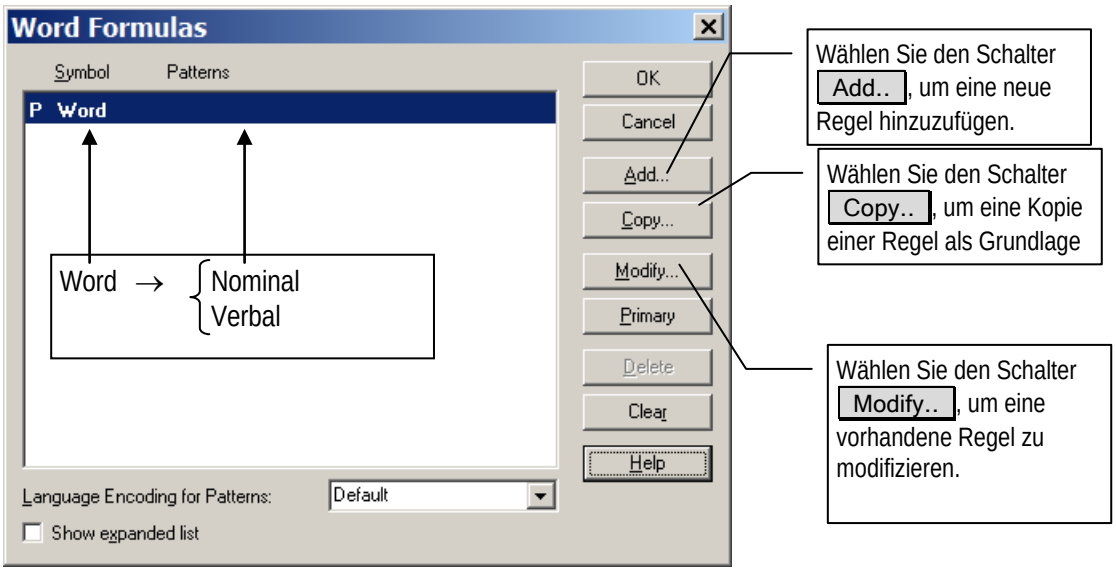
- Wählen Sie den Menübefehl Project >> Database Types ... Wählen Sie aus der Liste den Eintrag Interlinear und dann den Schalter Modify...
- Öffnen Sie das Register Interlinear, wählen Sie den Eintrag für den Parse-Prozess und dann wiederum den Schalter Modify...

In dem Eigenschaftsfenster für Parse sehen Sie links unten ein Feld mit der Beschriftung Enable word formulas und rechts einen – inaktiven – Schalter mit der Aufschrift Formulas.

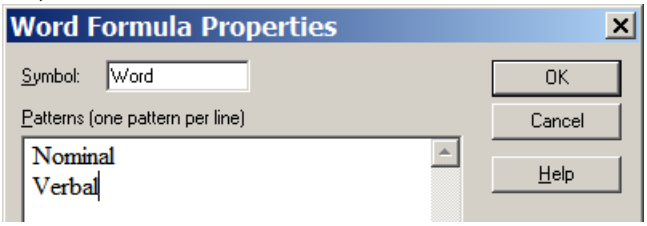


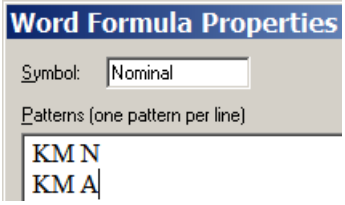
- Aktivieren Sie das Feld Enable word formulas.
- Klicken Sie dann den Schalter Formulas... an.

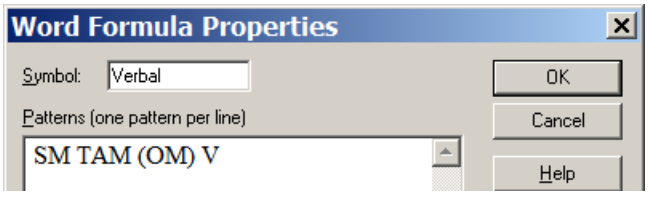
Das sich öffnende Dialogfenster für Word Formulas zeigt eine Liste der bereits definierten Formeln in zwei Spalten, wobei in der linken Spalte die Symbole stehen (entspricht den Symbolen links vom Pfeil) und in der rechten Spalte, jeweils in einer separaten Zeile, die *patterns* (entspricht den Symbolen rechts vom Pfeil). Festgelegt ist zunächst nur das Startsymbol Word (als Startsymbol markiert durch den Buchstaben P - für *primary*), dem jedoch noch keine Muster zugeordnet sind. Das müssen wir jetzt nachholen.



Um die erste Regel (Word → {Nominal, Verbal}) als Wortformel zu definieren, müssen wir den vorhandenen Eintrag für das Symbol Word modifizieren und die *Patterns* Nominal und Verbal hinzufügen. Im weiteren müssen dann die Wort-strukturen für Nominal- und Verbal-Formen spezifiziert werden.

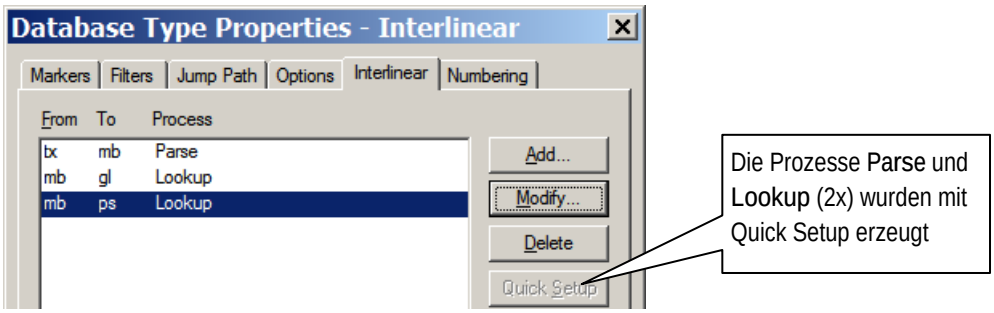






Leider gibt es bei der Verwendung dieser Wortstruktur-Grammatik ein gravierendes Problem, das mit einer Ungereimtheit im Zusammenwirken verschiedener Programmkomponenten zu tun hat.

Die automatische Anwendung einer Wortformel wie Verbal → SM TAM (OM) V setzt voraus, dass die kategoriale Information (hier: SM, TAM, OM, V) verfügbar ist. Wenn man jedoch bei der Einrichtung eines Datenbanktyps für die Interlinearisierung die Prozesse automatisch durch *Quick Setup* vornehmen lässt, werden diese in der falschen Reihenfolge erzeugt:



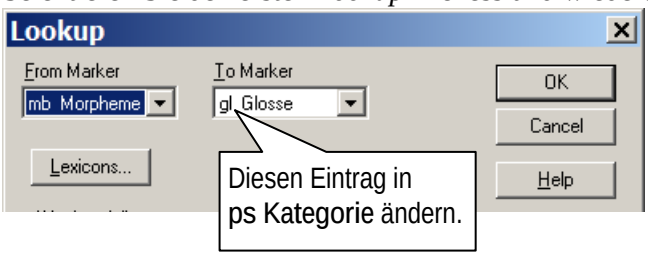
Diese Anordnung führt dazu, dass bei der Interlinearisierung die Glossen **vor** den Kategorien zu stehen kommen, und daher für die Anwendung der Wortformeln nicht zur Verfügung stehen.

Es gibt zwei Möglichkeiten mit diesem Problem umzugehen:

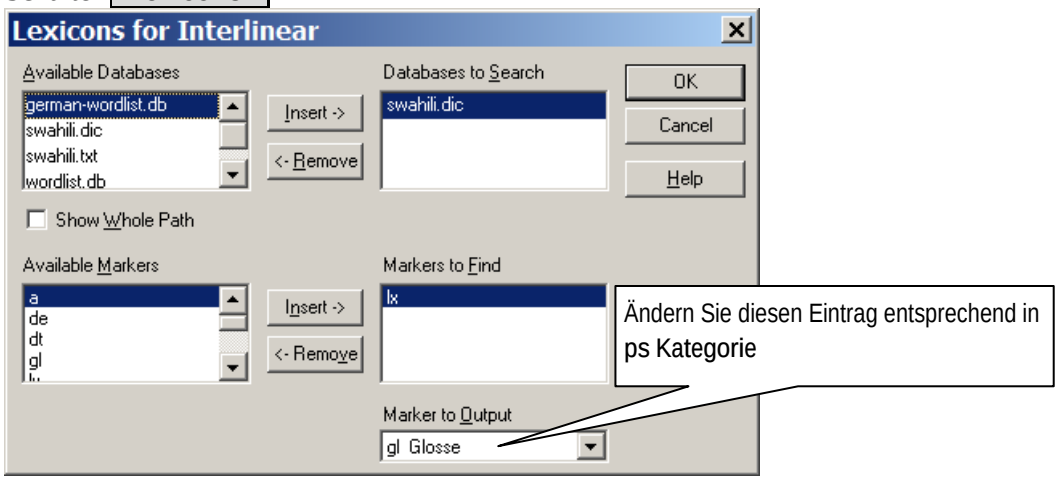
1. Man definiert die relevanten Interlinearisierungsprozesse (*Parse* und *Lookup*) selbst sozusagen manuell. Dann muss man aber genau wissen, was dabei im Detail zu tun ist.
2. Man modifiziert die von *Quick Setup* angelegten Prozesse derart, dass zuerst die Kategorien nachgeschlagen werden und dann erst die Glossen.

Der komplizierteste Prozess ist *Parse*. Es bietet sich daher eher die 2. Methode an. Was wir erreichen müssen ist, dass zuerst die Morpheme (mb) auf die Kategorien (ps) abgebildet werden und dann auf die Glossen (gl). Folgendes ist zu tun:

- Wählen Sie den Menübefehl Project >> Databasetypes…
- Wählen Sie den Eintrag Interlinear und dann den Schalter Modify… und dann das Register Interlinear.
- Selektieren Sie den ersten *Lookup*-Prozess und wiederum den Schalter Modify…



- Ändern Sie in der Liste To Marker den Eintrag in ps Kategorie. Wählen Sie dann den Schalter Lexicons...



- Ändern Sie den Marker to Output (Marker, der ausgegeben werde soll) von gl Glosse in ps Kategorie und bestätigen Sie 2 x mit OK

- Modifizieren Sie analog den 2. *Lookup-Prozess* derart, dass dadurch die Glossen ausgegeben werden.

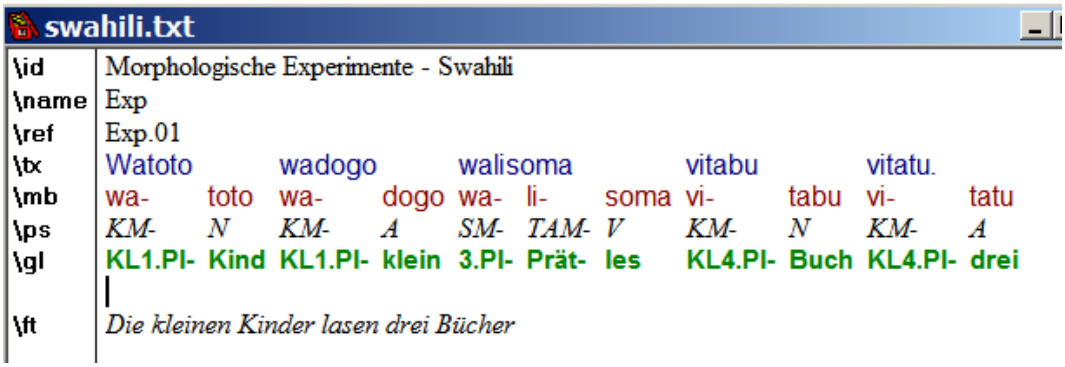
Als nächstes soll gezeigt werden, dass diese Wortstruktur-Grammatik mit den vorgenommenen Änderungen tatsächlich den gewünschten Effekt hat. Dazu soll in der Textdatenbank swahili.txt ein neuer Datensatz angelegt werden, und zwar mit folgenden Eigenschaften:

\id Morphologische Experimente - Swahili
\name Exp
\ref Exp.01

- Aktivieren Sie das Fenster swahili.txt und fügen Sie einen neuen Datensatz ein, entweder mit dem Menübefehl Database >> Insert Record... oder mit der Tastenkombination Strg+N.
- Machen Sie in den Feldern \name und \ref die entsprechenden Einträge.

Wenn Sie nach der Eingabe des Textes im \ref-Feld die Eingabetaste drücken, wird automatisch ein Textfeld (\tx) eingefügt.

- Tragen Sie in das Textfeld den Text *Watoto wadogo walisoma vitabu vitatu* ein und drücken Sie dann die Eingabetaste.
- Tragen Sie in das automatisch eingefügte Feld für die freie Übersetzung (\ft) den Text *Die kleinen Kinder lasen drei Bücher* ein.
- Interlinearisieren Sie dieses Beispiel: Wenn Sie alles richtig gemacht haben, erhalten Sie das richtige Ergebnis, ohne dass Sie Ambiguitäten manuell auflösen müssen.



swahili.txt

\id	Morphologische Experimente - Swahili										
\name	Exp										
\ref	Exp.01										
\tx	Watoto		wadogo		walisoma			vitabu		vitatu.	
\mb	wa-	toto	wa-	dogo	wa-	li-	soma	vi-	tabu	vi-	tatu
\ps	*KM-*	*N*	*KM-*	*A*	*SM-*	*TAM-*	*V*	*KM-*	*N*	*KM-*	*A*
\gl	KL1.Pl-	Kind	KL1.Pl-	klein	3.Pl-	Prät-	les	KL4.Pl-	Buch	KL4.Pl-	drei
\ft	*Die kleinen Kinder lasen drei Bücher*										

- Bearbeiten Sie auf die gleiche Weise die folgenden Sätze:
  Mpishi alinunua visu vikubwa vitatu. – Der Koch kaufte drei große Messer.
  Wapishi alivinunua – Die Köche kauften sie.
  Mpishi mpya atapika supu[8] – Der neue Koch wird Suppe kochen.
  Mganga mzee alimponya mgonjwa – Der alte Doktor heilte den Patienten

Komredi Kipepe na kisa cha Bi Arafa

Im Folgenden geht es darum, einen Original-Comicstrip zu bearbeiten (Vorlage s. separaten Text). Hierbei soll – technisch betrachet – zunächst gezeigt werden, wie in Toolbox ein längerer Text automatisch in Sätze aufgegliedert und mit Referenzfeldern versehen werden kann.

- Legen Sie zuerst einen neuen Datensatz mit den folgenden Daten an. Der Eintrag im Textfeld entspricht den Texten im Comicstrip.

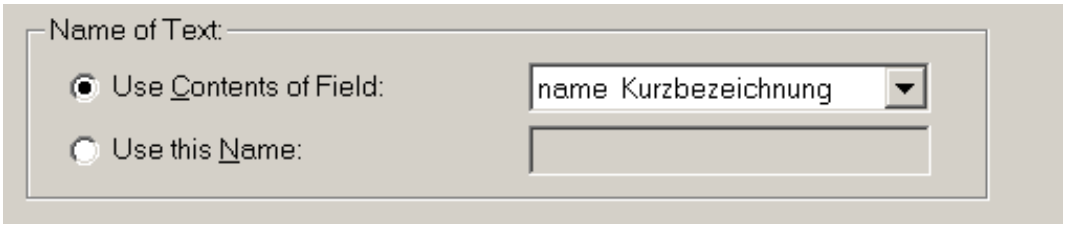
\id Komredi Kipepe na kisa cha Bi Arafa
\name Kipepe
\ref

\tx Komredi Kipepe na kisa cha Bi Arafa. Katika kijiji cha Wabush, Bi Arafa, mchawi na mganga mkuu, anajitayarisha kwenda kwenye tamasha la waganga kwenye msitu wa Gambush. Nina wasiwasi, Bi Arafa. Unajua ni hatari kwenda peke yako msitu wa Gambush. Usijali. Hapana, acha nikusindikize. Komredi, unajua wazi wa ni waganga pekee wanaoruhusiwa huko huoni kwamba. najua sana watu wa kawaida haturuhusiwi kwenye tamasha lenu mimi nitakusubiri nje. Sawa kama mwenyewe umeridhika! Mara. na mimi nitakusindikiza Bi Arafa. kumbe alikuwa Madenge! Wote watatu walianza safari ya kuelekea kwenye msitu wa Gambush.

- Stellen Sie sicher, dass das Textfenster swahili.txt aktiv und darin der neue Datensatz ausgewählt ist.
- Wählen Sie dann den Menübefehl Tools >> Break/Number Text...

Da wir die Numerierungseigenschaften bereits bei der Einrichtung des Datenbanktyps für die Interlinearisierung festgelegt haben, ist nur eine Einstellung zu machen:

- Wählen Sie als Grundlage für die Numerierung das Feld \name (Kurzbezeichnung, s. Bild).



- Bestätigen Sie die Einstellung mit OK

[8] Das Nomen *supu* ist unveränderlich.

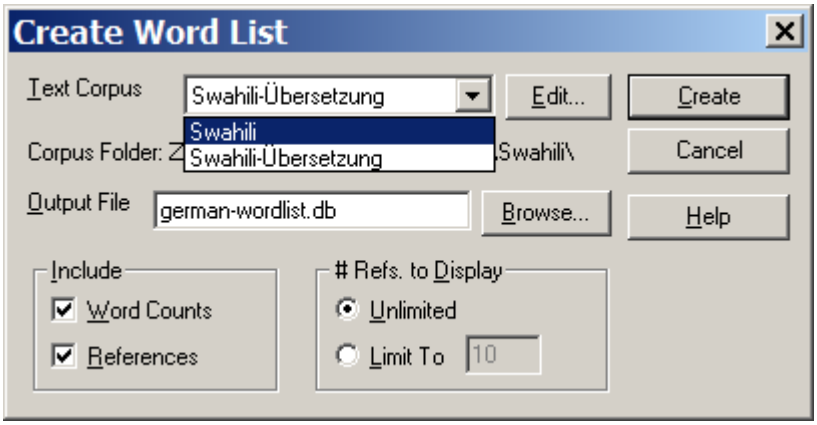
Der Text sollte jetzt wunschgemäß in Sätze zerlegt sein.

Wenn Sie die Schreibmarke an das Ende eines Textfeldes setzen und dann die Eingabetaste drücken, wird automatisch ein Feld für die freie Übersetzung eingefügt.

- Tragen Sie auf diese Weise für jeden Satz die Übersetzung aus der Textvorlage ein.

Ergänzung der Wortlisten

- Bringen Sie die Wortlisten auf den neuesten Stand: Wählen Sie zunächst den entsprechenden Menübefehl Tools >> Wordlist ...
- Wählen Sie dann aus der Liste das jeweilige Text-Corpus aus (Swahili bzw. Swahili-Übersetzung)
- Wählen Sie den Schalter Create.



Die Wortlisten werden um die neuen Einträge ergänzt.

Wir machen ein Wörterbuch

Die nächste Aufgabe, der wir uns stellen wollen, wird sein, die Voraussetzungen zu schaffen für ein Swahili-Deutsch Wörterbuch, das ein einigermaßen professionelles Aussehen haben soll, wie es der unten stehende Ausschnitt zeigt.

Das Toolbox-Wörterbuch für Swahili, mit dem wir bisher gearbeitet haben, hat vorrangig der morphologischen Analyse und der Interlinearisierung gedient. Die Lexikoneinträge bestehen daher im Wesentlichen aus Affixen und Wurzeln bzw. Stämmen. Für die Interlinearisierung sind die Glossen möglichst einfach gehalten und dürfen nicht mit genauen Bedeutungsangaben verwechselt werden. Soweit die Affixe betroffen sind, gehören Details ohnehin in die Grammatik und nicht in ein Wörterbuch. Bei den lexikalischen Morphemen jedoch erwarten wir in einem "richtigen" Wörterbuch mehr an Information.

Zitierformen

Verschiedene Sprachen haben verschiedene Konventionen, nach denen lexikalische Elemente (z.B. Wörter) in einem Wörterbuch einsortiert werden. Diese Formen entsprechen weitgehend den sog. "Nennformen" oder "Zitierformen". Im Deutschen ist die Zitierform der Substantive der Nominativ Singular (z.B. *Turm*), die der Verben der Infinitiv (z.B. *türmen*).

Nominalklassen

Für das Swahili als Bantusprache ist das Nominalklassensystem ein charakteristisches Merkmal. Diese Klassen werden – wie wir gesehen haben – durch Präfixe markiert. In der traditionellen Bantuphilologie werden mehr als 20 solcher Klassen unterschieden, die auch charakteristische semantische Gemeinsamkeiten aufweisen.

Traditionellerweise gehört z.B. *mtu* 'Person' zur Klasse 1, *watu* 'Personen' dagegen zur Klasse 2. Die Formen *mtu* und *watu* bilden jedoch offensichtlich ein Paar, das zur gleichen lexiko-semantischen Klasse gehört, dessen Formen sich hinsichtlich der Kategorie Numerus unterscheiden. Das Präfix *m-* drückt Singular aus, das Präfix *wa-* hingegen Plural. Es bietet sich daher an, entgegen der traditionellen Klassifikation, die Paare zu gemeinsamen Klassen zusammenzufassen. Das haben wir auch bereits getan, indem wir das Präfix *m-* (Allomorph *mw-*) als KL1.Sg und das Präfix *wa-* als KL1.Pl glossiert haben. Das Paar *kitu* 'Ding, Gegenstand' (trad. Kl.7) und *vitu* 'Dinge, Gegenstände' (trad. Kl. 8) haben wir in die neue Klasse 4 eingeordnet.

A - a

-a *Prt.* Allgemeine Relationspartikel vergleichbar mit der englischen Präposition *of*. **msitu wa Gambush** der Wald von Gambush.

a- *SM.* Subjektmarker der 3.Sg (er/sie) und Kongruenzmarker der Personenklasse (Kl. 1). **aliona** er sah mich.

acha *V.* erlauben, lassen.

anza *V.* beginnen, anfangen. **walianza safari** sie begannen eine Reise.

B - b

baba *N.* Vater, Vorfahr. *Pl:* **baba.** *Kl:* **5. baba wa mtoto** der Vater des Kindes.

babu *N.* Großvater. *Kl:* **5. baba ni mtoto wa babu** der Vater ist das Kind des Großvaters.

Bi Arafa *Name. Bi Arafa* ist ein Frauenname. **Bi Arafa ni mganga** Bi Arafa ist eine Heilerin.

Nominalklassen (Fortsetzung)

In Swahili Wörterbüchern ist die Zitierform für das Substantiv daher die Singularform eines Klassenpaares. Den Eintrag für BUCH findet man daher unter *kitabu*. Damit die Klassenzugehörigkeit eindeutig ist, muss in einem Lexikoneintrag auch noch angegeben werden, wie die Pluralform gebildet wird, meist durch Angabe des entsprechenden Klassenpräfixes (im Beispiel *vi-*).

Zitierformen anderer Lexemklassen

Die Zitierformen der anderen Lexemklassen entsprechen den Wurzeln bzw. Stämmen. Beim Verb ist es z.B. der Indikativstamm, der identisch ist mit dem Imperativ Sigular. Der Lexikoneintrag zu einer Verbform wie *ninafika* 'ich komme an' (Imp. *fika*!) findet sich im Wörterbuch also unter *fika*.

Adjektive (ebenso Demonstrativa) kongruieren mit dem Substantiv, das sie modifizieren, und haben daher variable Klassenpräfixe – ähnlich wie sich im Deutschen das Genus der attributiven Adjektive nach dem Genus des Substantivs richtet: ***wa****toto* ***wa****dogo* ***wa****tatu* 'drei kleine Kinder' aber ***vi****su* ***vi****dogo* ***vi****tatu* 'drei kleine Messer'. In Wörterbüchern werden Adjektive in der Stammform aufgeführt, wobei das Fehlen des Präfixes durch einen vorangestellten Bindestrich angezeigt wird: *-dogo* 'klein', *-tatu* 'drei'.

Definitionen

Für die morphologische Analyse und die Interlinearisierung werden in den Glossen möglichst knappe Bedeutungsangaben gemacht, weil es hier weniger um die genaue lexikalische Bedeutung, sondern mehr um die grammatische Funktion geht. Für ein "richtiges" Wörterbuch reicht das nicht aus. Es müssen weiter gehende Definitionen angeboten werden, die beispielsweise für das Textverständnis unerlässlich sind.

Neue Datenfelder

Um diese und weitere für ein Wörterbuch notwendige Informationen zur Verfügung stellen zu können, muss das Lexikon um eine Reihe von Datenfeldern erweitert werden, die für die Analyse und Interlinearisierung keine oder nur eine untergeordnete Rolle spielen. Das Programm Toolbox verfügt über eine Exportfunktion zur Erstellung eines derartigen Wörterbuchs, welche ein ehemals selbständiges Programm mit dem Namen *Multiple Dictionary Formatter* (MDF) integriert. Für dieses Programm wurde ein Satz von über 100 Markern definiert, die dem Programm bekannt sind und mit bestimmten Formatierungseigenschaften (Schriftart, Schriftschnitt, Farbe etc.) verbunden sind. Wir benötigen aus diesem Satz allerdings nur einen Bruchteil.

Multiple Dictionary Formatter: MDF

Lexique Pro: LP

Wir werden allerdings nicht die eingebaute MDF-Funktion verwenden, sondern ein separates Progamm namens Lexique Pro, weil dieses transparenter und flexibler zu handhaben ist. Für dieses Programm gelten also die gleichen MDF-Voraussetzungen. Es ist zwar möglich, bei Bedarf eigene Datenfelder zu definieren, grundsätzlich sollten jedoch die Marker aus MDF und deren vorgegebene Anordnung zugrundegelegt werden. Im folgenden Musterdatensatz sind die neuen Feldmarkierungen **fett** gedruckt (z.B. **\lc** – das ist die "Zitierform", hier die Singularform des Substantivs mit dem Klassenpräfix *m-*).

Musterdatensatz

\lx	pishi
\lc	mpishi
\ps	N
\gl	Koch
\dg	Koch, Köchin
\xv	Mpishi alinunua visu.
\xg	Der Koch kaufte ein Messer.
\lf	Vgl = -pika
\lg	kochen
\mr	m- pik –i
\pdl	Kl
\pdv	1
\pl	wa-
\dt	02/Jun/2005

Die neuen Felder und ihre Marker

\lc Engl.: *lexical citation* – Zitierform, nach der die Einträge im Wörterbuch auch sortiert werden. Das Feld kann leer bleiben, wenn es mit dem Lemmafeld identisch ist. Für Swahili-Substantive ist die Zugehörigkeit zu ihrer Klasse enorm wichtig, daher werden sie in Standardwörterbüchern des Swahili mit ihrem Klassenpräfix im Singular aufgeführt, z.B. *mpishi* 'Koch', und nicht als Wurzeln oder Stämme.

\dg Ersetzt \de (engl. *definition English*). Das **g** steht für *German*. *Lexique Pro* bietet die Möglichkeit, bestimmte Marker durchgehend zu "lokalisieren". Dieses Feld beinhaltet die eigentliche Wörterbuchdefinition (also die Bedeutungsangabe), die viel expliziter sein kann (und sollte) als eine Glosse. Insbesondere lassen sich so Bedeutungsvarianten angeben.

\lf Engl.: *lexical function* (lexikalische Funktion). Damit sind systematische Beziehungen zwischen Wörtern gemeint wie SYNONYMIE (Bedeutungsgleichheit), ANTONYMIE (Bedeutungsgegensätzlichkeit), HYPONYMIE (begriffliche Unterordnung), PARTONYMIE (Teil-Ganzes-Beziehung) etc. Eine lexikalische Funktion wird in der Form Funktion = Wert angegeben. Um im Lexikoneintrag z.B. für *–dogo* 'klein' das Antonym (Ant) *–kubwa* 'groß' anzuführen, müsste man die folgenden Angaben machen: [\lf Ant = -kubwa, \lg 'groß'].

Wir werden dieses Feld hauptsächlich als allgemeinen Querverweis missbrauchen. Dafür ist zwar ein eigener Marker – \cf – vorgesehen, der hat aber den Nachteil, dass als Verweisbezeichner das englische *see* eingefügt wird. Beispielsweise würde [\cf fika] als *see:* fika wiedergegeben. Mithilfe des \lf-Markers können wir hingegen die Ausgabe beeinflussen: [\lf Vgl = fika].

\lg Engl.: *lexical function gloss*; das ist nichts anderes als die Übersetzung des Eintrags im \lf-Feld.

Belege

\xv Engl.: ***e**xample **v**ernacular*. In diesem Feld ist ein typisches Beispiel anzugeben, das die Verwendung des Lexikoneintrags gemäß der Definition zeigt.

\xg Engl.: *example gloss German*. In diesem Feld steht die Übersetzung des Beispiels aus **xv**.

Paradigma

\pdl Engl.: *paradigm label*. Hier geht es um Angaben zu einzelnen Formen aus dem Flexionsparadigma eines Lexikoneintrags. Das *label* ist ein Bezeichner wie Singular, Plural, Passiv, Futur (oder eine Abkürzung davon: Sg, Pl, Pass, Fut). Um anzugeben, dass *kisu* eine Singularform ist, könnte man im Lexikoneintrag die Angabe [\pdl Sg, \pdv kisu] machen. Wir verwenden diese Möglichkeit vor allem um für Substantive die Nominal-Klasse anzugeben, beispielsweise für die Form *kisa* 'Geschichte': [\pdl Kl, \pdv 4].

\pdv Engl.: *paradigm value*. Hier wird die Form zu einem entsprechenden *paradigm label* angegeben: [\pdl Sg, \pdv kisu].

\pl Dieses Feld steht für Plural und wird hier verwendet, um bei den Substantiven die Pluralform anzugeben, z.B. *wapishi* zum Sg. *mpishi* in der 1. Nominalklasse.

\mr Engl.: *morphemic representation*. In diesem Feld kann, falls erfordlich, die morphologische Zusammensetzung eines Wortes verdeutlicht werden. Man könnte z.B. für *cha* in *kijiji cha Wabush* 'das Dorf der Wabush' einen eigenen Lexikoneintrag machen, in dem u.a. aufgeführt ist, dass sich *cha* aus dem Präfix *ki-* und der Relationspartikel *–a* zusammensetzt: [\mr ki- a]. Angezeigt wird dies dann als *Morph:* ki- a.

Modifikation der Lexikondatenbank

Wir werden – wie bereits gesagt – das Wörterbuch mit dem Programm *Lexique Pro* bearbeiten. Bevor wir aber mit der Einführung in dieses Programm beginnen, wollen wir die Lexikondatenbank swahili.dic noch für diesen Zweck aufbereiten, indem wir die neuen Datenfelder ergänzen. Für einige Marker ist es sinnvoll festzulegen, welcher Marker als nächstes folgen soll.

Marker	**Name**	**Sprache**	**Marker für Folgefeld**
lx	Lemma	Swahili	lc
lc	Zitierform	Swahili	[none]
dg	Definition	German	[none]
xv	Beispiel	Swahili	xg
xg	Beispiel-Glosse	German	[none]
lf	Funktion	Swahili	lg
lg	Funktions-Glosse	German	[none]
mr	Morphologie	Swahili	[none]
pdl	PD-Name	Default	pdv
pdv	PD-Wert	Default	[none]
pl	Plural	Swahili	[none]

Da die Marker für das Folgefeld (letzte Spalte in der Tabelle) bereits bekannt sein müssen, müssen diese zuerst definiert werden. Sie müssen also den Eintrag für die Zitierform (\lc) bereits gemacht haben, bevor Sie den Marker für das Lemma (\lx) modifizieren können. Gehen Sie folgendermassen vor:

- Aktivieren Sie das Fenster die Lexikondatenbank swahili.dic durch Anklicken und wählen Sie dann den Menübefehl Database >> Properties..., um das Eigenschafts-fenster der Datenbank zu erhalten.
- Wählen Sie den Schalter [Add...], um den neuen Marker lc hinzuzufügen.
- Definieren Sie entsprechend alle weiteren in der obigen Tabelle aufgelisteten Marker.
- Überprüfen Sie vor dem nächsten Schritt, ob alle Marker-Definitionen korrekt sind und insbesondere die Nachfolge-Marker festgelegt sind.
- Modifizieren Sie den Eintrag für *pishi* 'Koch' auf der Grundlage des oben gezeigten Musterdatensatzes. Achten Sie darauf, dass die Marker genau in der Reihenfolge des Musterdatensatzes eingegeben werden.

Das Lexikon-programm *Lexique Pro*

Während ein Projekt in Toolbox bearbeitet wird, sind die Dateien auf der Festplatte mit einem Schreibschutz versehen. Bevor diese Dateien von einem anderen Programm wie *Lexique Pro* bearbeitete werden können, muss dieser Schreibschutz entfernt werden. Dies geschieht automatisch, wenn man Toolbox "geregelt" verlässt.

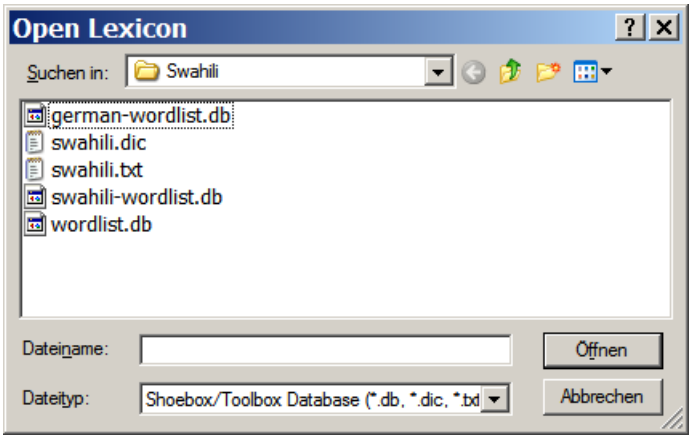
- Schließen Sie Toolbox mit dem Menübefehl File >> Exit oder auf einem anderen Wege. Sie brauchen die offenen Fenster vorher **nicht** zu schließen.

Lexique Pro ist ein eigenständiges Programm, das die Aufgabe der MDF-Funktion von Toolbox übernimmt, aber in der Handhabung viel benutzerfreundlicher ist. Mit *Lexique Pro* kann man in Toolbox erstellte Datenbanken im MDF-Format einlesen und bearbeiten. Die so veränderten Lexikondatenbanken können in Toolbox weiterverwendet werden, beispielsweise für die Interlinearisierung weiterer Texte.

- Starten Sie das Programm *Lexique Pro*. Sie finden es im Startmenü an der gleichen Stelle wie Toolbox. Es erscheint die Startseite des Programmes.

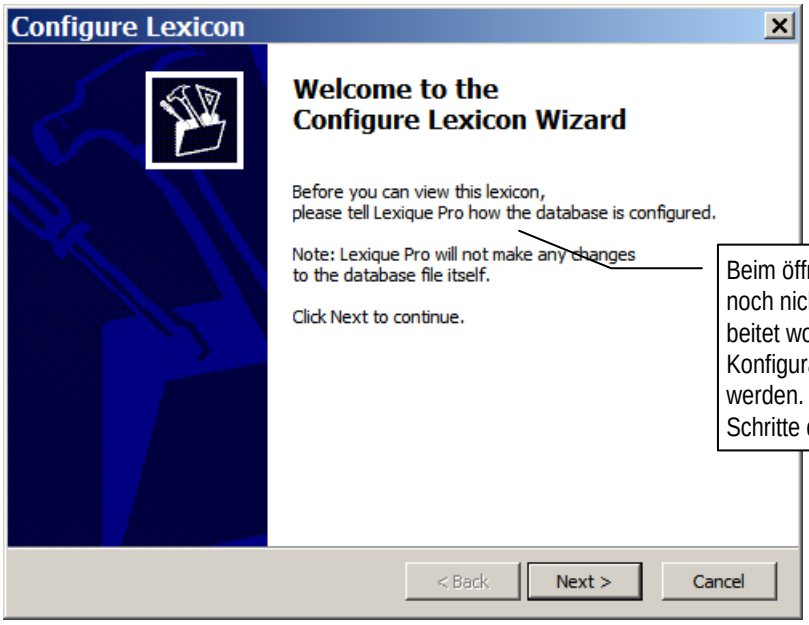
Hier klicken um ein Lexikon zu öffnen.

- Klicken Sie auf das Link Open Lexicon, um ein Lexikon zu laden.

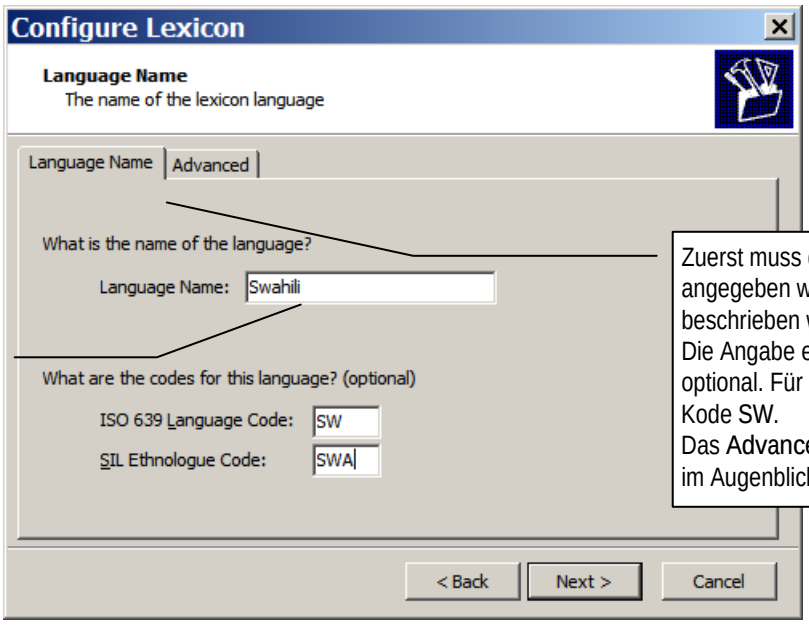


- Navigieren Sie ggf. in das Verzeichnis Swahili in Z:\Toolbox. In der Voreinstellung werden nur Dateien vom Typ Shoebox/Toolbox Database mit den Erweiterungen *.db, *.dic bzw. *.txt angezeigt. Wählen Sie swahili.dic und den Schalter Öffnen, um diese Datenbank zu öffnen.

Als nächstes muss diese Datenbank für *Lexique Pro* konfiguriert werden. Dabei werden eine Reihe von Konfigurationsschritten durchlaufen. Die meisten dieser Einstellungen können auch nachträglich modifiziert werden. Dabei wird eine Konfigurationsdatei angelegt. Die zu öffnende Lexikondatenbank wird bei dieser Konfiguration **nicht** verändert.

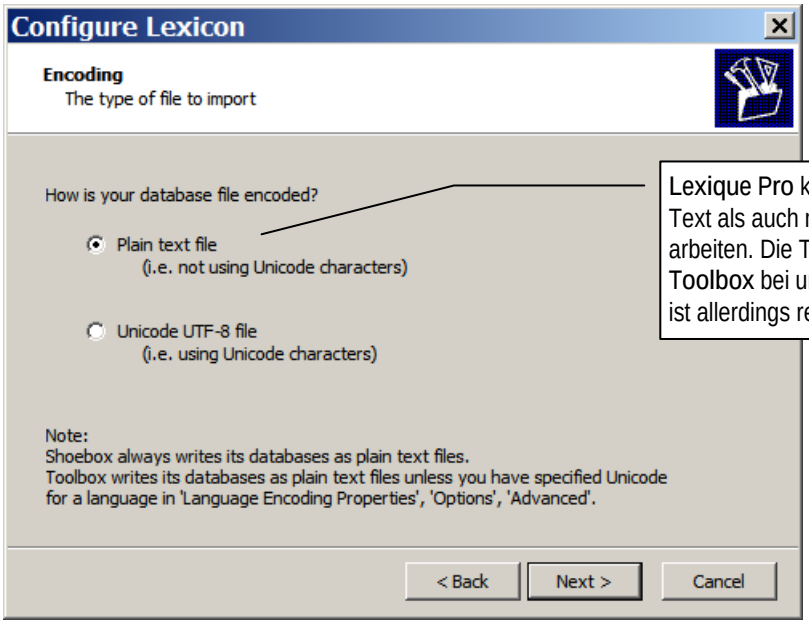


Beim öffnen eines Lexikons, das noch nicht mit *Lexique Pro* bearbeitet worden ist, muss zuerst eine Konfigurationsdatei angelegt werden. Dabei werden mehrere Schritte durchlaufen.

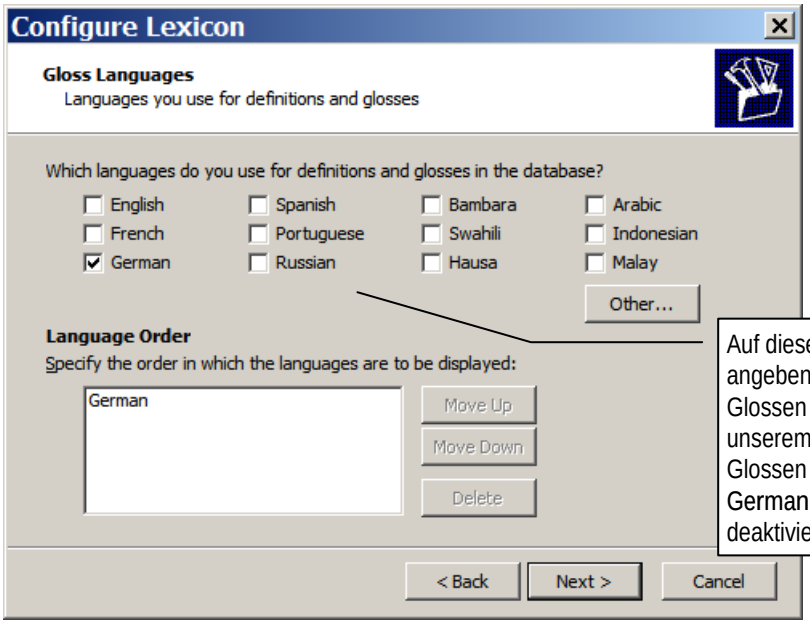


Wählen Sie **Swahili** als Sprachnamen.

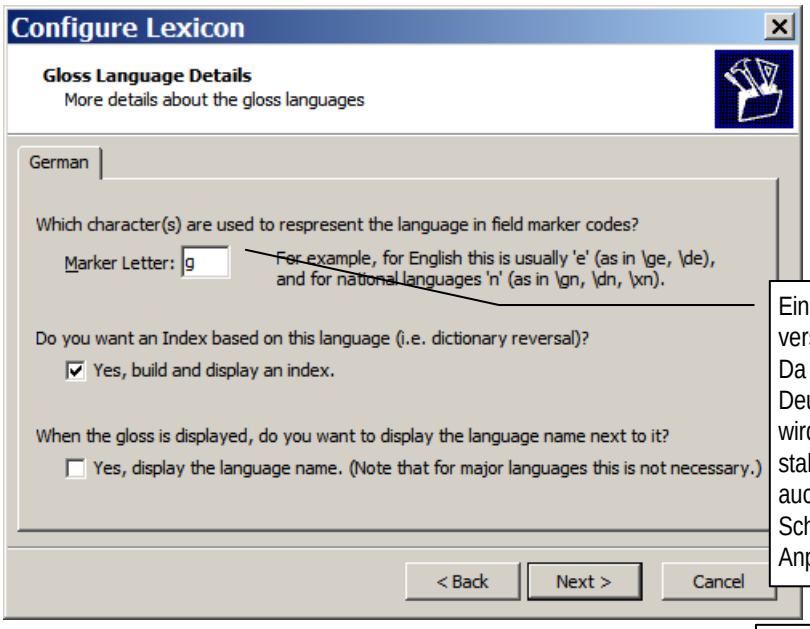
Zuerst muss der Name der Sprache angegeben werden, die im Lexikon beschrieben wird.
Die Angabe eines Länderkodes ist optional. Für **Swahili** wäre der ISO-Kode **SW**.
Das **Advanced** Register wollen wir im Augenblick ignorieren.



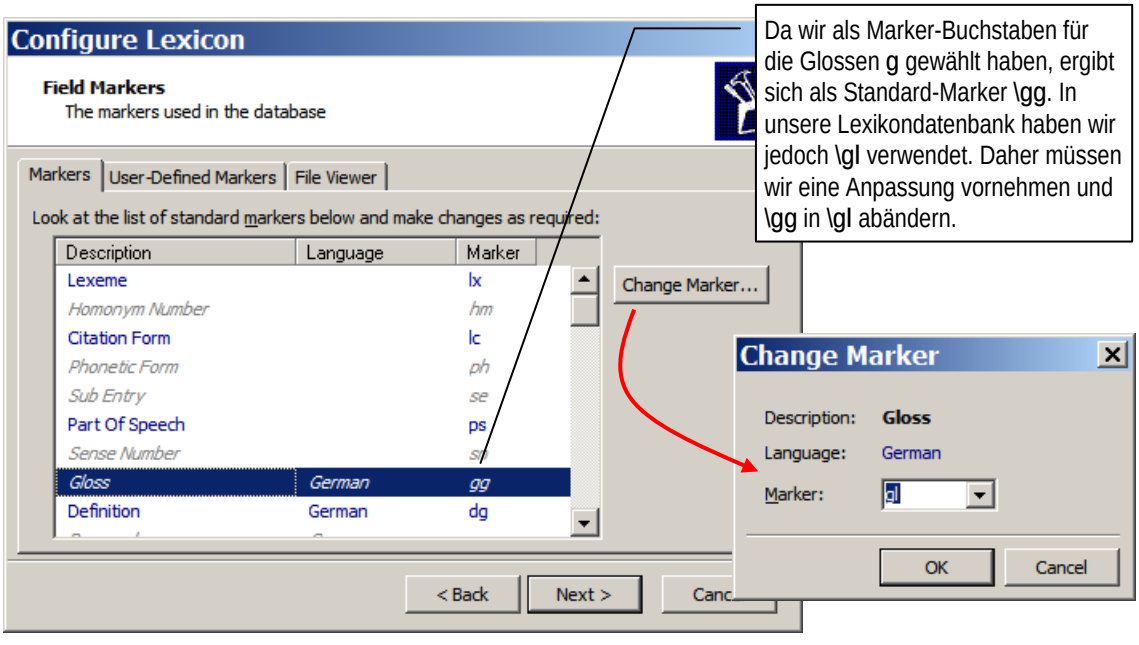
Lexique Pro kann sowohl mit reinem Text als auch mit Unicode-Texten arbeiten. Die Textausgabe von **Toolbox** bei unseren Einstellungen ist allerdings reiner Text.



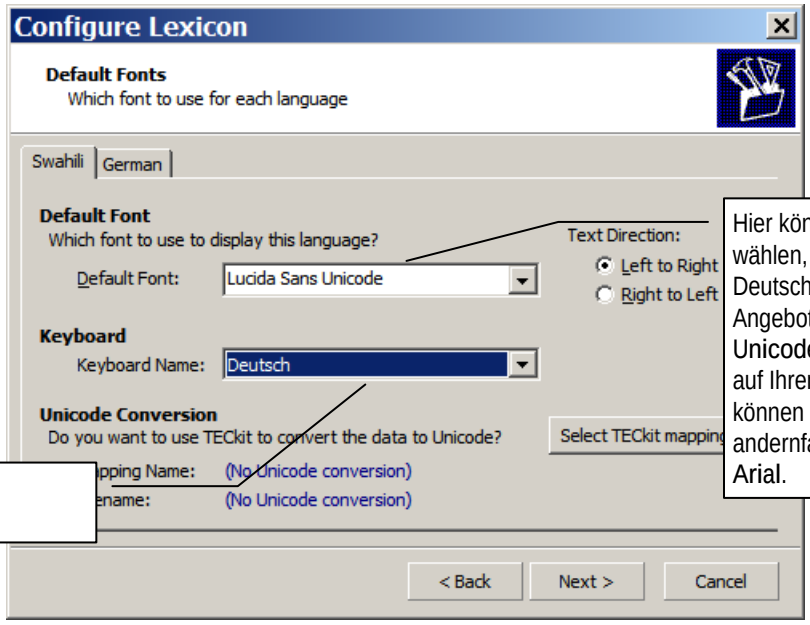
Auf dieser Seite können Sie angeben, welche Sprachen in den Glossen verwendet werden. Da in unserem Lexikon nur deutsche Glossen vorkommen, müssen Sie **German** aktivieren und alle anderen deaktivieren.



Eine Reihe von Markern kommen in verschiedenen Sprachvarianten vor. Da wir als Sprache für die Glossen Deutsch (German) gewählt haben wird zur Differenzierung der Buchstabe **g** angeboten. Dies sollten wir auch so akzeptieren. Im nächsten Schritt muss dann allerdings eine Anpassung vorgenommen werden.

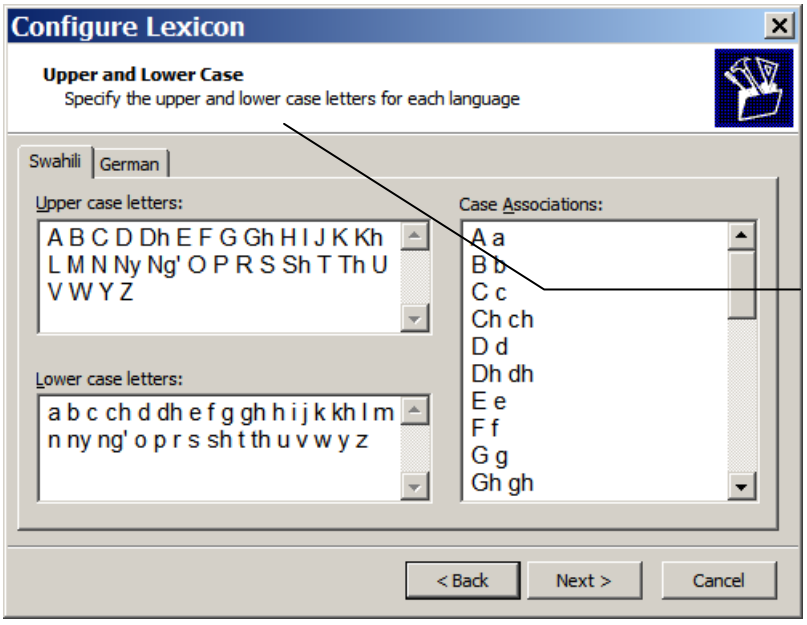


Da wir als Marker-Buchstaben für die Glossen **g** gewählt haben, ergibt sich als Standard-Marker **\gg**. In unsere Lexikondatenbank haben wir jedoch **\gl** verwendet. Daher müssen wir eine Anpassung vornehmen und **\gg** in **\gl** abändern.

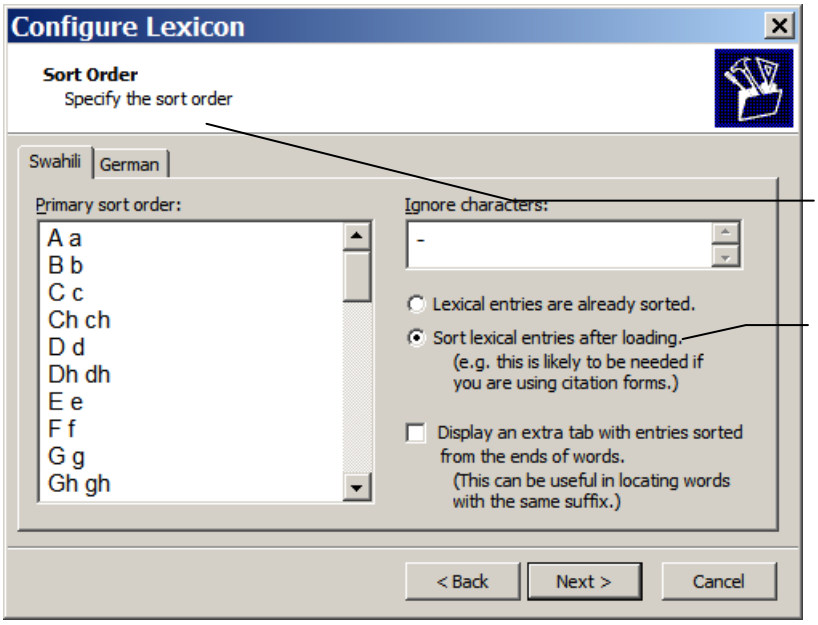


Hier können Sie den Zeichensatz wählen, der für Texte in Swahili bzw. Deutsch verwendet werden soll. Angeboten wird Lucida Sans Unicode. Falls dieser Zeichensatz auf Ihrem Rechner vorhanden ist, können Sie das beibehalten, andernfalls ersetzen Sie ihn durch Arial.

Wählen Sie unbedingt als Tastaturlayout Deutsch!

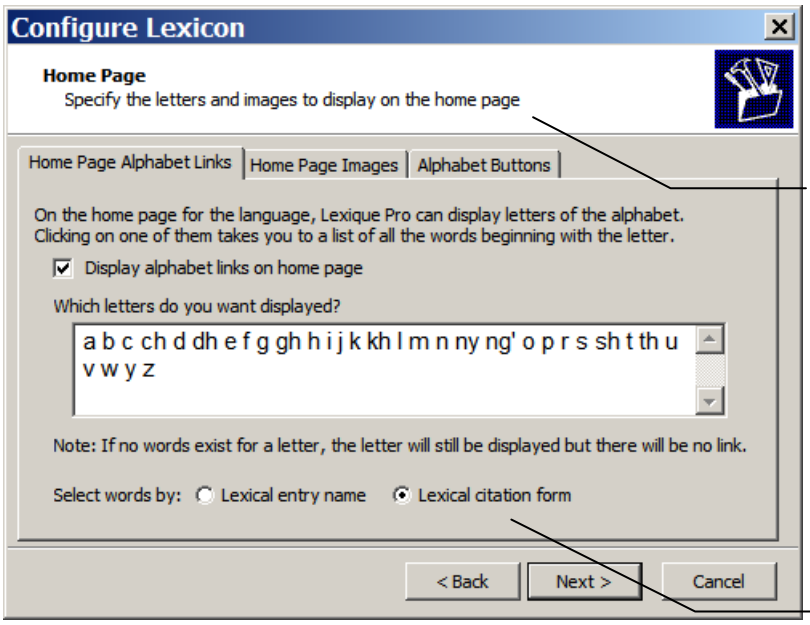


Diese Angaben werden aus der importierten Datenbank übernommen und müssen nicht verändert werden.



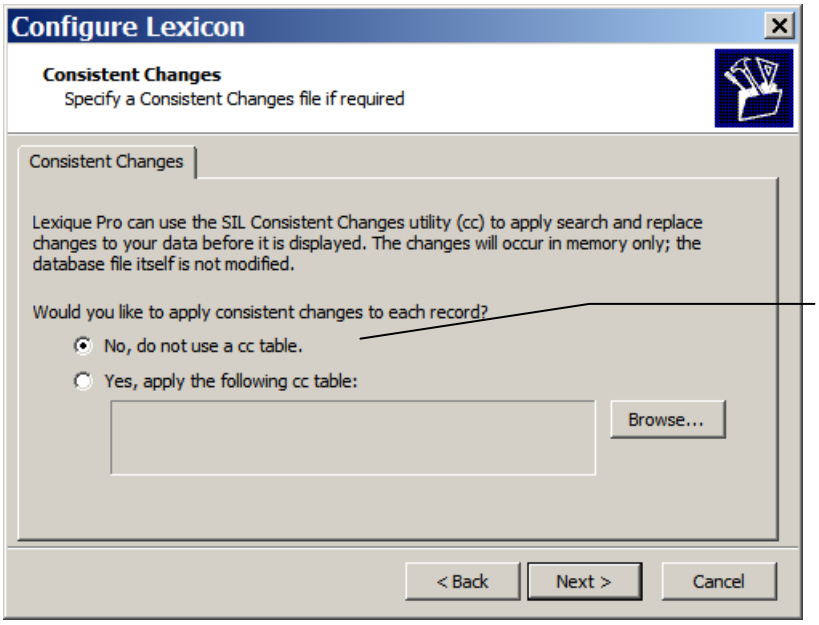
Auch die Sortierreihenfolge wird übernommen.

Da in der Datenbank die Einträge nach den Lemmata (lx) sortiert sind, wir aber Zitierformen (lc) verwenden, muss beim Import unbedingt neu sortiert werden!

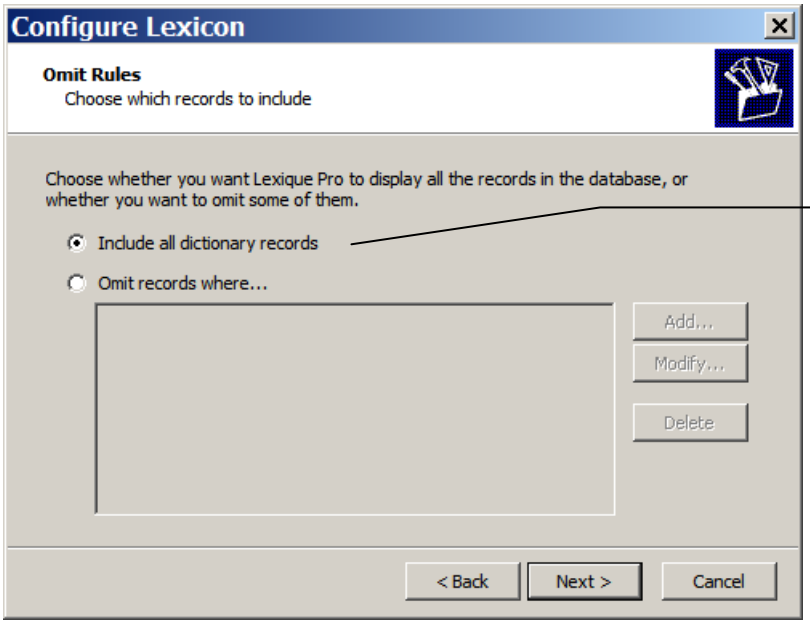


Bei diesen Optionen geht es um die Gestaltung der Startseite des Programms. Damit können wir uns zu einem späteren Zeitpunkt befassen. Wir behalten die Voreinstellungen bei.

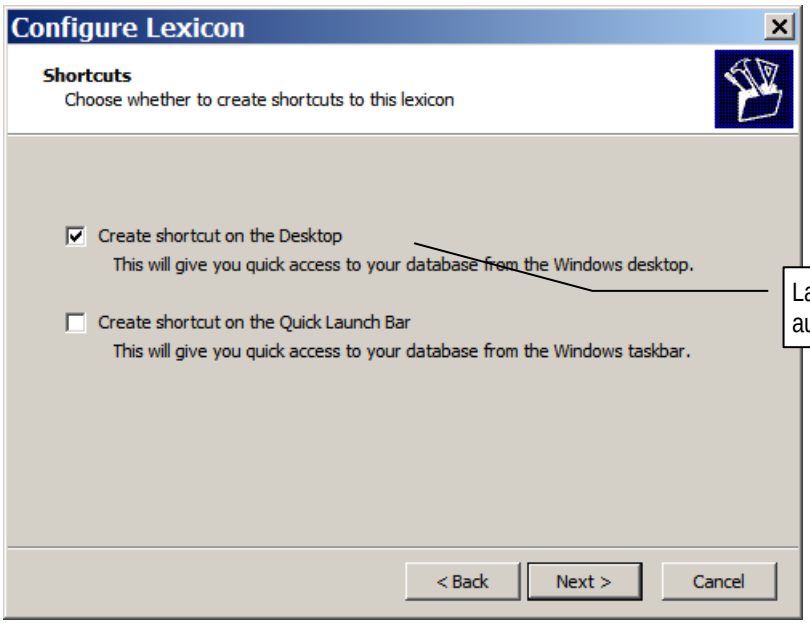
Die Auswahl soll nach der **Zitierform** vorgenommen werden.



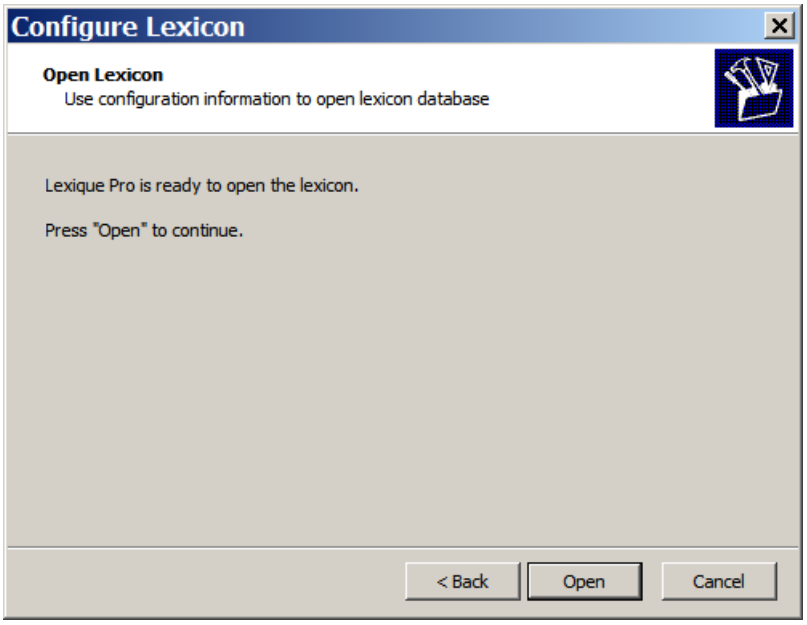
Wir verwenden keine **Consistent Changes**



Wir übernehmen alle Datensätze.



Lassen Sie eine Verknüpfung auf dem Desktop anlegen.

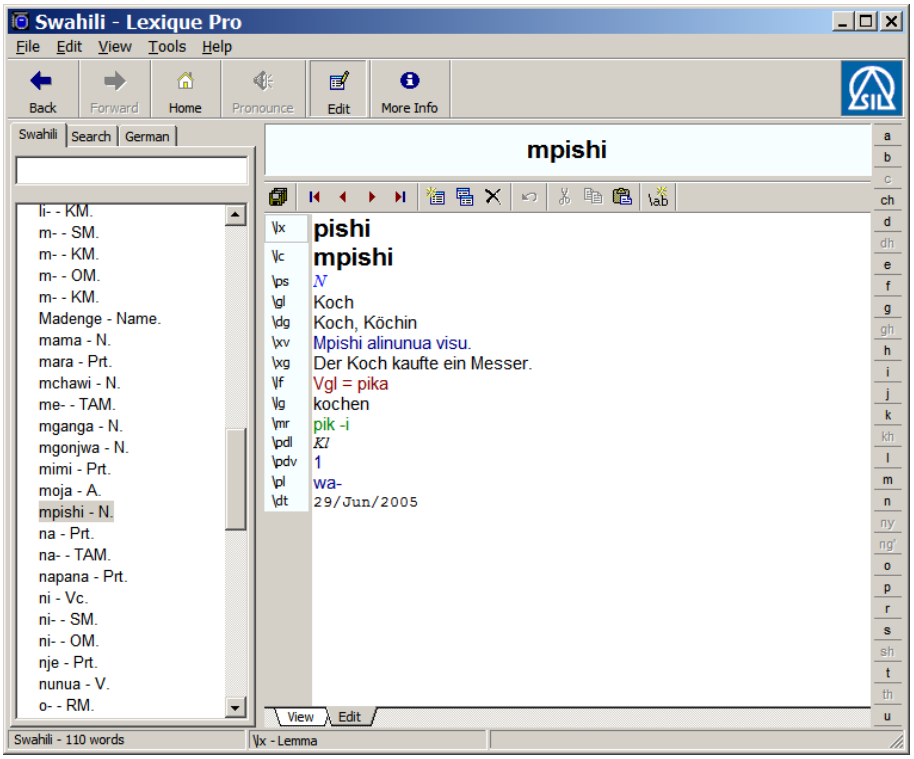


Navigation in Lexique Pro

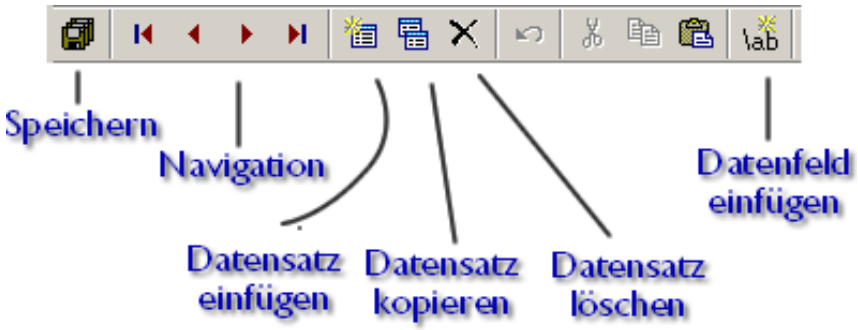
Nach dem erfolgreichen Start von *Lexique Pro*, sehen sie die umseitig abgebildete Startseite. Diese Seite ist wie ein moderner Internet-Browser bzw. eine Hilfedatei aufgebaut. Oben finden Sie eine Menüleiste mit den Menüpunkten File (Datei), Edit (Bearbeiten), View (Ansicht), Tools (Extra) und Help (Hilfe). Darunter ist eine Schalterleiste zur Navigation (Back, Forward, Home) und für bestimmte Aktionen (Pronounce, Edit). In der Mitte befindet sich das Hauptinformationsfenster. Links davon finden Sie ein Navigationsfenster, in dem die Lexikoneinträge mit Angabe der Kategorie alphabetisch aufgelistet sind. Wenn Sie auf einen dieser Einträge klicken, wird der zugehörige Datensatz im Hauptfenster angezeigt. Auf der rechten Seite finden sie die Anfangsbuchstaben senkrecht angeordnet. Wenn Sie auf einen dieser Buchstaben klicken, werden alle Einträge, die mit diesem Buchstaben beginnen, im Hauptfenster angezeigt. Diese Liste enthält wiederum Links, die Sie zu den zugehörigen Datensätzen führt.

- Versuchen Sie, das Programm etwas zu erforschen, in dem Sie alle Navigationsmöglichkeiten ausprobieren.

- Suchen Sie in der Liste den Datensatz für mpishi und lassen Sie ihn im Hauptfenster anzeigen.
- Klicken Sie dann auf den Schalter  (Edit)



Die Darstellung hat sich deutlich verändert. Im Hauptfenster können zwei Ansichten gewählt werden: Edit (zur Bearbeitung des Datensatzes) und View für eine Vorschau. In der Bearbeitungsansicht wird der Datensatz in seiner zugrunde liegenden Struktur dargestellt. Oberhalb des Hauptfensters befindet sich eine Schalterleiste mit einer Reihe von Optionen für die Bearbeitung von Datensätzen. Wenn Sie mit der Maus über die Symbole fahren, wird Ihnen ihre Funktion angezeigt.



Zitierformen für Adjektive

Mit Ausnahme unseres Musterdatensatzes für mpishi haben wir für unsere Lexikoneinträge noch keine Zitierformen definiert. Wo diese identisch mit den Lemmata sind, ist dies auch nicht erforderlich.

Wir wollen mit den Adjektiven beginnen. Wie bereits ausgeführt kongruieren Adjektive (ebenso Demonstrativa) mit dem Substantiv, das sie modifizieren, und haben daher variable Klassenpräfixe – ähnlich wie sich im Deutschen das Genus der attributiven Adjektive nach dem Genus des Substantivs richtet: ***wa**toto **wa**dogo **wa**tatu* 'drei kleine Kinder' aber ***vi**su **vi**dogo **vi**tatu* 'drei kleine Messer'. In Wörterbüchern werden Adjektive in der Stammform aufgeführt, wobei das Fehlen des Präfixes durch einen vorangestellten Bindestrich angezeigt wird: *-dogo* 'klein', *-tatu* 'drei'.

Aufgabe

Fügen Sie für alle in der Datenbank enthaltenen Adjektive die Zitierformen ein. Dies ist eine leichte Aufgabe, da nur ein Bindestrich vor den Stamm gesetzt werden muss. Achten Sie darauf, dass der Bearbeitungsmodus (Edit) aktiviert ist.

- Suchen Sie in der Navigationsleiste im Register Swahili nach dem ersten Adjektiv. Es handelt sich um *dogo* 'klein'. Klicken Sie auf den Eintrag, damit der Datensatz im Hauptfenster angezeigt wird.
- Stellen Sie die Schreibmarke hinter das Lemma (hier also hinter *dogo*). Wenn Sie die Eingabetaste drücken wird automatisch ein Feld für die Zitierform eingefügt.
- Tragen Sie als Wert *–dogo* ein.
- Aktivieren Sie das Vorschaufenster (View). Sie sehen, dass der Datensatz jetzt mit der Zitierform angezeigt wird. Diese ist auch im Titelfenster zu sehen.
- Aktivieren Sie wieder das Bearbeitungsfenster.
- Tragen Sie auf die gleiche Weise die Zitierformen aller anderen Adjektive ein.

Achtung: einige Adjektive sind unveränderlich, nehmen also kein Präfix zu sich. Dazu gehören *peke (yako), sawa, wazi*.

Aufgabe Suchen Sie im Lexikon alle Substantive (Kategorie: N) und nehmen Sie folgende Ergänzungen vor:

1. Tragen Sie – wo erforderlich – die Zitierformen ein.
2. Machen Sie einen Eintrag über die Klassenzugehörigkeit [\pdl Kl, \pdv …]
3. Machen Sie – wo erforderlich – einen Eintrag für die Pluralform.

Die Zitierform für Substantive ist – wie bereits ausgeführt – die Singularform. Folgende Klassen kommen im Lexikon vor:

Klasse 1 (traditionell 1/2): Sg. *m/mw-* Pl. *wa-*

-ganga, -gonjwa, -toto, -tu

Klasse 2 (traditionell 3/4): Sg. *m/mw-* Pl. *mi-*

-situ

Klasse 4 (traditionell 7/8): Sg. *ki-* Pl. *vi-*

-jiji, -sa, -su, -tabu

Klasse 5 (traditionell 9/10): unveränderlich, kein Präfix, keine eigene Pluralform

baba, mama, tamasha, hatari, kawaida, supu, wasiwasi

Für diese Einträge braucht also keine Zitierform definiert zu werden. Die Klassenzugehörigkeit muss allerdings angegeben werden.

Aufgabe Die Stammform *–gonjwa* ist eigentlich ein Adjektiv mit der Bedeutung 'krank'.

1. Machen Sie einen neuen Eintrag für *–gonjwa* als Adjektiv (Kategorie: A) und der Bedeutung 'krank'.
2. Fügen Sie im Eintrag für das Substantiv *mgonjwa* 'Kranker' einen Querverweis ein, der auf das Adjektiv verweist: [\lf Vgl = -gonjwa, \lg 'krank']