

Sprachtechnologie für Sotho- Tswana: Die Modellierung von Ton

Sabine Zerbian
Department Linguistik
Universität Potsdam
sabine.zerbian@uni-potsdam.de



Sprachtechnologie (ST)

- technische Entwicklung und sozioökonomische Orientierung zur Informationsgesellschaft
→ Sprache und Text (Träger von „Information“) haben an Wertigkeit zugenommen
- Verarbeitung von elektronisch verfügbarer Information ist Kern der Computerlinguistik
- Verschiedene Sprachtechnologien stehen zur Verfügung
- Ziel: Benutzerfreundlichkeit erhöhen

Wo wird ST benutzt?

- Korrekturprogramme (Spell-checker)
- Computergestützte Lexikographie
- Volltextsuche
- Sprachsynthese
- Spracherkennung
- Sprachlehr- und lernsysteme
- Maschinelle Übersetzung



Google™
Afrika Borwa

[Advanced Search](#)
[Distribution for Printables](#)

Google search

Ke ikutlwa ke le lehlohonolo

Fuputsa: web maqephe hotswa Afrika Borwa

Google.co.za e fihle ka se: [English](#) [Afrikaans](#) [IsiZulu](#) [IsiXhosa](#)

[Hobla Kaha Google](#) - [Google.com.in English](#)

©2010 Google





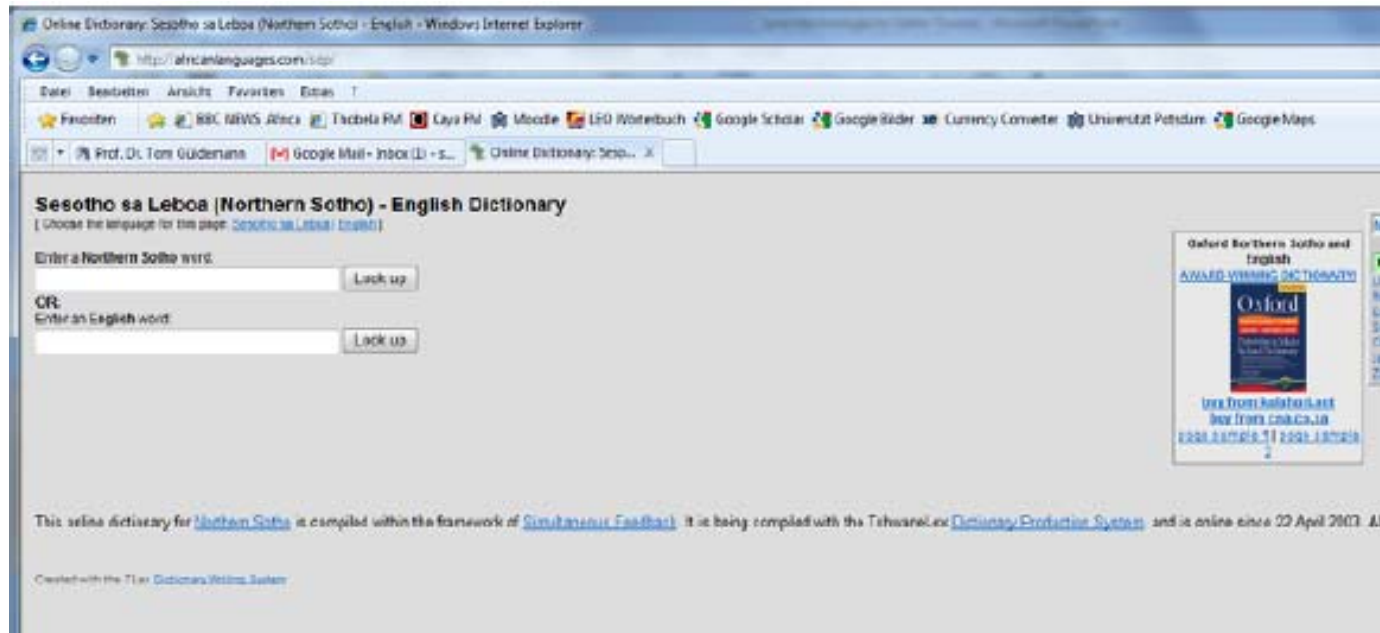
The screenshot shows the Mozilla Add-ons website for the "Northern Sotho spell checker" add-on. The page features the Mozilla logo and the Firefox logo. A search bar is visible with the text "search for add-ons". The add-on is listed as "Northern Sotho spell checker" with ID "20100125" by "Dwayne Bailey, Translate.org.za". The description states: "Your Northern Sotho spell checker allows you to check spelling in Northern Sotho and avoid typing errors. Now you don't need to see your Northern Sotho text underlined in red and you can be proudly South African." A "Jetzt herunterladen" button is present. The add-on is marked as "Dieses Add-on empfohlen". The metadata table shows:

| | |
|------------------|---|
| Aktualisiert | 25.01.2016 |
| Webseite | http://www.translate.org.za/ |
| Funktioniert mit | Firefox 2.0b1 - 36.* |
| Bewertung | ★★★★★ 2 Bewertungen |
| Downloads | 355 |

Below the table, it says "Mehr über dieses Add-on".

<https://addons.mozilla.org/de/firefox/addon/46519/>





<http://africanlanguages.com/sdp/>





<http://ctext.nwu.ac.za/ProductsSoftware.html>



The screenshot displays the website for Airports Company South Africa. At the top right is the logo for AIRPORTS COMPANY SOUTH AFRICA. Below the logo is a navigation bar with 'LATEST NEWS' and a link to 'Basic Prayers Service at O.R. Tambo International Airport'. A search bar is located below the navigation bar with the text 'Flight information: +27 (0)86 727 7888' and a 'Search' button. The main content area is divided into two sections. On the left, under the heading 'Passengers', there is a sub-heading 'For detailed information on each of our airports, click on the relevant link below:' followed by a grid of 12 airport images with labels: O.R. Tambo International, Cape Town International, Durban International, Bloemfontein International, Phakisa International, Port Elizabeth International, Upington International, East London Airport, George Airport, Kimberley Airport, and King Shaka International. On the right, there is a 'Live flight info' section with radio buttons for 'Arrival' (selected) and 'Departure', and checkboxes for 'Domestic' and 'International'. Below this are input fields for 'From', 'To', and 'Airline', each with a '...list' dropdown arrow. There is also a 'Date' dropdown menu set to '21 June 2019' and an 'Arrival time' section with 'from' and 'to' time pickers (00:00) and a '...search' button. An 'OR' section follows with a 'flight number' input field and a '...search' button. At the bottom of the right sidebar, there are two links: '+ Receive flight information via SMS ...more' and '+ Season flight schedule ...more'. A vertical sidebar on the left of the flight info section contains links for 'Live flight info', 'Travel links', and 'Parking calculator'.

ACSA: +27 (0)86 727 7888



Wer engagiert sich für weniger erforschte Sprachen?

- Initiativen wie
 - AfLaT (African Language Technology)
<http://aflat.org/node/1>, jährl. Konferenz, seit 2009
 - *Spoken Language Technologies for Underresourced Languages*; biannualer Workshop, seit 2008
- Uni-Projekte: z.B. SFB 632, elektronische Korpora für Gur-, Kwa- und tschadische Sprachen

Herausforderungen in afrikanischen Sprachen

- Multilingualität (SA: 11 offizielle Sprachen)
- unterschiedliche Sprachfamilien
- qualifizierte Studenten, Post-docs, Forscher
- Anzahl der Sprecher/ Nutzer
- Kürzere Vergangenheit in der Forschung
- Weniger Geld
- Computeraffinität der Zielgruppe

Sprachtechnologie in Südafrika

- President Mbeki's 2002 State of the Nation Address mandated a national strategic initiative with the major objective to facilitate national economic and social development through human capital development and needs-based research and innovation, leading to products and services based on Information and Communication Technology

(Meraka Institute, CSIR, website)

Gesellschaftlicher Kontext für ST

- Zugang zu verlässlicher und aktueller Information keine Selbstverständlichkeit
- Computer-Infrastruktur wenig entwickelt
- Verbreitung von Mobiltelefonen bietet Möglichkeit für sprachbasierten Informationszugang
- Sprachbasierter Informationszugang auch für Analphabeten und nicht-computer-affine Nutzer

Technische Herausforderungen (ST)

- Kodifizierung des linguistischen Wissen erfordert z.T. Grundlagenforschung
- Erstellung und Entwicklung von grundlegenden Ressourcen, wie z.B. Wortlisten, Phonemsets, Aussprachewörterbücher, Korpora
- Entwicklung sprecher-unabhängiger ASR-Systeme/ TTS-Systeme für vers. Sprachen
- Design von benutzbaren Interfaces
- Auswahl an geeigneten Anwendungsgebieten

Wer engagiert sich in Südafrika?

- Council for Scientific and Industrial Research (CSIR); <http://www.meraka.org.za/humanLanguage.htm>
- Ctext @ North-West University http://ctext.nwu.ac.za/Index_e.html
- University of Pretoria
- UNISA
- University of Limpopo



Resourcen 2006-2009

The image shows two screenshots of the LWAZI website. The left screenshot is the main homepage, and the right screenshot is the 'Downloads' page.

Homepage Content:

- Header:** LWAZI logo, Department of Arts and Culture, Republic of South Africa.
- Navigation:** Home, About Lwazi, Areas of research, Get involved!, Demos, Downloads, Publications, Contact us.
- Main Text:**

A telephone-based, speech-driven information system

More and more South Africans have access to telephones, while most do not have reliable access to the internet or even printed media. A telephone-based, speech-driven information system can build on existing infrastructure and communication methods to connect all South Africans to the benefits of information technology.

The Lwazi project was commissioned by the Department of Arts and Culture to provide South Africans with access to government information and services in any of the 11 official languages, using a landline or mobile telephone, free of charge. The Lwazi service was developed by the Human Language Technology (HLT) Research Group of the Meraka Institute of the CSIR.
- Areas of research:**

The Lwazi telephone-based, speech-driven information system will allow every South African to access government information and services for all South Africans and will showcase the potential of human language technologies (HLTs) in South Africa.

Three areas of research:

 - Application selection and Human Factors
 - Scientific and Technical Outputs Development
 - Electronic Linguistic Resource Collection

Downloads Page Content:

- Section:** Text-to-Speech (TTS) Resources
- Table:**

| Audio and alignments | Voices |
|--|--|
| <ul style="list-style-type: none"> Afrikaans: ZIP file, 51M English: ZIP file, 39M isiNdebele: ZIP file, 67M isiXhosa: ZIP file, 62M isiZulu: ZIP file, 61M Sepedi: ZIP file, 37M Sotho: ZIP file, 22M Swswana: ZIP file, 35M Swati: ZIP file, 63M Tshivenda: ZIP file, 48M Xitsonga: ZIP file, 38M | <ul style="list-style-type: none"> Afrikaans: ZIP file, 21M English: ZIP file, 69M isiNdebele: ZIP file, 110M isiXhosa: ZIP file, 88M isiZulu: ZIP file, 90M Sepedi: ZIP file, 58M Sotho: ZIP file, 34M Swswana: ZIP file, 50M Swati: ZIP file, 104M Tshivenda: ZIP file, 80M Xitsonga: ZIP file, 61M |

Footer: All information copyright © CSIR. Terms & Conditions | Tel: +27 (0)12 041 2352 | Email: lwazi@meraka.csr.co.za. An initiative funded by the Department of Arts and Culture.

<http://www.meraka.org.za/lwazi/>

Beitrag der Sprachwissenschaft?

- Keynote-Ansprache AfLaT 2010 auf Malta:
Justus Roux: "*Do we need linguistic knowledge for speech technology applications in African languages?*"
- Kodifizierung des linguistischen Wissen erfordert z.T. Grundlagenforschung
- Informationen in der Literatur teilweise erhältlich (z.B. Khumalo 1981/82 für Ton im Zulu, Chebanne et al. 1997 für Ton im Tswana)

Fallstudie:

Ton im Sotho-Tswana TTS-System

- Zusammenarbeit auf dem Gebiet der Tonmodellierung mit Etienne Barnard, Daniel van Niekerk und Mpho Raborife vom Meraka-Institut, CSIR, Pretoria
- Beitrag der Linguistik zur Sprachtechnologie
- Beitrag der Sprachtechnologie zur Linguistik

Fallstudie: Bantu Ton

- Jede Silbe für einen Ton (H oder L) spezifiziert

bóna ‚sehen‘ / **boná** -‘sie‘

- Ton hat lexikalische Bedeutung

bá-bóna - ‚sie sehen‘ vs.

bá-boná - ‚indem sie sehen‘

- Ton hat grammatische Bedeutung

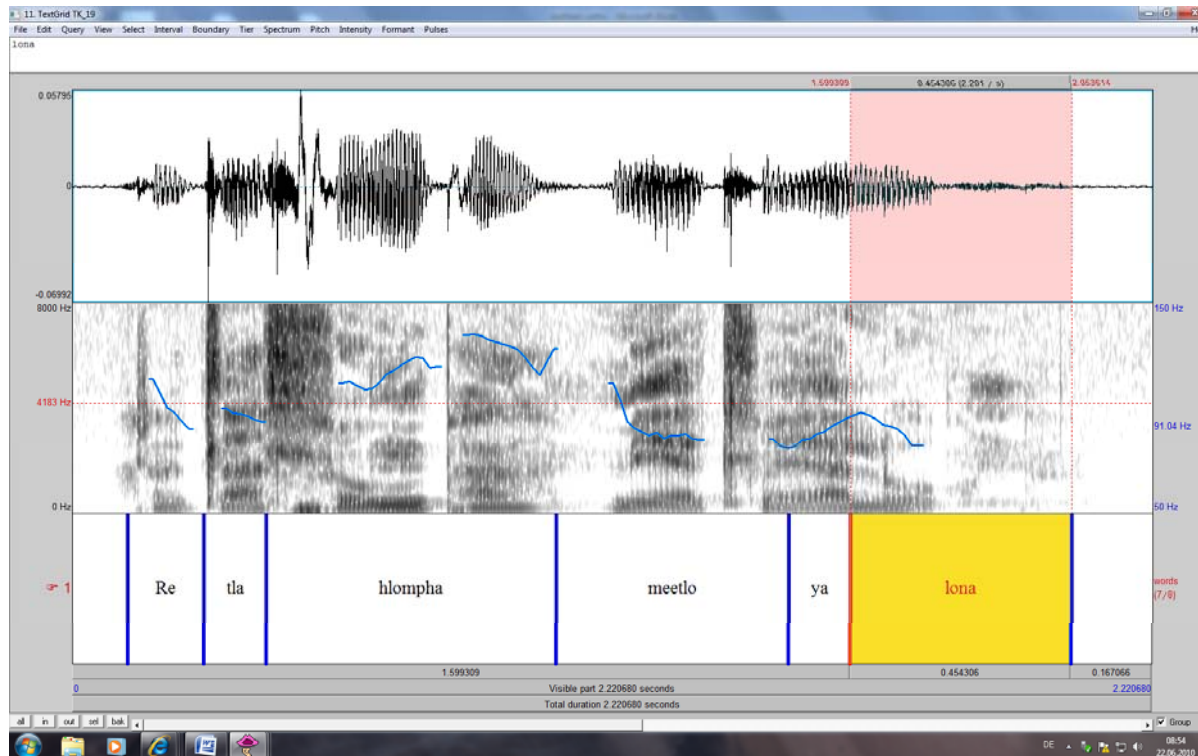
Tonassimilation: mo-sádi ó-bóná...

Meeussen's Rule: mo-sádi wámé

Extratonalität: Ò-bóná mo-sádi.

- Ton ändert sich im Satzkontext

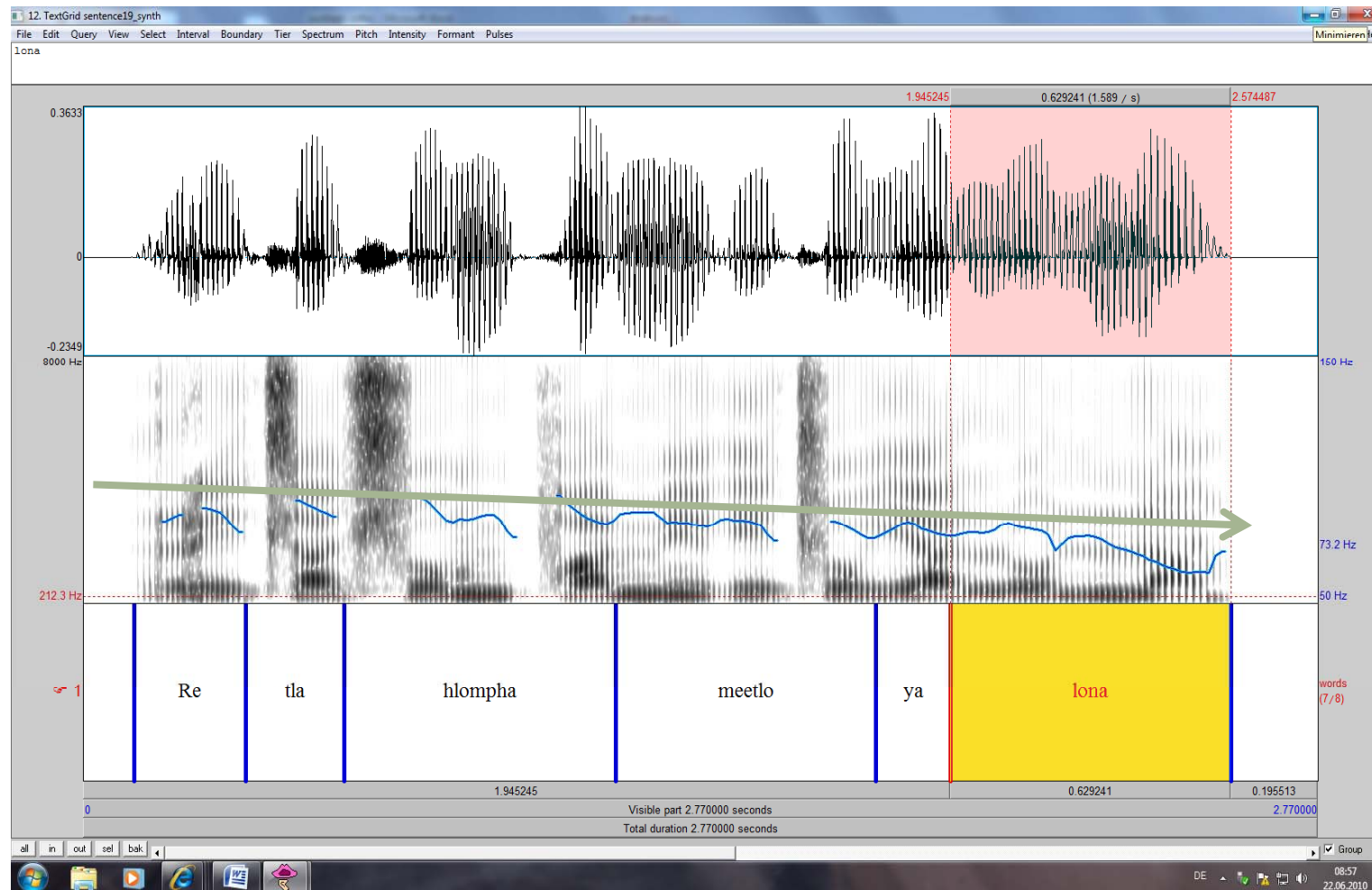
- Ton d/ Lexikon, Morphologie und Syntax bestimmt
- Weitere intonatorische Merkmale: Deklination, Längung der vorletzten Silbe



Re tla hlômpha me-êtlo ya lona
 1PL FUT respect CL4.tradition AGR PRN.2PL
 “We will respect your traditions”

Herausforderung in TTS

- Penultimate lengthening & Deklination weitestgehend integriert
- Kurze Historie der Generierung von Ton für SA Bantu TTS:
 - Versuche, Oberflächenton direkt aus dem akustischen Signal abzuleiten (Kuun et al. 2005)
 - Statistische Vorhersage von Ton (Roux, p.c.)
 - Hidden Markov Model-basiertes Sprachsynthese System (van Niekerk & Barnard, Ms.)



Re tla hlômpha me-êtlo ya lona
 1PL FUT respect CL4.tradition AGR PRN.2PL
 “We will respect your traditions”



Linguistischer Ansatz: Ton in TTS

Input:

Ba tloga ba thobile kgobe ka mootlwa.

„Sie gehen, nachdem sie den Mais mit einem Dorn aufgespießt haben.“

1. Modul: Shallow Parser (POS-Tagger und morphologische Analyse)

- **ba** independent subject concord class 2
 - **tloga** verb Present Tense, main clause
 - **ba** dependent subject concord class 2
 - **thob-ile** verb Perfect Tense, dependent clause
 - **kgobe** noun class 9
 - **ka** preposition
 - **mo-otlwa** noun class 3
- Ton bestimmt mit, welche Informationen enkodiert werden

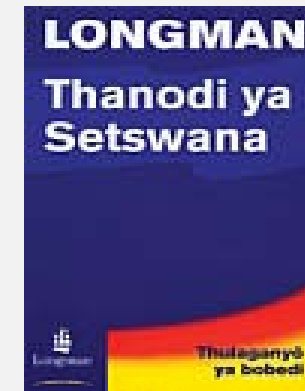
2. Modul: Tonmarkiertes Lexikon

- **Lexikalischer Hochton**

Bá tloga ba thobile kgobe ká moótlwa.

- **Morphologischer Hochton**

Bá tloga bá thobíle kgobe ká moótlwa.



→ Argumente von Lexikographen (!) gegen Markierung von Ton im Lexikon:

- Probleme bei der Darstellung mehrerer Diakritika
- Platzgründe bei gedruckten Exemplaren
- fehlende Relevanz von Ton

3. Modul: Tonregelalgorithmus

- **Iterative High Tone Spread**

Bá tloga bá thobílé kgobe ká moótlwa.

- **High Tone Spread**

Bá tlóga bá thobílé kgobe ká moótlwá.

- **Finality Rule**

Bá tlóga bá thobílé kgobe ká moótlwa.

→ Tonregeln in der linguistischen Literatur beschrieben und relativ leicht programmierbar (cf. Raborife 2009, in progress)

4. Modul: Tonlabels

Bá tlóga bá thobílé kgobe ká moótlwa

H H L H L HH L L H LH L

- Mini-Pilotstudie (15 Sätze, 3 Transkribierer, siehe Raborife 2009):

vorhergesagter Ton vs transkribierter Ton

→ 94.8% Übereinstimmung

Beitrag der
Linguistik zur
Sprachtechnologie

5. Modul:Phonetische Implementierung

- Akustisches Korrelat von Ton ist Grundfrequenz (F_0)
- Ton ist relativ
- H = höhere durchschnittliche Grundfrequenz in der Silbe

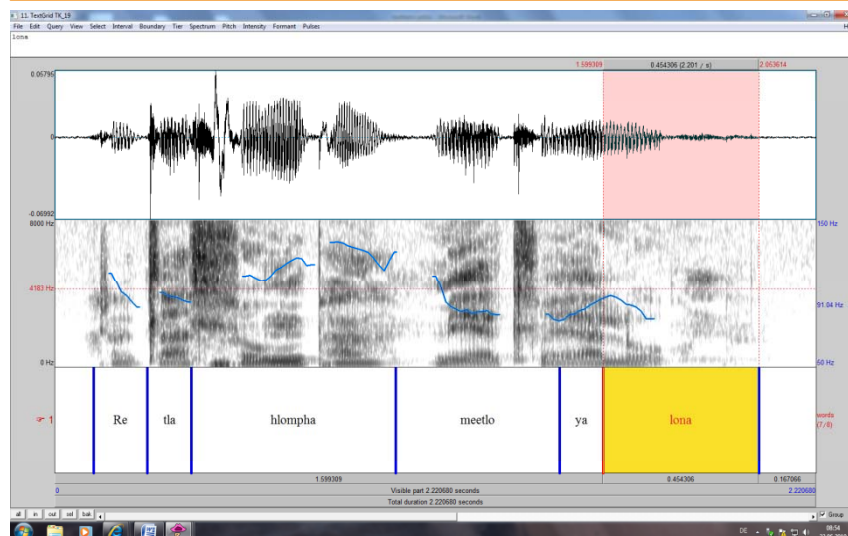
Phonetische Analyse (Barnard & Zerbian 2010):

- Veränderung in durchschnittlicher Grundfrequenz zwischen Silben ist stärkerer Indikator
- Intrasilbischer Trend von F_0 ist entscheidend

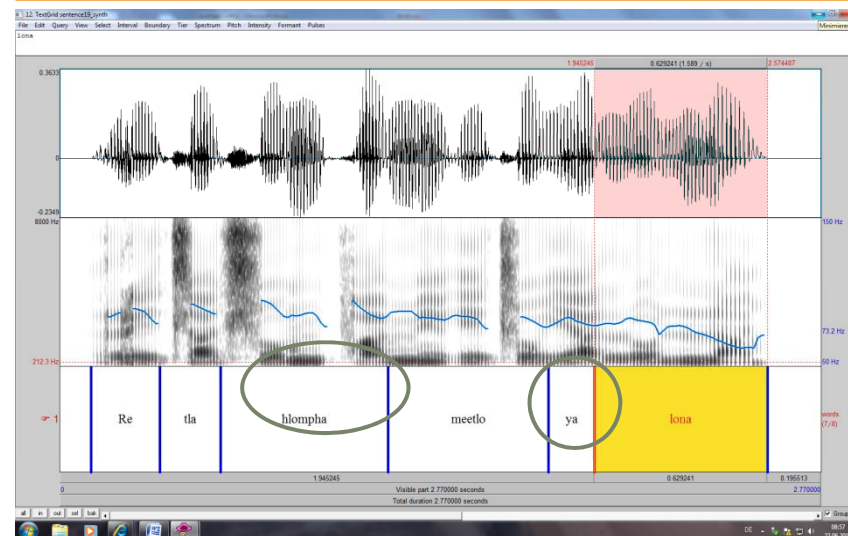
Beitrag der
Sprachtechnologie
zur Linguistik

Vergleich I

Natürliche Sprache



Synthetisierte Sprache

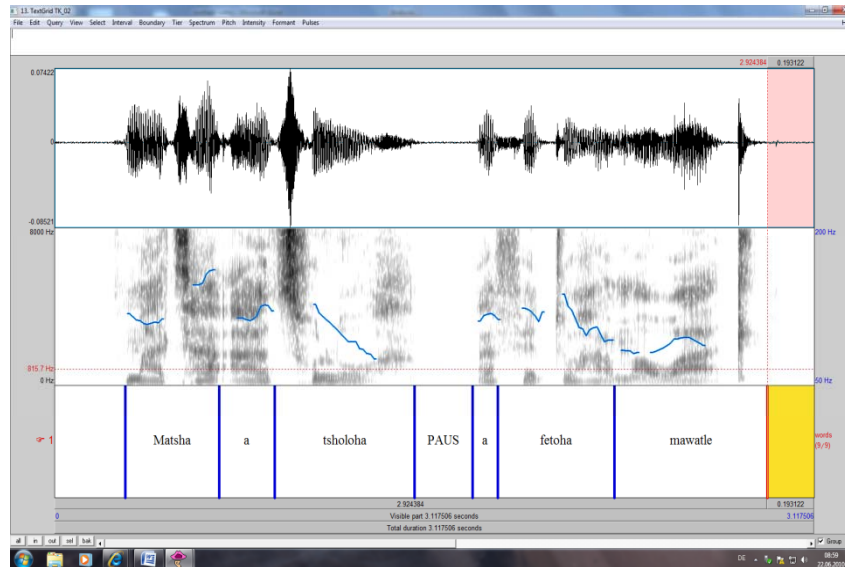


Vorhersage

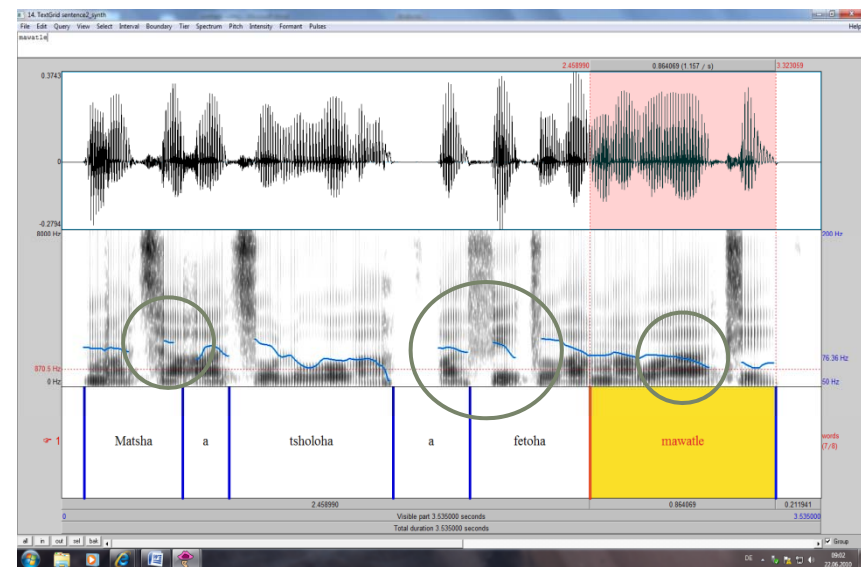
Re tla hl_ompha me-êtlo yá loná
 1PL FUT respect CL4.tradition AGR PRN.2PL
 “We will respect your traditions”

Vergleich II

Natürliche Sprache



Synthetisierte Sprache



Vorhersage

Ma-tshá á tshóloha, á fêtoha ma-wátlê
 CL6-lake SC6 pour SC6 change CL6-sea
 “The lakes are overpouring, they are turning into seas”

Weitere Anwendung der Module

- Sprachbeschreibung, Module 1-4 in eventuell anderer Reihenfolge
→ Computermodule testen linguistische Modelle
- Sprachunterricht (siehe Mascher, Ms.; gewürdigt mit dem PanSALB Multilingualism Award Setswana in Jo'burg, Feb. 2010)



Literatur

- Barnard, E. & Zerbian, S. 2010. From Tone to Pitch in Sepedi. *Proceedings of the Workshop on Spoken Language Technologies for under-ressourced languages* (SLTU 10).
- Kuun, C., V. Zimu, E. Barnard and M. Davel, "[Statistical investigations into isiZulu intonation](#)", In *Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 111-115, Langebaan, South Africa, November 2005.
- Mascher, D. A five-year course in Tswana. Manuscript, Vryburg, Südafrika.
- Raborife, M.I. 2009. The Implementation of Sesotho Tonal Rules in a Text-To-Speech system. Honours report, School of Computer Science, University of the Witwatersrand, Johannesburg.
- Raborife, M.I. in progress. The Implementation of further Sesotho Tonal Rules. Master's report, School of Computer Science, University of the Witwatersrand, Johannesburg.
- Van Niekerk, D.R. & Barnard, E. 2009. Phonetic alignment for speech synthesis in under-ressourced languages. In *Proceedings of Interspeech 2009*. Brighton, UK.
- Van Niekerk, D.R. & Barnard, E. An intonation model for TTS in Sepedi. Manuscript, CSIR, Pretoria.
- Zerbian, S. & Barnard, E. 2008. Phonetics of intonation in South African Bantu languages. *SALALS* 26(2): 235-254.