

ProcedureProcedure

**Title of DocumentBriefing: Lemmatisation**

**Name of AuthorW. Pienaar**

**2020/09/10**

TABLE OF CONTENTS

[1. Introduction 3](#_Toc35005364)

[2. Initial setup 3](#_Toc35005365)

[TeamViewer 3](#_Toc35005366)

[Dropbox 4](#_Toc35005367)

[3. Software used for annotation 4](#_Toc35005368)

[Lara2 4](#_Toc35005369)

[Steps during the annotation process: 4](#_Toc35005370)

[4. Protocols for annotation 5](#_Toc35005371)

# Introduction

During the NCHLT: Text I project (2010-2013), various resources for South African Languages were developed and are available for download from the language Resource Management Agency (http://rma.nwu.ac.za). Among these resources were annotated, monolingual, stratified corpora for ten South African languages, with a size of fifty thousand tokens each, annotated on token, orthographic, morphological and morphosyntactic layers. This project also developed protocols and tag sets for each level and language in line with internationally accepted universal tag sets. Although the annotated corpora are valuable resources for linguistic and computational linguistic investigations, there is significant scope for the improvement and further use of these annotated data sets. This is especially true for the conjunctively written South African languages, due to the fact that the type-to-token ratio for these languages is significantly higher because of their morphological complexity.

In this project which will be developed for the South African Centre for Digital Language Resources (SADiLaR), we will extend the existing annotated data with an additional 50,000 tokens (calculated on the English source text) for four conjunctive languages, viz. isiNdebele, isiXhosa, isiZulu and siSwati. As with the NCHLT: Text I project, CTexT will collect parallel corpora with English as the source language. By utilising a parallel dataset, research on differential and contrastive linguistics can also be conducted, and possible technology transfer between languages can also be explored. The annotation of the corpora will be conducted on three levels of annotation, namely:

1. morphological analysis
2. part of speech tagging
3. lemmatisation

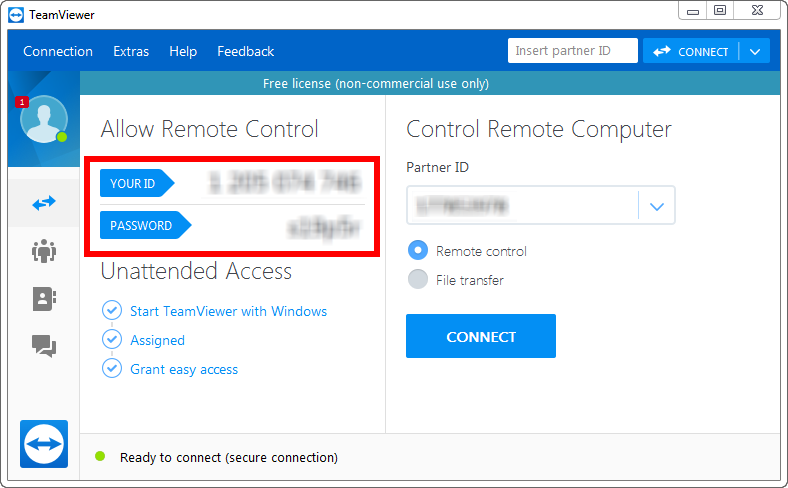
In level 1 the morphological analyses were done and in level 2 the part of speech. We are currently busy with level 3 (lemmatisation). The pre-tagged lemmatisation data are based on the morphological analysis done during level 1.

# Initial setup

Before the first use, some initial setup steps need to be performed by CTexT.

## TeamViewer

1. Download TeamViewer from <https://www.teamviewer.com/en/download/windows/>.
2. Install TeamViewer on your computer.
3. Open TeamViewer on your computer.
4. When someone from CTexT makes contact to commence with the setup, provide the person with your TeamViewer ID and Password.



1. Your computer can now be remotely controlled by the person responsible for doing the setup.

## Dropbox

All the following steps will be done remotely via TeamViewer:

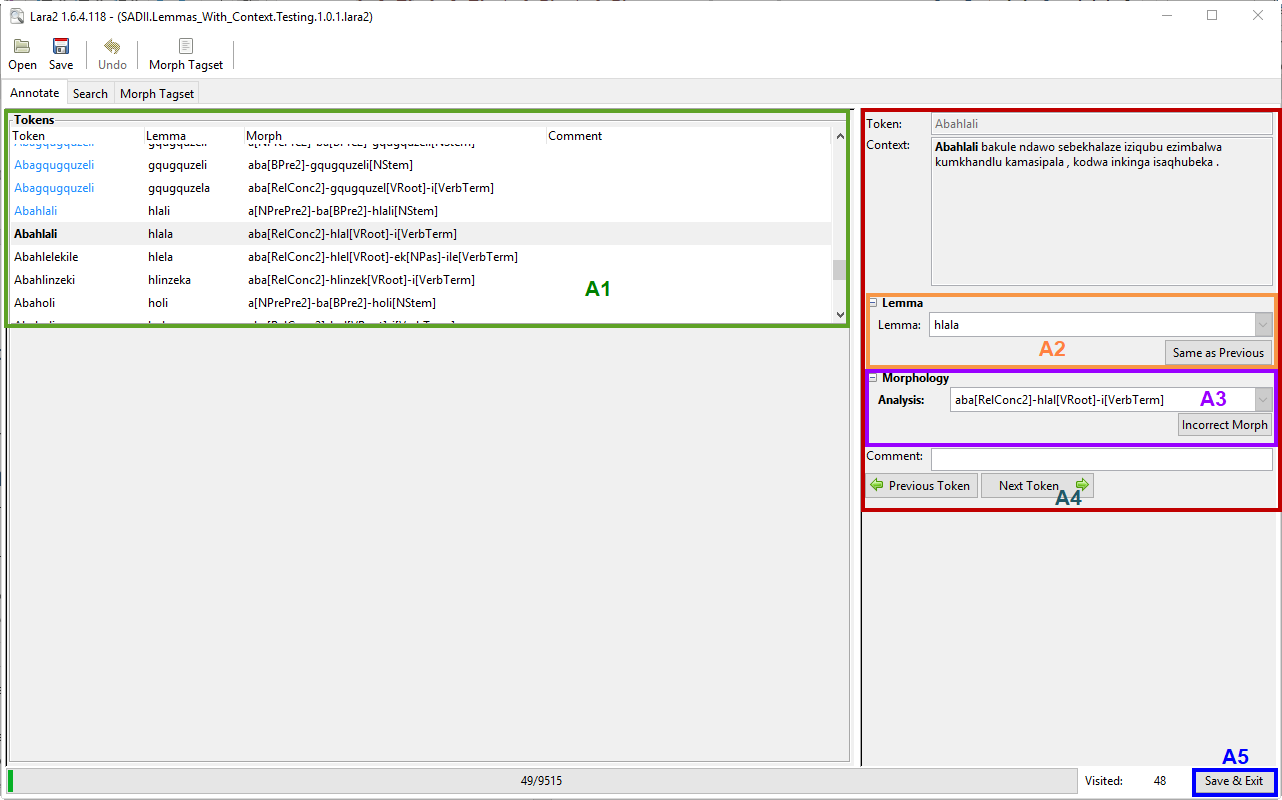
1. Install Dropbox.
2. Create a local folder on your computer. This folder will be the working folder.

Please ensure that you always work in the created Dropbox folder. In this way we have access to the files, and ensure that it is backed up safely.

# Software used for annotation

## Lara2

In this project we make use of the Lara2 software environment. Please refer to the Lara2 help file for more in-depth information about the program. Ensure that the version installed is 1.6.4.118 or later. Take note that the interface will look a little different than the interface for the morphological and part of speech analysis.



## Steps during the annotation process:

* Check for each generated lemma if it is the correct one for the token and its morphological analysis. The example sentence will provide some context.
* A lemma can either be changed from the Lemma column in the pane on the left hand side (A1) or in the text box on the right hand side (A2).
* If you notice an error in the morphological analysis, click the *Incorrect morph* button (A3) to add “Incorrect morph” to the comment field as well as highlight the token in red. Alternatively the analysis can either be corrected in the Morph column in the pane on the left hand side (A1) or in the text box on the right hand side (A3).
* To proceed to the next token either click the *Next Token* button (A4), use the enter button on the keyboard or the down arrow on the keyboard.
* When you are finished, click the *Save & Exit* button (A5) in the bottom right corner to save the file and exit the program.

# Protocols for annotation

Refer to the annotation protocol for the specific language you are working on.