

CTAP user guide

Author: Nadezda Okinina

<https://kommul.eurac.edu/ctapWebApp>

Contents

1. Introduction
2. How to use the online browser-based CTAP tool
 - 2.1. Register
 - 2.2. Upload a corpus
 - 2.3. Select features for analysis
 - 2.4. Set up the analysis
 - 2.5. Run the analysis and download the results
 - 2.6. Visualise the results
 - 2.7. Dashboard
3. Description of linguistic complexity features implemented in CTAP
 - 3.1. Lexical complexity features
 - 3.2. Morpho-syntactic complexity features
 - 3.3. Syntactic complexity features

1. Introduction

CTAP is a tool that allows to measure linguistic complexity (or readability) of English, German, and Italian texts.

It was developed for English by Xiaobin Chen at the Tübingen University Computational Linguistics department¹ and then extended for German by Zarah Weiss, also in Tübingen; after that it was extended for Italian by Nadezda Okinina at the Institute for Applied Linguistics in Eurac Research².

CTAP was conceived to be extendable to any number of languages. For the obvious reason of differences in language systems and in available resources, not every feature can be calculated for every language. Further extension for the Italian language is foreseen in the near future, your collaboration is very welcome. If you need a feature that is not yet implemented into CTAP, write to applnglt@scientificnet.onmicrosoft.com. Or if you come across a resource (a list of words, for example) that could potentially be used for the calculation of a complexity feature for Italian, write to applnglt@scientificnet.onmicrosoft.com. If you learn about the existence of a gold standard or you want to propose a method for evaluation, please, let us know.

CTAP is open source and distributed under the BSD licence.

The code can be found on Github: <https://github.com/commul/ctap>

1 <https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/departments-of-linguistics/>

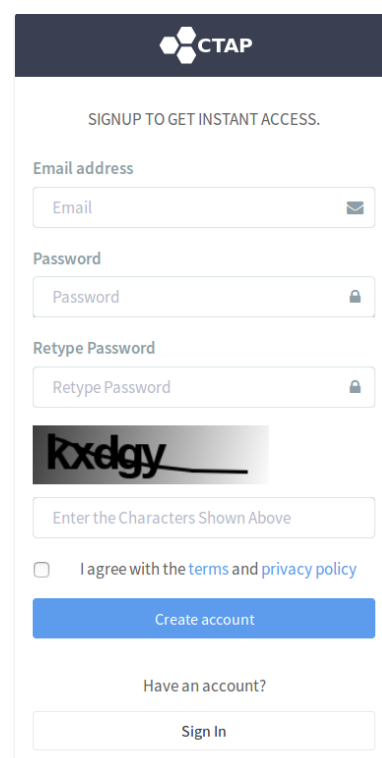
2 <http://www.eurac.edu/en/research/autonomies/commul/Pages/default.aspx>

2. How to use the online browser-based CTAP tool

CTAP is a browser-based tool and can be used via <https://kommul.eurac.edu/ctapWebApp>.

2.1. Register

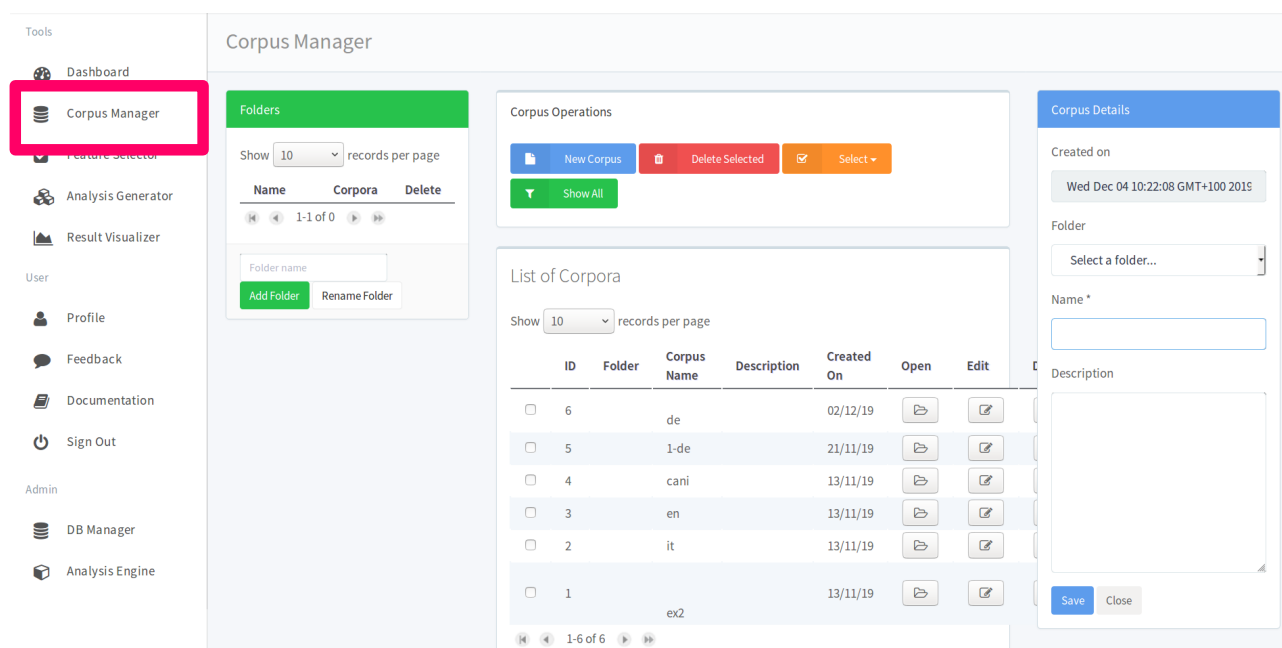
If you are not registered as a user yet, go to the URL <https://kommul.eurac.edu/ctapWebApp> and create an account.



The image shows the CTAP sign-up form. It has a dark header with the CTAP logo. Below the header, it says "SIGNUP TO GET INSTANT ACCESS." The form includes fields for "Email address", "Password", and "Retype Password". There is a CAPTCHA section with a logo and a text input field. Below the CAPTCHA, there is a checkbox for "I agree with the terms and privacy policy" and a blue "Create account" button. At the bottom, there is a "Sign In" button for users who already have an account.

2.2. Upload a corpus

Once signed in, you will see a menu on the left. The second item from the top is the *Corpus Manager*. Go to this section and create a new corpus by clicking on the button *New Corpus*. You will have to give a name to the corpus you are creating. The description is optional. The corpus will be added to the *List of Corpora*. Now you can push the *Open* button and add files to the corpus. Click on the *Import* button and then on the *Add files* button. This will allow you to choose files from your file system. If you have few files to analyse, you can also drag and drop them in the *Drop files here* field.



The image is a screenshot of the CTAP Corpus Manager interface. On the left is a sidebar menu with items: Dashboard, Corpus Manager (highlighted with a red box), Feature Selector, Analysis Generator, Result Visualizer, User, Profile, Feedback, Documentation, Sign Out, Admin, DB Manager, and Analysis Engine. The main area is titled "Corpus Manager" and contains three panels. The "Folders" panel on the left shows a table with columns "Name", "Corpora", and "Delete". It has a "Show 10 records per page" dropdown and a "1-1 of 0" indicator. Below the table are "Add Folder" and "Rename Folder" buttons. The "Corpus Operations" panel in the middle has buttons for "New Corpus", "Delete Selected", "Select", and "Show All". The "List of Corpora" panel on the right shows a table with columns: ID, Folder, Corpus Name, Description, Created On, Open, and Edit. It also has a "Show 10 records per page" dropdown and a "1-6 of 6" indicator. The table lists six corpora with IDs 1 through 6. The "Corpus Details" panel on the far right shows a form for creating a new corpus, with fields for "Created on" (Wed Dec 04 10:22:08 GMT+100 2019), "Folder" (Select a folder...), "Name" (Name *), and "Description". It has "Save" and "Close" buttons at the bottom.

Corpus files content

Those files have to be in **plain text** format, containing texts in English, German or Italian that are **longer than 51 characters**. If a text is too short, it won't be analysed by the tool.

Corpus files language

If a text is in **several languages**, the analysis result will be **erroneous**, because most natural language processing components on which CTAP depends are language-specific. In order to run the analysis you will have to choose **only one language**.

The files you have just added will appear in the *List of Texts*. By clicking the *Edit* button you may modify their text content.

2.3. Select features for analysis

Now that you have a corpus to analyse, you need to select the complexity features that you want to extract for your texts. In the menu on the left, click on *Feature Selector*, and then on *New Feature Set*. Give a name to your feature set. The description is optional. The feature set will appear in the *List of Feature Sets*. Click the *Open* button: on the right you will see a list of all the complexity features available in the tool.

The complexity features available in the tool are now a little bit less than 400: around 270 for English, around 264 for German, around 245 for Italian. Some features are available in 3 languages, others in 2, others in 1. The numbers are impressive, but many features are variants of each other or just the same calculations are applied to different language elements (different parts of speech, for example).

By default, you will only see the first 10 features of the list and will have to go to its bottom and click on the right arrow in order to see more and more features. But you can select the number of features that will be displayed at the same time from the drop down list.

However, in order to analyse your corpus, you are only interested in features for 1 of the available languages. Please, go to the *Select* menu and **choose the language** you need from the drop down list. The features corresponding to this language will be highlighted in blue. If you want to know what a specific feature is doing, click on the information button *I* in the *Details* column and read its description. Once you have decided what features to use, click on the + button in the *Add* column: the feature will be added to the *List of Features Selected* on the left side of the screen. By pushing the *Add Selected* button you will add all the selected features to your list.

In order to select all the features available for one language, first select *show all records per page* from the dropdown menu, then select the language from the orange *Select* dropdown menu, then push the blue button *Add selected*.

The screenshot displays the 'Feature Set Manager: guide-feature-set' interface. On the left is a sidebar with navigation links: Tools (Dashboard, Corpus Manager, Feature Selector, Analysis Generator, Result Visualizer), User (Profile, Feedback, Documentation, Sign Out), and Admin (DB Manager, Analysis Engine). The main area is divided into two panels: 'Selected Features' (green header) and 'Available Features' (blue header). The 'Selected Features' panel shows a 'List of Features Selected' table with columns: ID, Feature Name, Languages, Details, and Remove. It includes a search bar and a 'Show 10 records per page' dropdown. The 'Available Features' panel shows a 'List of Available Features' table with the same columns, plus an 'Add' column. It also has a search bar and a 'Show 10 records per page' dropdown. In the 'Available Features' table, the feature 'Cohesive Complexity Feature: Breindl Additive and Concessive Connectives per Token' (ID 392) is highlighted in blue, indicating it is selected for the current language (de). The 'Add' column for this feature shows a '+' button. The 'Selected Features' table is currently empty.

The *Reverse* option in the *Select* dropdown menu allows to deselect all the selected features and select all the others, that have not been selected yet.

2.4. Set up the analysis

We have created the 2 elements that will allow us to construct an analysis: a corpus and a feature set. Now go to the *Analysis Generator* in the left side menu and click on *New Analysis*. A menu will appear on the right, where all the fields will be compulsory to fill, except the description.

Important notice:

Please pay attention to the **language** you are choosing, because the **default language is English**, and if you don't change it to the language you really need, NLP tools for English will be applied to the uploaded texts independently from the text language. **Results for non-English texts will be wrong.**

!!!!!!!!!!!!!!

In the *Select a Corpus* menu you should choose a corpus to analyse and in the *Select a Feature Set* menu you should select a feature set. The new analysis will be added to the *List of Analysis* on the left.

The screenshot shows the 'Analysis Generator' interface. On the right, the 'Analysis Detail' panel is visible. The 'Language' dropdown menu is highlighted with a red box, showing 'English' selected. Below it, the 'Tag Filter' is set to 'No Tag Filter'. The 'Select a Corpus' dropdown shows 'f 1-de' and the 'Select a Feature Set' dropdown shows 'punct test'.

2.5 Run the analysis and download the results

Push on the arrow in order to run the analysis. A progress bar will appear on top. When the whole progress bar becomes blue, it means the analysis is complete and results can be downloaded. On the right side of the screen you can see a drop down menu that allows you to choose the form of the .CSV file that you will download. I invite you to try both the long and the wide formats in order to understand the difference between them and which one you prefer.

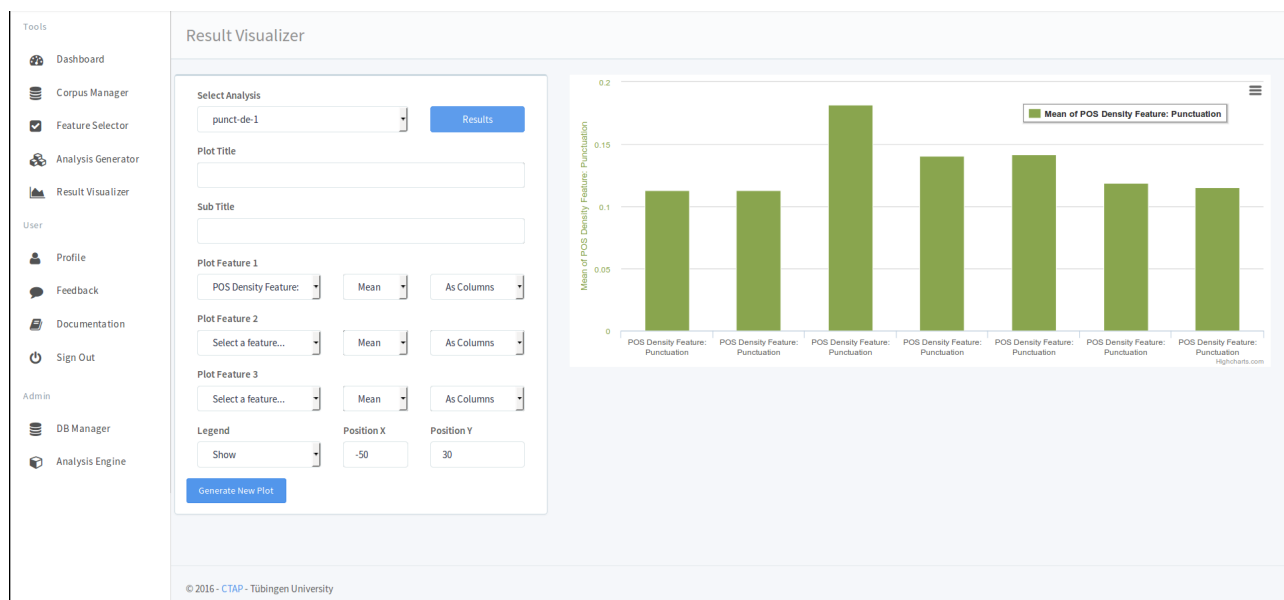
The screenshot shows the 'Analysis Generator' interface. A progress bar at the top indicates 'punct de 1: 100%'. Below it, the 'List of Analysis' table is visible. The 'Download results as' dropdown menu is highlighted with a red box, showing 'Long' selected. The table has columns for ID, Analysis Name, Corpus, Feature Set, Language, Created On, Run, Pause, Stop, Results, Edit, and Delete.

ID	Analysis Name	Corpus	Feature Set	Language	Created On	Run	Pause	Stop	Results	Edit	Delete
2	de-all	1-de	test	DE	02/12/19						
1	punct-de-1	1-de	punct	DE	02/12/19						

2.6. Visualise the results

Apart from downloading the results, you can visualise them as a graph. Go to the *Result Visualizer* menu on the left. Select the analysis you are interested in from the drop down list under the Select Analysis title. Select the features you want to plot, the type of representation etc. Push the [Generate New Plot](#) button. A graphic will be generated on the right. You will be able to download the image in different formats by clicking on the 3 horizontal bars in its upper right corner.

You can also download the analysis results in .CSV wide format by clicking on [Results](#).



2.7. Dashboard

In order to get a general picture of the work you have already done in CTAP, go to the *Dashboard* tab on top of the top left menu. You will see the corpora, the feature sets and the analysis you have created. On the right you will see a *System Provided Feature List* with all the available complexity features and their explanations.

The screenshot shows the 'Dashboard' interface. At the top are three summary cards: 'CORPORA' (2), 'FEATURE SETS' (2), and 'ANALYSES' (2). Below are three lists: 'Corpus List', 'Feature Set List', and 'Analysis List'. On the right is a 'System Provided Feature List' with details for various features.

ID	Corpus Name	Description	Created On
2	f		03/12/19
1	1-de		02/12/19

ID	Feature Set Name	Description	Created On
2	punct		02/12/19
1	test		25/11/19

ID	Analysis Name	Description	Corpus	Feature Set	Language	Created On
2	de-all		1-de	test	DE	02/12/19
1	punct-de-1		1-de	punct	DE	02/12/19

ID	CF Name	Description
401	Lexical Sophistication Feature: SUBTLEX Word Familiarity Per Million Words (LW Token)	Calculates lexical sophistication of the text. Three sophistication measures are calculated from ...
400	Lexical Richness: Type Token Ratio (Uber)	Calculates the type token ratio of a text. A word type is a non-duplicated token. This feat...
399	Syntactic Complexity Feature: Complex Nominals per Clause	Calculates the syntactic complexity of the text. This feature calculates the complex nominals ...
398	Number of Syntactic Constituents: Clausal Passive Subject	Calculates the number of a specific syntactic constituents in the text. This feature counts the n...
397	SD Token Length in Syllables	Calculates the standard deviation of token length in number of syllables. AAE Dependency: ...
396	Lexical Sophistication Feature: Google Books Word Frequency (LW Token)	Calculates lexical sophistication of the text. Three sophistication measures are calculated from ...
		Calculates negation adverb

Important notice: As CTAP is still under development, it may be reinitialised from time to time. Therefore it is strongly recommended to **download all the analysis results** that are important to you as soon as you have generated them.

If CTAP is reinitialised, you will also have to create an account over again.

3. Explanation of linguistic complexity features implemented in CTAP

Complexity features implemented in CTAP can be divided in 3 groups:

- lexical features
- morpho-syntactic
- syntactic features

Lexical features can also be divided in several groups:

- Lexical sophistication
- Lexical density
- Lexical variation
- Number and percentage of tokens and word types with two or more syllables
- Mean token length and its standard deviation in letters and syllables

Lexical sophistication: the proportion of relatively unusual or advanced words in a learner's text.

Lexical density: the ratio of lexical words (as opposed to grammatical words) to the total number of words in a text.

Spoken texts have a lower lexical density than written texts.

Lexical variation: the diversity of the vocabulary used in the text.

Many features of lexical complexity can be calculated for **all words** (AW), only for **lexical words** (LW) or only for **function words** (FW).

This document does not explain all the features implemented in CTAP, but gives a general understanding of some of them.

3.1. Lexical complexity features

- Number of word types (lexical variation)

number of different words used in the text.

Example: A cat is only a cat. - 4 word types: a, cat, is, only

- Number of word types with more than 2 syllables (lexical variation)

- Lexical Richness: Type Token Ratio (TTR) (lexical variation)

Type is a word and token is its occurrence in the text.

The more types there are in comparison to the number of tokens, the more varied is the vocabulary, i.e. there is greater lexical variety.

$$TTR = \text{number of types} / \text{number of tokens}$$

- Lexical Richness: Type Token Ratio (Root TTR) (lexical variation)

Type is a word and token is its occurrence in the text.

The more types there are in comparison to the number of tokens, the more varied is the vocabulary, i.e. there is greater lexical variety.

Root TTR = number of types / square root of number of tokens OR square root of (number of types / number of tokens)

- Lexical Richness: Type Token Ratio (Log TTR) (lexical variation)

Type is a word and token is its occurrence in the text.

The more types there are in comparison to the number of tokens, the more varied is the vocabulary, i.e. there is greater lexical variety.

Logarithmic TTR = $\log(\text{number of types}) / \log(\text{number of tokens})$

- Lexical Richness: Type Token Ratio (Corrected TTR)

Type is a word and token is its occurrence in the text.

The more types there are in comparison to the number of tokens, the more varied is the vocabulary, i.e. there is greater lexical variety.

Corrected TTR = number of types / square root of the double of the number of tokens OR square root of (number of types / the double of the number of tokens)

- Lexical Richness: Type Token Ratio (Uber)

square root of the logarithm of the number of tokens / logarithm of (number of types / number of tokens)

- Lexical Sophistication feature: Age of Acquisition

Based on a list of words: words learned during a certain age (0-2, ..., 13 and older) Need to know exactly what list is used in CTAP.

- Mean Token Length in Letters

- Mean Token Length in Syllables

- SD Token Length in Syllables

Standard Deviation from the mean of the token length shows how diversified the words' lengths are. The higher the SD token length, the more diversified the token lengths are, going from very short to very long words.

- Percentage of Tokens with More Than 2 Syllables

3.2. Morpho-syntactic complexity features

- POS Density Features

- singular noun: *number of singular nouns/total number of words in a text*
- singular proper noun: *number of singular proper nouns/total number of words in a text*
- plural noun: *number of plural nouns/total number of words in a text*
- plural proper noun
- wh pronoun : (was, wer ?)
- possessive wh pronoun
- wh adverb

- wh determiner
- preposition
- particle
- functional words
- conjunction
- coordinating conjunction
- pronoun
- personal pronoun
- possessive pronoun
- superlative adjective
- foreign word
- adverb
- superlative adverb
- adverb RB
- adjective
- verb
- past participle verb
- base form verb
- past tense verb
- non 3th person singular verb
- determiner
- determiner DT
- adjective jj
- interjection
- predeterminer
- To
- gerund verb
- existential there
- symbol
- cardinal number
- possessive ending
- lexical words
- modal
- ...

- Imageability (now only available for German)

Imageability scores are based on judgments of how easy it is to create an image of a word. A word such as beach is highly imageable, whereas a word such as philology is not very imageable.

Paivio et al. (1968):

"Words differ in their capacity to arouse mental images of things or events. Some words arouse a sensory experience such as a mental picture or sound very quickly and easily, whereas others may do so with difficulty after a long delay or not at all."

- Contextual Diversity

Contextual Diversity is the number of contexts in which a word has been seen.

- Logarithmic Contextual Diversity

- Lexical Variation Feature:

The lexical variation features show how diversified is the use of each part of speech category in a text.

For example, if instead of naming things by their names a language learner names them all "thing", the noun variation feature of his text will be low.

Here are some examples of available lexical variation features:

- Noun : number of noun types / total number of lexical (non-grammatical) tokens
- Adverb: number of adverb types / total number of lexical (non-grammatical) tokens
- Adjective: number of adjective types / total number of lexical (non-grammatical) tokens
- Modifier: number of adjectives and adverbs / total number of lexical (non-grammatical)

tokens

- Lexical: number of lexical word types / total number of lexical (non-grammatical) tokens
- Verb: number of verb types / total number of lexical (non-grammatical) tokens
- Verb variation 1 : number of verb types / number of verb tokens
- Verb variation 2 : number of verb types / total number of lexical (non-grammatical) tokens
- corrected verb variation 1 : number of verb types / square root of the double of the number

of verb tokens

- squared verb variation 1 : square of the number of verb types / total number of lexical (non-grammatical) tokens

3.3. Syntactic complexity features

- Number of Sentences

Calculates the number of sentences in a text.

- Mean Sentence Length

- in tokens
- in letters
- in syllables

- SD Sentence Length

- in tokens
- in letters
- in syllables

SD means 'Standard Deviation': standard deviation from the mean length of a sentence in an essay.

Example:

There are 3 sentences in an essay. They consist of 2 words, 5 words and 3 words. The mean sentence length is $3.(3) : (2+5+3)/3$ where $(2+5+3)$ is the sum of the lengths of all the sentences and 3 is the number of sentences.

There is standard deviation from this mean: how different the lengths are. If the lengths are 3, 4 and 3, the mean is the same, but the standard deviation is lower, because the values are more similar to each other.

"If the SD Sentence Length is low" means that the essay contains lots of sentences with similar length. To solve this issue, one needs to mix a variety of sentences : simple sentences, complex sentences and compound sentences.

"Low Sentence Length SD" would make the essays boring to read.

- Number of Syntactic Constituents

- Coordinate Phrases
- Fragment T-units
- Fragment Clauses
- T-units
- Verb Phrase
- Sentences
- Clauses
- Dependent Clauses
- Complex Nominal
- Complex T-units