CTAP user guide

Contents of the user guide:

Introduction

CTAP is a tool that allows to measure linguistic complexity of English, German, and Italian texts. It was developed in the Tübingen University Computational Linguistics department[1] for the English language and then extended for German, also in Tübingen, and for Italian, at the Institute for Applied Linguistics in Eurac Research[2]. The tool is still under development, it contains a number of imperfections, but can already be used with caution.

"Used with caution" means that before trusting a number obtained for a certain complexity measure of a certain text, it is preferable to make a test run on a text for which you know the right value. If the 2 values match or are close enough, it means that you can proceed your analysis with this complexity measure. But it will be necessary to repeat the procedure for the next measure. Why is this necessary? Unfortunately, no gold standard exists for the complexity measures and therefore no tests were integrated into the tool. Your help is needed in order to confirm, which measures work well and which measures need to be revised. Please, whenever you have tested a certain complexity measure, write to nadezda.okinina@eurac.edu. It would be of great help for her. Of course, certain tests have already been run, but they are not exhaustive.

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 measure can be calculated for every language. Furthermore, CTAP being still under development, not all possible measures have yet been integrated into it. Further extension for the Italian language is foreseen in the near future, your collaboration is very welcome. If you need a measure that is not yet implemented into CTAP, write to nadezda.okinina@eurac.edu. Or if you come across a resource (a list of words, for example) that could potentially be used for the calculation of a complexity measure for Italian, write to nadezda.okinina@eurac.edu. If you learn about the existence of a gold standard, please, let us know.

How to use the on line browser-based CTAP tool

CTAP is a browser-based tool.
1 In order to use it, go to the URL https://kommul-dev.eurac.edu/ctapWebApp and create an account.
2 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.
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 it too short, it won't be analysed by the tool. If a text is in

---

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

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

several languages, the analysis result will be erroneous, because most natural language processing components on which CTAP depends are language specific and when performing an 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.

3 Now that you have a corpus to analyse, you need to select the complexity features that you want to apply to 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, which are now a little bit less than 400: 270 for English, 159 for German (plus some features that don't work in the current version: *Lexical Richness: Type Token Ratio (Uber),* all the *Dependency Locality Theory* features *(20)*), 149 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: let us select 100. You will notice that features are listed in complete disorder: features for different languages, features belonging to different groups are all mixed. However, in order to analyse your corpus, you are only interested in features for 1 of the 3 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.

4 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. 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 and the results 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.

5 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.

6 Apart from downloading the results, you can visualise them as a graph. Go to the *Result Visualizer* menu on the left. Select the analysis your 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*.

7 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.

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.

Explanation of linguistic complexity features implemented in CTAP

Complexity features implemented in CTAP can be devided in 3 groups:
- lexical features
- syntactic features

Lexical features can also be divided in several groups:
- Lexical sophistication
- Lexical density
- Lexical variation
- Contextual diversity

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.

Contextual diversity: the diversity of contexts in which a word has been used in a text.

Many measures of lexical complexity can be calculated for **all words**, only for **lexical words** or only for **function words**.
Here are the abbreviations:
AW: all words
LW: lexical words
FW: function words

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

Lexical complexity measures

- 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 = number of types / number of tokens

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

Type is a word and token is its occurence 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 occurence 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(number of types) / log(number of tokens)

- Lexical Richness: Type Token Ratio (Corrected TTR)

Type is a word and token is its occurence 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

- 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

Syntactic complexity measures

- 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 Syntatic Constituents
> - Coordinate Phrases
> - Fragment T-units
> - Fragment Clauses
> - T-units
> - Verb Phrase
> - Sentences
> - Clauses
> - Dependent Clauses
> - Complex Norminal
> - Complex T-units