LATIC – A Linguistic Analyzer for Text and Item Characteristics

Documentation (Version 1.3.0)

Nadine Cruz Neri[1] & Florian Klückmann[2]

[1] Faculty of Education, Educational Psychology, Universität Hamburg

[2] Freelancing software engineer

Hamburg, 21st February 2023

Author Note

We would like to thank Judith Keinath for her editorial support.

Questions and concerns regarding this documentation as well as the software should be addressed to us via mail (hello@latic.software). You can also leave a comment on GitHub: https://github.com/floriankluecKmann/LATIC/issues.

Recommended citation: Cruz Neri, N., & Klückmann, F. (2023). *LATIC – A Linguistic Analyzer for Text and Item Characteristics* (Version 1.3.0) [Computer Software]. https://github.com/floriankluecKmann/LATIC

List of Contents

LATIC – A Linguistic Analyzer for Text and Item Characteristics

In recent decades, the role of linguistic text and item characteristics for comprehension and performance has been investigated in different domains, such as mathematics (e.g., Shaftel et al., 2006), science (e.g., Cruz Neri et al., 2021) and reading materials in general (e.g., Heppt et al., 2015). The goal for most researchers is to understand which linguistic text and item characteristics act as facilitators or inhibitors of individuals' comprehension (e.g., White, 2012). For this purpose, it is essential to analyze texts and items with regard to different linguistic characteristics of interest. LATIC, which stands for „Linguistic Analyzer for Text and Item Characteristics", is a software currently in development that analyzes such linguistic characteristics automatically and replaces error-prone manual tagging. LATIC intends to combine objectively quantifiable text and item characteristics, such as word count and average sentence length, with the value of natural language processing, enabling users to analyze parts of speech.

2 Technical Details

LATIC is a Java application that allows to analyze and count text and item characteristics in English, French, German and Spanish based on the Stanford CoreNLP 4.5.1 (Manning et al., 2014). The Stanford CoreNLP 4.5.1 (ibid.) uses natural language processing (for a detailed description see Chowdhury, 2005) to annotate words and symbols. The dependency parsing is based on the Universal Dependencies Project (2014–2020).

The annotation in English is based on the Penn Treebank Tagset (Santorini, 1990). In addition to the Stanford CoreNLP (Manning et al., 2014), LATIC enables the counting of other text and item characteristics, such as word count, number of sentences, and word length.

The analysis of texts and items in LATIC varies depending on (1) the length of the entered text or item, (2) the selected characteristics, and (3) the performance of the computer. The analysis usually only takes a few seconds.

LATIC can be used on Windows, Linux and macOS. To use LATIC, we recommend at least 8 GB RAM.

3 Languages

LATIC enables the user to tag and count linguistic characteristics by means of the Stanford CoreNLP 4.5.1 (Manning et al., 2014) in English, French, German and Spanish. The Stanford CoreNLP 4.5.1 (ibid.) can further process texts and items in Arabic, Chinese, Hungarian and Italian. Implementing the remaining languages into LATIC is possible. However, this requires the support of people who speak one of these languages at a very good level. If you are interested in a collaboration, feel free to contact the first author.

4 Instruction

4.1. Downloading and Starting LATIC

Before using LATIC, you must download the software. Visit the website https://download.latic.software; you will be forwarded to the latest LATIC release on GitHub. Under "Assets", you can download the .zip folder for your operating system. Download the folder "LATIC_Windows.zip" for Windows operating systems. Download the folder "LATIC_Ubuntu_Debian.zip" for Ubuntu or Debian based Linux operating systems. Download the folder "LATIC_jar.zip" for macOS, other Linux operating systems like OpenSUSE or Fedora or in case, you want to start LATIC via the terminal or rather console.

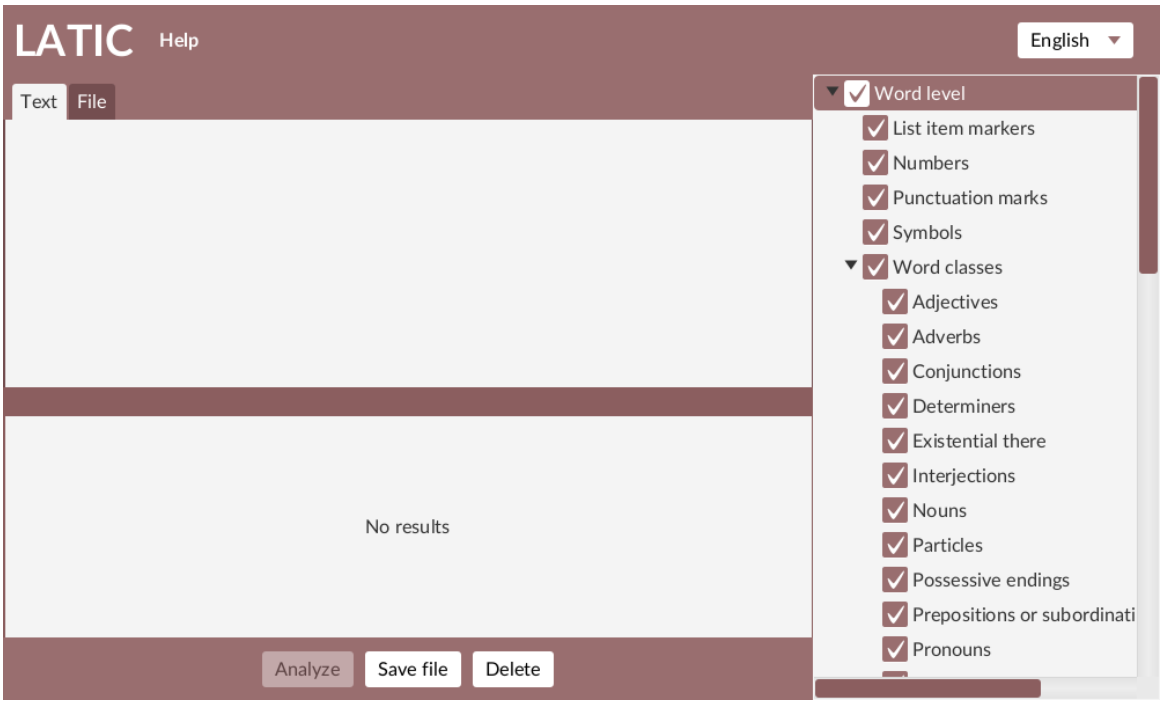
Next, you should right-click on the folder and click on "Extract here". Lastly, you can open the folder. In this folder, you find further instructions on how to start LATIC.

4.2 Usage

When you open LATIC, you will see the following user interface (Figure 1). Follow these steps in order to analyze the entered items and their item characteristics:

Figure 1

User Interface of LATIC



1. On the upper right, choose the language of the text or item.
2. Enter the text or item in the upper text field. Alternatively, you can select a file that shall be analyzed.
3. In the column on the right, select the characteristics that shall be analyzed. You can choose between different characteristics at the word, sentence and text level.
4. Click the "Analyze" button.
5. The results of the analysis will now be displayed.
6. To save the results table, click "Save file".
7. To delete the results, click "Delete". A pop-up window will appear asking you to confirm your decision to delete the results.
8. In case you need help, click „Help“ and select one of the available options.

4.3 General Information

In order to obtain reliable results, it is essential to keep a few factors in mind. First, you need to make sure to use correct spelling, including the upper and lower case. Otherwise, annotation errors are more likely to occur.

Second, we would recommend to avoid abbreviations. This way you get reliable results regarding specific text and item characteristics, such as word and sentence length.

Third, LATIC annotates every word or rather every letter and character combination as a unit. Hence, terms consisting of several words are annotated and counted repeatedly. For instance, in the sentence “The *Red Sea* has a high salt content.” both *Red* and *Sea* receive the tag NNP (singular proper noun). Accordingly, LATIC will count two proper nouns in this sentence.

Fourth, it is necessary to use a space between cardinal numbers and measuring units (e.g., "The equator is more than *40,000 km* long."). This way both the cardinal numbers and the measuring units are recognized as independent units and receive their own tag. Measuring units with potencies (e.g., "The area is 20 m^2.") are usually not recognized as measuring units and should be spelled out in order to be correctly annotated.

Fifth, the Stanford CoreNLP 4.5.1 (Manning et al., 2014) does not recognize email addresses and websites. In our test runs, they were often incorrectly annotated as adjectives.

5 Analysis of Item Characteristics

The following instructions, including the tagset and all functions of LATIC, only apply to the English language.

5.1 Analysis at the Word Level

The different parts of speech at the word level are tagged by means of the Stanford CoreNLP 4.5.1 (Manning et al., 2014). Every single word and punctuation mark is tagged individually (see Table 1 for an overview). LATIC enables the counting of the tags.

Table 1

Tagset of LATIC

Parts of speech	Further description	Tag	Examples
Adjectives	Adjectives	JJ	My *favorite* color is *red*.
	Comparative adjectives	JJR	Apples are *healthier* than sweets.
	Superlative adjectives	JJS	Your parents own the *largest* house I have ever been in.
Adverbs	Adverbs	RB	She smiled *shyly*.
	Comparative adverbs	RBR	Yesterday I ran *faster* on my run.
	Superlative adverbs	RBS	You are the one I liked the *most*.
	Wh-adverbs	WRB	*Why* did the chicken cross the road?
Cardinal numbers		CD	My dad picked *21* apples today.
Conjunctions	Coordinating conjunctions	CC	Cats *and* dogs are their favorite animals.
	Subordinating conjunctions	IN	*Despite* not sleeping well, I feel good.
Currencies		$	These shoes cost 10 €.
Determiners	Determiners	DT	Can I have *some* water?
	Predeterminers	PDT	That is nearly *double* the price!

Parts of speech	Further description	Tag	Examples
	Wh-determiners	WDT	*Which* one do you like best?
Existential there		EX	*There* have been worse times.
Interjections		UH	"*Aw*, that's sweet!", replied the girl.
List item markers		LS	How to bake a cake: *1*. Mix all the ingredients. *2*. Put the mixture in the oven. *3*. Let it bake for 45 minutes.
Modals		MD	I *will* be home today.
Nouns	Singular nouns	NN	The *sun* goes down in the *evening*.
	Plural nouns	NNS	Some *animals* hibernate during winter.
Particles		RP	I do not want to fall *over*.
Possessive ending		POS	Students' textbooks are expensive.
Prepositions		IN	The keys were *in* her pocket.
Pronouns	Personal pronoun	PRP	Bees produce honey to feed *themselves*.
	Possessive pronoun	PRP$	The kid forgot *his* teddy at home.

Parts of speech	Further description	Tag	Examples
	Possessive wh-pronoun	WP$	The kid, *whose* mother is a nurse, broke his leg.
	Wh-pronoun	WP	To *whoever* may read this: Keep going.
Proper nouns	Singular proper noun	NNP	My friend is planning a vacation on *Hawaii*.
	Plural proper nouns	NNPS	He is rooting for the *Lakers*.
Punctuation marks		PUNCT	. … ? ! , : ; (
Symbols		SYM	= + / \ ~ # > ^
To		TO	I need *to* go grocery shopping today.
Unknown or uncertain		X	
Verbs	Base form	VB	It is crucial to *keep* in touch.
	Gerund or present participle	VBG	The party guests are *dancing*.
	Past participle	VBN	The song was *played* in the radio.
	Past tense	VBD	The movie *started* at 7 o'clock.
	Singular present tense (third person)	VBZ	He usually *plays* the violin at noon.
	Singular present tense (non-third person)	VBP	We *listen* to a lot of music during road trips.

5.1.1 Adjectives

Adjectives are parts of speech that modify nouns (Cabredo Hofherr & Matushansky, 2010). Ordinal numbers (e.g., "The *third* place goes to him.") are identified as adjectives. Adjectives are given the tag JJ (e.g., "Lizzy was a *clever* girl."). Comparative adjectives are given the tag JJR (e.g., "He is *older* than his brother."). Superlative adjectives are given the tag JJS (e.g., "The Mount Everest is the *highest* mountain above sea level."). LATIC combines all tags into one when counting the adjectives in the entered text or item.

5.1.2 Adverbs

Adverbs are a part of speech that usually modify verbs or verb phrases as well as adjectives and other adverbs (Huddleston & Pullum, 2005). Non-cardinal numbers are tagged as adverbs (e.g., "I told you *twice.*"). The only known exception is the phrase "Once upon a time…". Adverbs are given the tag RB (e.g., "He feels *well.*"). Comparative adverbs are given the tag RBR (e.g., "Children typically learn languages *faster* than adults."). Superlative adverbs are given the tag RBS (e.g., "There is a waiting list of at *least* 10 people."). Wh-adverbs are given the tag WRB (e.g., "*When* did you have breakfast?"). LATIC combines all tags into one when counting the adverbs in the entered text or item.

5.1.3 Cardinal Numbers

Cardinal numbers indicate quantities and may function as adjectives, pronouns or determiners (Universal Dependencies Project, 2014–2020). The Stanford CoreNLP 4.5.1 (Manning et al., 2014) claims to identify cardinal numbers both spelled out and in the form of Arabic or Roman numerals. However, in testing the software, we found that the CoreNLP 4.5.1 (ibid.) had difficulties identifying Roman numerals and spelled-out numbers. Hence, we recommend writing cardinal and decimal numbers in their Arabic form for reliable results. Non-cardinal numbers are tagged as adjectives (e.g., "The students are in *second* grade.") or

adverbs (e.g., “The kid fell *twice*.”). Cardinal numbers are given the tag CD (e.g., “The arithmetic mean is *9.25*.”).

5.1.4 Conjunctions

Conjunctions depict the connection(s) of two or more (subordinate) clauses (Huddleston & Pullum, 2005). There are two types of conjunctions to be distinguished: coordinating conjunctions and subordinating conjunctions.

5.1.4.1 Coordinating Conjunctions. Coordinating conjunctions typically connect clauses of equal rank, such as a main clause with another main clause or a subordinate clause with another subordinate clause (Danlos et al., 2018). Coordinating conjunctions are given the tag CC (e.g., “Do you prefer your tea with sugar *or* milk?”).

5.1.4.2 Subordinating conjunctions. Subordinating conjunctions are parts of speech that introduce subordinate clauses. The Stanford CoreNLP 4.5.1 (Manning et al., 2014) combines subordinating conjunctions and prepositions (see section 5.1.13). We assume that the CoreNLP 4.5.1 (ibid.) combines these two parts of speech for two reasons. First, some words, such as *before* and *since*, are polyfunctional and hence, may function as a preposition or as a conjunction. Second, some authors argue that prepositions and conjunctions should be classified as one class rather than two separate ones (for a discussion, see Haumann, 1997). Subordinating conjunctions (and prepositions) are given the tag IN (e.g., “The little boy cried *because* he laughed so hard.”).

5.1.5 Determiners

Determiners are only found in noun phrase structures. They define noun phrases as definite (e.g., “They liked *the* performance.”) or indefinite (e.g., “I want *a* cookie.”) (Huddleston & Pullum, 2005). Determiners include articles, quantifiers, interrogative determiners and more. As pointed out by the Universal Dependencies Project (2014–2020), the difference between determiners and pronouns (see section 5.1.14) is not always clearly

defined. Determiners are given the tag DT (e.g., "*The* staff has *a* meeting once *a* week."). Predeterminers are given the tag PDT (e.g., "They spent *all* the money."). Wh-determiners are given the tag WDT (e.g., "*Which* is your favorite subject at school?"). LATIC combines all tags into one when counting the determiners in the entered text or item.

5.1.6 Existential There

The word *there* refers to a dummy pronoun or the phrase meaning "in or at that place" (Huddleston & Pullum, 2005, p. 249). It is called existential since it declares the existence of living beings and other things. The existential there is given the tag EX (e.g., "*There* is good reason to be proud of yourself.").

5.1.7 Interjections

Interjections are expressions that serve the purpose of expressing individuals' emotions or reactions. While primary interjections only belong to the category of interjections (e.g., "*Huh*?"), secondary interjections are words that usually are categorized as different parts of speech ("*Thank God*!") (Ameka, 1992). The Stanford CoreNLP 4.5.1 (Manning et al., 2014) primarily tags primary interjections. Secondary interjections are rarely tagged correctly. Interjections are given the tag UH (e.g., "*Wow*, you look great today!").

5.1.8 List Item Markers

According to Santorini (1990), the tag of list item markers includes letters as well as numbers that define a list. However, in testing the software, we found that only Arabic numbers were identified as list item markers by the Stanford CoreNLP 4.5.1 (Manning et al., 2014). Furthermore, the list item markers indicated with a period were not tagged correctly when exceeding the fifth marker. List item markers indicated by closing brackets are tagged correctly. Neither bullet points nor letters were correctly identified as list item markers.

List item markers are given the tag LS. Example:

1) Get your ingredients.

2) Turn on the stove.

3) Cook the pasta according to instructions.

5.1.9 Modals

Modals are a form of auxiliary verbs that usually mark mood (e.g., imperative), verb tenses, and (active/ passive) voice (Huddleston & Pullum, 2005). Modals are given the tag MD (e.g., "*May* I have a look at it?").

5.1.10 Nouns

Nouns usually refer to physical objects, entities, concepts, actions, people and events (Huddleston & Pullum, 2005). Singular nouns are given the tag NN (e.g., "A rolling *stone* gathers no *moss*."). Plural nouns are given the tag NNS (e.g., "*Humans* and *dolphins* are *mammals*."). LATIC combines both tags into one when counting the nouns in the entered text or item.

5.1.11 Particles

According to the Universal Dependencies Project (2014–2020), particles are parts of speech that (1) only convey meaning when linked to another word or phrase, and (2) do not meet the criteria for other main classes of parts of speech. Particles are considered a big source of ambiguity when tagging parts of speech automatically by means of natural language processing (Jianjun et al., 2011). However, the Stanford CoreNLP 4.5.1 (Manning et al., 2014) achieves satisfactory results in tagging particles correctly (ibid.). Particles are given the tag RP (e.g., "Don't give *up*!").

5.1.12 Possessive Endings

The possessive ending refers to the noun endings *'s* (singular) and*'* (plural). It is tagged separately from the noun. Possessive endings are given the tag POS (e.g., "The flower*'s* petal has beautiful colors.").

5.1.13 Prepositions

Prepositions express relations between phrases, such as spatial, causal, or temporal relations. The Stanford CoreNLP 4.5.1 (Manning et al., 2014) combines prepositions and subordinating conjunctions (see section 5.1.4.2). Prepositions (and subordinating conjunctions) are given the tag IN (e.g., "The dog was sitting *on* the chair.").

5.1.14 Pronouns

Pronouns are a subclass of a noun and function as the head of a noun phrase (Huddleston & Pullum, 2005). Depending on the type of pronoun, the Stanford CoreNLP 4.5.1 (Manning et al., 2014) tags pronouns differently. LATIC combines all tags into one when counting the pronouns in the entered text or item.

5.1.14.1 Personal Pronouns. Personal pronouns refer to a grammatical person (Huddleston & Pullum, 2005). Reflexive pronouns (e.g., "Timothy taught *himself* to cook.") are tagged as personal pronouns, too. Personal pronouns are given the tag PRP (e.g., "*I* love this song.").

5.1.14.2 Possessive Pronouns. Possessive pronouns indicate possession in a broad sense, for example of an object, an animal or an idea (Huddleston & Pullum, 2005). Possessive pronouns are given the tag PRP$ (e.g., "Birds spread *their* wings to fly."). Possessive wh-pronouns are given the tag WP$ (e.g., "*Whose* lunch is this?").

5.1.14.3 Wh-pronouns. Wh-pronouns are given the tag WP (e.g., "*Who* let the dogs out?"). Possessive wh-pronouns are given the tag WP$ (see section 5.1.14.2).

5.1.15 Proper Nouns

Proper nouns refer to the name of a specific living being, a place, an object, etc. (Huddleston & Pullum, 2005). The Stanford CoreNLP 4.5.1 (Manning et al., 2014) should be able to tag common proper nouns. Singular proper nouns are given the tag NNP (e.g., "I live in *Hamburg*, *Germany*."). Plural proper nouns are given the tag NNPS (e.g., "Lucky Luke

arrested the *Daltons*."). LATIC combines both tags into one when counting the proper nouns in the entered text or item.

5.1.16 Punctuation Marks

Punctuation marks, which are non-alphabetical and non-numeral characters, indicate the syntactical structures of a text (Universal Dependencies Project, 2014–2020). Punctuation marks are given the tag PUNCT (e.g., "Wow*!* LATIC recognizes different punctuation marks*,* such as commas*,* colons and exclamation points*!*").

5.1.17 Symbols

Symbols differ from words regarding their function, form or both (Universal Dependencies Project, 2014–2020). Symbols are given the tag SYM (e.g., "The law students read *§§* 2–12."). Currencies – if not spelled out – are given the tag $ (e.g., "This book costs 20 *$*."). LATIC combines both tags into one when counting the symbols in the entered text or item.

5.1.18 To

The word "to" is an infinitive marker. In case of the word "to" being a preposition (e.g., "Let's go *to* the mall."), it is tagged as such. The infinitive marker "to" is given the tag TO (e.g., "I have *to* wake up early tomorrow.").

5.1.19 Unknown and Uncertain

Parts of speech that cannot be tagged as a specific part of speech will be tagged as unknown words or rather as belonging to an uncertain category. They are given the tag X.

5.1.20 Verbs

Verbs express actions (Huddleston & Pullum, 2005). The Stanford CoreNLP 4.5.1 (Manning et al., 2014) differentiates between tenses, which locate actions and/ or events relative to the present (Comrie, 1985). LATIC combines all tags into one when counting the verbs in the entered text or item.

5.1.20.1 Base Form. The base form, sometimes referred to as the plain form, is identical with a verb's lexical base. However, it is not a form of the present tense, but rather a part of the imperative and the subjunctive (Huddleston & Pullum, 2005). Verbs in their base form are given the tag VB (e.g., "*Stay* safe.").

5.1.20.2 Gerund or Present Participle. The gerund's function is similar to a noun, while the participle's function is similar to an adjective. These tags are combined since English verbs do not have "different forms corresponding to the two uses" (Huddleston & Pullum, 2005, p. 32). Hence, the gerund and the present participle are sometimes referred to as the gerund-participle. Verbs in the gerund and the present participle are given the tag VBG (e.g., "The child is *running* around.").

5.1.20.3 Past participle. The past participle usually accompanies perfect and passive constructions (Huddleston & Pullum, 2005). Verbs in the past participle are given the tag VBN (e.g., "She had *called* her best friend.").

5.1.20.4 Past Tense. The past tense primarily expresses actions that happened in the past (Huddleston & Pullum, 2005). Verbs in the past tense are given the tag VBD (e.g., "He *bought* his mother a cake for her birthday.").

5.1.20.5 Singular Present Tense. The present tense indicates action of the present time (Huddleston & Pullum, 2005). Verbs in the third person singular present tense are given the tag VBZ (e.g., "He *plays* with his puppets."). Verbs in the non-third person singular present tense are given the tag VBP (e.g., "I *go* to school."). LATIC combines both tags into one when counting the verbs in the singular present tense in the entered text or item.

5.1.21 Word Length

The average word length is calculated by dividing the total number of (1) characters or (2) syllables by the total word count. See Table 2 for examples.

Table 2

Examples for the Analyses of Average Word Length

Examples	Average word length
Cats have four legs. 4 characters 4 characters 4 characters 4 characters 1 syllable 1 syllable 1 syllable 1 syllable	4.0 characters 1.0 syllables
An apple a day keeps the 2 characters 5 characters 1 character 3 characters 5 characters 3 characters 1 syllable 2 syllables 1 syllable 1 syllable 1 syllable 1 syllable	3.63 characters 1.38 syllables
doctor away. 6 characters 4 characters 2 syllables 2 syllables	

5.2 Analysis at the Sentence Level

5.2.1 Number of Sentences

A sentence finishes with the following punctuation marks: periods (.), interrogation points (*?*), and exclamation points (*!*). Sentences divided by line breaks, colons (*:*) or semicolons (*;*) are not counted as two (or more) sentences, but are rather as one. In case you would like to count sentences ending with colons and semicolons as separate sentences, it is recommended to change these punctuation marks into periods. See Table 3 for examples.

Table 3

Examples for the Analyses of the Number of Sentences

Examples	Number of sentences
LATIC is an analysis tool. It stands for "Linguistic Analysis Tool for Item Characteristics".	2
LATIC is an analysis tool: It stands for "Linguistic Analysis Tool for Item Characteristics".	1

5.2.2 Sentence Length

The average sentence length is calculated by dividing the (1) number of characters (with or without spaces), (2) the number of syllables or (3) the word count by the number of sentences. Punctuation marks indicating the end of a sentence will not be included in the analyses of sentence length. See Table 4 for examples.

Table 4

Examples for the Analyses of Average Sentence Length

Examples					Sentence length
I 1 character 1 syllable	have 4 characters 1 syllable	an 2 characters 1 syllable	exam 4 characters 2 syllables	tomorrow. 8 characters 3 syllables	24.0 characters 20.0 characters without spaces 8.0 syllables 5.0 words
It 2 characters 1 syllable	is 2 characters	my 2 characters	birthday 8 characters	today. 5 characters	27.0 characters 24.0 characters without spaces 7.0 syllables 6.0 words

5.3 Analysis at the Text Level

5.3.1 Connectives

Connectives or cohesive devices are words or expressions that link sentences or parts of sentences (Breindl et al., 2014). There are different types of connectives: For instance, they can indicate temporal ("*As soon as* I get home, I will eat something.") or causal links ("I'm eating an apple *because* I am hungry."). Connectives are recognized and counted in LATIC using an algorithm. Correlative conjunctions (e.g. "The performer *not only* sang live, *but also* wrote the songs.") are counted as one connective.

5.3.2 Readability Indices

Readability indices are measurements to estimate the linguistic difficulty of a text or an item. In the English language, LATIC can calculate seven readability indices.

5.3.2.1 Automated Readability Index (ARI). The ARI was developed by Senter and Smith (1967). The values provide an estimate on which US grade level is necessary to understand a text or an item and vary between one and 14. The formula for calculating the ARI is:

$$ARI = 4.71 * \frac{number\ of\ characters}{word\ count} + 0.5 * \frac{word\ count}{number\ of\ sentences} - 21.43$$

5.3.2.2 Coleman-Liau Index (CLI). The CLI was developed by Coleman and Liau (1975). The values produce an estimate on which grade level is needed to understand the text or item. The values vary between one and 16 (ibid.), while values of 13 and higher indicate the need for a college education (Huang et al., 2015). The formula for the index is:

$$CLI = 0.0588 * \left(\frac{number\ of\ characters}{word\ count} * 100\right) - 0.296 * \left(\frac{number\ of\ sentences}{word\ count} * 100\right) - 15.8$$

5.3.2.3 Flesch-Kincaid Grade Level (FKGL). The FKGL provides an estimate on how many years of schooling a person needs to understand a text or an item (Flesch, 1979; Kincaid et al., 1975). The values range from one to 18 and above (see Kincaid et al., 1975). The formula for calculating the index is:

$$FKGL = 0.39 * \frac{word\ count}{number\ of\ sentences} + 11.8 * \frac{number\ of\ syllables}{word\ count} - 15.59$$

5.3.2.4 Flesch Reading Ease (FRE). The Flesch Reading Ease was developed by Flesch (1948). The formula for calculating the index is:

$$FRE = 206.835 - (1.015 * sentence\ length) - (84.6 * word\ length\ (syllables))$$

The values range from zero to 100 and can be interpreted as follows (see Table 5).

Table 5

Interpretation of FRE Values

Difficulty	FRE value
Very difficult	under 29
Difficult	30–49
Medium	50–79
Easy	80–89
Very easy	90–100

5.3.2.5 Gunning Fog Index (GFI). The GFI was developed by Gunning (1952) and produces an estimate on which grade level is needed to understand a text or an item. The values range between six and 17. The formula for calculating the index is:

$$GFI = 0.4 * \left(\frac{word\ count}{number\ of\ sentences} + \frac{words\ with\ \geq 3\ syllables}{word\ count}\right)$$

5.3.2.6 LIX. LIX is short for *Læsbarhedsindex* (in English: readability index) and was developed by Carl-Hugo Björnsson (1968). The formula for calculating the LIX is:

$$LIX = \frac{word\ count}{number\ of\ sentences} + \frac{long\ words}{word\ count} * 100$$

The values of the LIX vary between zero and 100. Table 6 depicts how the LIX values may be interpreted according to Björnsson (1968).

Table 6

Interpretation of LIX Values

Difficulty	LIX value
Very easy	under 29
Easy	30–39
Medium	40–49
Difficult	50–59
Very difficult	over 60

Note. The second example stems from Wikipedia ("DNA", 2021).

5.3.2.7 SMOG. The SMOG is short for "Simple Measure of Gobbledygook" and was developed by McLaughlin (1969). The values range from one to 19 and above and indicate the school years needed to be completed to understand a text or an item (ibid.). The formula for calculating the SMOG is:

$$SMOG = 1.043 * \sqrt{30 * \frac{words\ with\ \geq 3\ syllables}{number\ of\ sentences}} + 3.1291$$

5.3.3 Lexical Diversity

Lexical diversity, also called type-token-ratio, refers to the number of different words divided by the total word count of a text or an item (Johansson, 2009). See Table 7 for examples.

Table 7

Examples for the Analyses of Lexical Diversity

Examples	Lexical diversity
I do not always use LATIC, *but when* I do, I *enjoy it*.	Number of different words: 10
	Word count: 13
	Lexical diversity: 0.77
The itsy bitsy spider climbed up the *water spout*.	Number of different words: 18
Down came the *rain and washed* the spider *out*.	Word count: 38
Out came the *sun* and *dried* up *all* the rain.	Lexical diversity: 0.47
And the itsy bitsy spider climbed up the spout *again*.	

Note. The words that count towards the number of different words are in italics.

5.3.4 Syllable Count

The syllable count feature gives out the total number of syllables in an entered text or item.

5.3.5 Tagged Text or Item

The tagged text or item option allows depicting the tags for every single part of speech in the entered text or item. An example is illustrated in Figure 2.

Figure 2

Example of the Tagged Text or Item Option

Mammals/NNS are/VBP warm/JJ -/PUNCT blooded/JJ vertebrates/NNS ./PUNCT

Humans/NNS and/CC many/JJ animals/NNS belong/VBP to/IN the/DT class/NN of/IN Mammalia/NNP ./PUNCT

5.3.6 Word Count

LATIC identifies words as the juxtaposition of characters, separated by spaces and punctuation marks.

6 Evaluation

To evaluate LATIC, different features were tested and optimized separately (see also Cruz Neri et al., 2022).

6.1 Part of speech tagging

The evaluation of the part of speech tagging was done using the Stanford CoreNLP 4.2.0 (Manning et al., 2014). We further used the MULTEXT-East 4.0 corpus (Erjavec et al., 2010). This corpus, consisting of the book 1984 by George Orwell, was manually tagged and is freely available online (Erjavec, 2012). It used a specific tagset (Ide et al., 2009) tagging only parts of speech. Punctuation marks are not tagged at all.

First, we used the initial 10,006 tokens of the MULTEXT-East 4.0 corpus (Erjavec et al., 2010) to evaluate the tagging of the Stanford CoreNLP 4.2.0 (Manning et al., 2014). While testing the Stanford CoreNLP 4.2.0 (ibid.), we found that correct spelling is essential for reliable results. Hence, we removed misspelled words and expressions, obtaining a final sample of $N = 9{,}989$ tokens.

Second, we matched the tagsets in order to be able to compare the tokens of the MULTEXT-East 4.0 corpus (Erjavec et al., 2010) and those of the Stanford CoreNLP (Manning et al., 2014). For instance, the Stanford CoreNLP (ibid.) uses four different tags for different types of pronouns, while the MULTEXT-East 4.0 corpus (Erjavec et al., 2010) uses approximately 45 different tags. After matching the tagsets, the tokens determined by the MULTEXT-East 4.0 corpus (ibid.) corresponded 86.61 % ($n = 8{,}651$) to the tags given out by the Stanford CoreNLP 4.2.0 (Manning et al., 2014).

Third, the comparison needed to be corrected. Some tags were not convertible. Although the MULTEXT-East 4.0 corpus (Erjavec et al., 2010) has a more differentiated tagset (Ide et al., 2009), for instance, particles and the infinitive “to” were not tagged specifically. Furthermore, the Stanford CoreNLP (Manning et al., 2014) tags every word and

punctuation mark individually, whereas the MULTEXT-East 4.0 corpus (Erjavec et al., 2010) does not. For instance, in the MULTEXT-East 4.0 corpus (ibid.) the word "dark-haired" is tagged as an adjective, while the Stanford CoreNLP 4.2.0 (Manning et al., 2014) tags it as an adjective – punctuation mark – adjective. After considering these circumstances and including them into the evaluation, 92.75 % tokens were correctly identified by the Stanford CoreNLP 4.2.0 (ibid.).

The errors in tagging made by the Stanford CoreNLP 4.2.0 (ibid.) were analyzed based on the first 5,000 tokens. By far the most errors in tagging occurred in the distinction between specific types of determiners and pronouns (approximately 37.84 % of all errors). However, this was to be expected since the distinction between pronouns and determiners is often ambiguous (Universal Dependencies Project, 2014–2020). The second and third most frequent errors were found in the distinction between adjectives and (1) prepositions and subordinating conjunctions (approximately 11.28 % of all errors), and (2) all types of verbs (approximately 8.52 % of all errors).

The tagging accuracy of different taggers varies between studies, which may be contributed to the used data sets and the genre of texts and items. For instance, taggers achieve an accuracy of up to 90.3 % in bug reports (Tian & Lo, 2015) and up to 94.6 % in biomedical texts (Ling et al., 2008). Drawing on prior research (e.g., Khin & Aung, 2016), the tagging by the Stanford CoreNLP (Manning et al., 2014) implemented in LATIC achieves good results.

6.2 Syllable Count

LATIC uses an algorithm provided by Wormer (2021) to count the syllables. This algorithm was tested on $N = 9{,}107$ common English words. The syllables were counted correctly for $n = 8{,}723$ words (95.78 %). In a second step, we optimized the algorithm by

adding a list of words for which syllables were not counted correctly. After optimizing the algorithm, the syllable count was correct for all $N = 9{,}107$ test words (100 %).

6.3 Connectives

The counting of the connectives is based on a self-created algorithm. Three different texts (total length: 2,349 words) were used to evaluate the algorithm. LATIC recognized $n = 309$ out of $N = 331$ connectors (93.35 %) occurring in the three texts. After optimizing the algorithm, LATIC now recognizes $n = 326$ of $N = 331$ connectors (98.49 %).

7 References

Ameka, F. (1992). Interjections: The universal yet neglected part of speech. *Journal of Pragmatics, 18*(2–3), 101–118. https://doi.org/10.1016/0378-2166(92)90048

Björnsson, C. H. (1968). *Läsbarhet* [Readability]. Gad.

Breindl, E., Volodina, A., & Waßner, U. H. (2014). *Handbuch der deutschen Konnektoren 2* [Handbook of German Connectives 2]. De Gruyter.

Cabredo Hofherr, P., & Matushansky, O. (2010). *Adjectives: Formal analyses in syntax and semantics*. John Benjamins Publishing Company.

Chowdhury, G.G. (2005). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), 51–89. https://doi.org/10.1002/aris.1440370103

Coleman, M., & Liau, T.L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283–284. https://doi.org/10.1037/h0076540

Comrie, B. (1985). *Tense*. University Press.

Cruz Neri, N., Guill, K., & Retelsdorf, J. (2021). Language in science performance: Do good readers perform better? *European Journal of Psychology of Education, 36*(1), 45–61. https://doi.org/10.1007/s10212-019-00453-5Cruz Neri, N**.**, Klückmann, F., & Retelsdorf, J. (2022). LATIC – A linguistic analyzer for text and item characteristics. *PLoS ONE, 17*(11), e0277250. https://doi.org/10.1371/journal.pone.0277250

Danlos, L., Rysová, K., Rysová, M., & Stede, M. (2018). Primary and secondary discourse connectives: Definitions and lexicons. *Dialogue & Discourse,* 9(1), 50–78. https://doi.org/10.5087/dad.2018.10

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic resources for central and eastern European languages. *Language Resources & Evaluation, 46*, 131–142. https://doi.org/10.1007/s10579-011-9174-8

Erjavec, T., Barbu, A.-M., Derzhanski, I., Dimitrova, L., Garabík, R., Ide, N., Kaalep, H.-J., Kotsyba, N., Krstev, C., Oravecz, C., Petkevič, V., Priest-Dorman, G., QasemiZadeh, B., Radziszewski, A., Simov, K., Tufiş, D., & Zdravkova, K. (2010). *MULTEXT-East "1984" annotated corpus 4.0*. Slovenian resource repository CLARIN. https://www.clarin.si/repository/xmlui/handle/11356/1043

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Flesch, R. (1979). *How to write in plain English: A book for lawyers and consumers*. Harper.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

Haumann, D. (1997). *The syntax of subordination*. De Gruyter.

Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly, 50*(1)*,* 61–82. https://doi.org/10.1002/rrq.83

Huang, G., Fang, C.H., Agarwal, N., Bhagat, N., Eloy, J.A., & Langer, P.D. (2015). Assessment of online patient education materials from major ophthalmologic associations. *JAMA ophthalmology*, *133*(4), 449–454. https://doi.org/10.1001/jamaophthalmol.2014.6104

Huddleston, R., & Pullum, G.K. (2005). *A student's introduction to English grammar.* Cambridge University Press.

Ide, N., Priest-Dorman, G., Erjavec, T., & Varadi, T. (2009, October 6). *MULTEXT-East morphosyntactic specification, version 4. English specifications*. Natural Language Server. http://nl.ijs.si/ME/Vault/V4/msd/html/msd-en.html

Jianjun, M., Degen, H., Haixia, L., & Wenfeng, S. (2011, May 1). *POS tagging of English particles for machine translation* (pp. 57–63). Proceedings of the Thirteenth Machine Translation Summit, Xiamen, China.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers, 53*(2008), 61–79.

Khin, N.P.P., & Aung, T.N. (2016). Analyzing tagging accuracy of part-of-speech taggers. In T. Zin, J.W. Lin, J.S. Pan, P. Tin, & M. Yokota (Eds.), *Genetic and evolutionary computing. Advances in intelligent systems and computing* (pp. 347–354). Springer.

Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., & Chissom, B.S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.

Ling, M.H.T., Lefevre, C., & Nicholas, K. R. (2008). Parts-of-speech tagger errors do not necessarily degrade accuracy in extracting information from biomedical text. *The Python Papers, 3*(1), 65–80.

Manning, C.D., Surdeani, M., Bauer, J., Finkel, J., Bethard, S.J., & McClosky, D. (2014, June 23–24). *The Stanford CoreNLP Natural Language Processing Toolkit*. Proceedings of 52nd Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, United States. https://www.aclweb.org/anthology/P14-5010.pdf

McLaughlin, G.H. (1969). SMOG grading – a new readability formula. Journal of Reading, 12(8), 639–646.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

Senter, R.J., & Smith, E.A. (1967). Automated readability index. Wright-Patterson Air Force Base.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105–126. https://doi.org/10.1207/s15326977ea1102_2

Tian, Y., & Lo, D. (2015, March 2–6). *A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports* (pp. 570–574). *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, Montreal, QC, Canada. https://ieeexplore.ieee.org/document/7081879

Universal Dependencies (2014–2020). *Universal POS tags*. https://universaldependencies.org/u/pos/

White, S. (2012). Mining the text: 34 text features that can ease or obstruct text comprehension and use. *Literacy Research and Instruction, 51*(2), 143–164. https://doi.org/10.1080/19388071.2011.553023

Wormer T. (2021). *Syllable* [cited 2022 Sep 19]. Database: GitHub repository. Retrieved from: https://github.com/words/syllable