1. Linguistic analysis

First, the tool carries out linguistic analysis that identifies the word class of each word in a text (token) and assigns it the dictionary form (headword) using the TreeTagger (Schmid 1994). Each token is then compared with the headword and its specific inflectional form (exponence) is identified, with the following linguistic analysis, that accounts for both regular and irregular forms.

Premises
- The aim is to count exponences, i.e. forms, thus no reference will be made to their functions, i.e. to the semantic of syntactic properties they encode.
- The present operationalization only applies to written morphology, and exemplification is limited to verbs. However, it can be extended to other word classes and to oral forms.

Assumptions
- Inflected words can be analyzed as a lexical base (the stem) plus one or more exponents, which together constitute its exponence.
- For any lexeme it is possible to identify a default stem (DS), defined as the stem that is common to most cells of that lexeme's paradigm in the target language. If two or more stems occupy exactly the same number of cells, then decision as to which one should count as default can be made on theoretical grounds or by flipping a coin.

Procedure
- Identify the default stem.
- Identify exponences by describing how inflected word forms relate to the default stem, using the following notation format.

Examples from English

	notation	sample WF(s)	DS	exponence(s)
WF is identical to DS	Ø	cut (present or past tense)	cut	Ø
WF consists in DS + additional graphemes at the end of the DS	additional graphemes	cuts risen, taken fallen	cut rise, take fall	s n en
WF consists in DS + additional graphemes in the middle of the DS	_additional graphemes_	none?		
WF consists in DS minus some graphological material at the end of the DS	£[deleted grapheme(s)]	hid	hide	£e
WF consists in DS minus some graphological material in the middle of the DS	_£[deleted grapheme(s)]_	fed led	feed lead	_£e_ _£a_
WF consists in DS + additional graphemes replacing parts of the DS at the end of the DS	[replaced graphemes]/[new graphemes]	bought thought sought left spelt told, sold	buy think seek leave spell tell, sell	uy/ought ink/ought eek/ought eave/eft l/t ell/old

WF consists in DS + additional graphemes replacing parts of the DS in the middle of the DS	_[replaced graphemes]/[new graphemes]_	found, ground drove, rode	find, grind drive, ride	_i/ou_ _i/o_
multiple aspects		kept, felt	keep, feel	_£e_t
multiple aspects		broke, stole	break, steal	_ea/o_e
multiple aspects		sworn, torn	swear, tear	_ea/o_n

Examples from German

	notation	sample WF(s)	DS (typically, infinitive minus -en)	exponence(s)
WF is identical to DS	∅	none?		
WF consists in DS + additional graphemes at the end of the DS	additional graphemes	backte	back	te
WF consists in DS + additional graphemes at the beginning of the DS	additional graphemes_	gebacken geschenkt	back schenk	ge_en ge_t
WF consists in DS minus some graphological material at the end of the DS	£[deleted grapheme(s)]	none?		
WF consists in DS minus some graphological material in the middle of the DS	_£[deleted grapheme(s)]_	none?		
WF consists in DS + additional graphemes replacing parts of the DS at the end of the DS	[replaced graphemes]/[new graphemes]	none?		
WF consists in DS + additional graphemes replacing parts of the DS in the middle of the DS	_[replaced graphemes]/[new graphemes]_	bot lieh	biet leih	_ie/o_ _ei/ie_
multiple aspects		bat, fiel	bitt, fall	_i/a_£t, _a/ie_l£
multiple aspects		bricht, brichst	brech	_e/i_t, _e/i_st
multiple aspects		gedurft	dürf	ge__ü/u_t
multiple aspects		ließ maß	lass mess	_a/ie_ß _/a_ß

NB: If exponence consists in additional graphemes added to the DS's right margin, these will appear as a simple suffix. When graphemes are added in other positions in the DS, this will be noted as follows: exp_ for graphemes added at the left margin; _exp_ for graphemes affecting the central part of the DS (no difference is made regarding their exact position w.r.t. the base).

NB: completely or highly suppletive forms will be listed as such, with no reference to their DS. For these cases, an exhaustive list will be provided.

Qualifications
- The procedure is purely descriptive. Even though the description of differences between DS and WF may be made using terms like 'addition', 'replacement', 'reduction' etc., no claim is made as to their historical, theoretical or psycholinguistic reality.
- Given that the aim is to assess morphological complexity, exponences that are the exclusive result of systematic orthography rules will be considered to be allographs and assimilated to other exponences representing the same phonological material (e.g. *paid, said, arrived* will all be considered simple cases of -ed exponence like *played*). This slightly compromises the coherence of the 'graphical morphology' construct, but treating differently written forms as different morphological exponences would artificially inflate morphological complexity values by adding orthographical complexity, which should be seen as a more serious threat to validity. However, allographs that cannot be explained by systematic orthography rules will be treated as allomorphs, e.g. *bought – caught*.
- When used for analyzing learner varieties, the procedure implies some form of "comparative fallacy" (Bley-Vroman 1983): default stems are identified by reference to the whole paradigm in the target language. This is true, but at least the target language provides an explicit and shared framework for describing exponences across learners and stages. However, users may wish to arrive at different exponences, based on careful analyses of individual interlinguistic systems, where for example the default stem is *went* and it is inflected as in *wents, will went, wented*. In such cases one only needs to be explicit and be able to defend one's operationalization in analytical and theoretical terms.

Arguments for this approach
- It is relatively simple
- It is basically in line with standard grammatical descriptions and with an 'item-and-process' approach to morphology
- It provides an explicit procedure which should produce high interrater agreement regarding word segmentation
- It doesn't draw any line between morphological processes creating new stems and those consisting in pure affixation. This implies for example that there is no need for a distinction between regular / irregular verbs, or between small and large inflection classes.

Questions and answers
Q: Why limit the procedure to inflectional forms, disregarding the semantic and syntactic properties they encode?
A: Mainly because the measure is intended to be used for describing also L1 and L2 acquisition, where it is often unclear exactly what functions are encoded by a given grammatical form. Even in native languages with extensive grammars, sometimes accounts differ as to what properties are expressed by a given morphological formative.

Q: why are exponences represented as actual strings of graphemes instead of more abstract operations like 'fronting a vowel', 'diphtonging a vowel', 'doubling a consonant' etc'
A: The construct concerns the actual forms inflected words may take, and this can be done only by reference to the actual graphological material they contain. Saying that a vowel is fronted or diphtonged refers to abstract representations of general processes, not to concrete representations of a word's shape.

2. Mathematical analysis

Second, after the text has been linguistically analyzed and exponences have been extracted, the tool computes the Morphological Complexity Index (MCI). This is operationalized by randomly drawing sub-samples of N forms of a word class (e.g. verbs) from a text and computing the average within- and across-sample range of inflectional exponences. Thus, MC = (within-subset variety + between-subset diversity/2) – 1.
The field ‘segment size’ specifies the number N of forms constituting each sub-sample; the field ‘random trials’ indicates for how many times pairs of N-forms subsamples are extracted from the text.
N.B.: While the approach to computing the MCI is the same as that proposed in Pallotti (2015), the actual mathematical formula has been slightly changed

How to cite
Pallotti, G., Brezina, V. (2023) Computing the Morphological Complexity Index. Available from http://corpora.lancs.ac.uk/vocab/analyse_morph.php