RiCL Research in Corpus Linguistics

The *Construction Complexity Calculator* (ConPlex): A tool for calculating Nelson's (2024) construction-based complexity measure

Christopher R. Cooper
Rikkyo University / Japan

Abstract – The current study aims to increase the accessibility of Nelson's (2024) recently suggested construction-based complexity measure by providing a tool that can calculate the measure for single or multiple texts. To validate the tool, complexity scores for the *International Corpus Network of Asian Learners of English* corpus (ICNALE) were compared with Nelson's (2024) results. In addition, complexity scores were calculated for a new dataset, the *Common European Framework of Reference English Listening Corpus* (CEFR), along with the *MERLIN* corpus, which includes learner writing samples from learners of Czech, German, and Italian. Complexity scores generally increased across CEFR levels in all of the datasets. However, the complexity scores in the current study tend to be higher than the original study due to differences in the sentence splitting approach. The sentence tokenisation method used is deemed to be more appropriate, and it may be concluded that the *Construction Complexity Calculator* (ConPlex) tool accurately calculates Nelson's measure. It is hoped that the tool will allow researchers to calculate the complexity of constructions at the text level for a wide range of research purposes.

Keywords – constructions; complexity; corpus tool creation; corpus tool validation; language development

1. INTRODUCTION

For open science to live up to its name, the code used for data analysis should be shared, and researchers should be actively involved in the development of freely available tools (Mizumoto 2024). Several tools are already available for measuring an array of indices that can be utilised to assess complexity and other elements of the production of second language (L2) users and other texts. These include *Coh-Metrix* (McNamara *et al.* 2014), the tool for the *Automatic analysis of Syntactic Sophistication and Complexity* (TAASSC; Kyle 2016), and the tool for the *Automatic Analysis of Lexical Diversity* (Kyle *et al.* 2021). The underlying code has also been made available for the latter two

DOI 10.32714/ricl.13.02.05

tools on GitHub.[1] The present study is a further contribution to open science and aims to 1) design a tool —the *Construction Complexity Calculator* (ConPlex)— to calculate a recently suggested construction-based complexity measure by Nelson (2024) and 2) to investigate the applicability of the tool to various types of texts in multiple languages, namely English, Czech, German, and Italian. The paper does not only describe the development of the tool in detail, along with its validation by comparing complexity scores output by the tool with some of Nelson's (2024) results but also highlights what researchers should be aware of when using the tool in their research.

The paper is organised as follows. Section 2 discusses the notion of 'complexity' in linguistics. Section 3 offers information on the corpora used to validate the accuracy of ConPlex and describes how the tool has been produced. Section 4 deals with the tool validation results as well as its potential uses and limitations. Finally, Section 5 offers some concluding remarks.

2. Construction-based complexity

There has been some recent debate about the types of measure that should be used to represent the construct of 'complexity' in linguistics or second language acquisition (SLA), and there has been no consensus about a common measure so far (Ehret *et al.* 2023: 2). In their theoretical and methodological overview, Bulté *et al.* (2024) assert that there should be a more restricted interpretation of complexity and suggest a set of core measures that should be used to increase replicability and knowledge accumulation. They put forward a list of eight core measures of complexity which include moving-average type-token ratio (MATTR) for several indices, and various ratio-based measures at the word, phrase, clause, T-unit, and AS-unit level. Bulté *et al.*'s (2024) manuscript has provoked several 'open peer commentary' responses in the same journal. For instance, the response by Biber *et al.* (2024: 1–2) points out that the 'omnibus' measures suggested by Bulté *et al.* (2024) disregard the syntactic functions of grammatical structures. They liken this to a biologist taking a reductionist method and operationalising the complexity of forests by simply calculating the average height of trees and the mean number of branches per tree. Biber *et al.*'s comment seems a valid point, as fine-grained measures can reveal intricate details about the grammatical

[1] https://github.com/kristopherkyle

complexity of a language. However, there is also a place for omnibus measures that represent the interaction of multiple features. If omnibus methods that match a theoretical construct in the field are selected, they may have the benefit of being applied to a wider range of texts. Omnibus measures could potentially be applied to multiple registers or languages without the need to create a taxonomy of grammatical possibilities in the target domain. The most appropriate measures might not necessarily be the ones suggested by Bulté *et al.* (2024). Appropriate measures can be selected to match the theoretical beliefs of the researcher and to answer specific research questions. There has also been disagreement regarding the use of the sentence as a syntactic unit. On the one hand, Bulté *et al.* (2024) claim that there is no linguistic definition of the sentence that is agreed upon and suggest that it is an unusable unit for oral data or for analysing texts produced by writers whose punctuation skills are limited. On the other hand, Lu (2024) points out that the sentence as a unit is intuitively useful in writing.

Nelson (2024) takes an alternative approach to the measuring of complexity that is grounded in Complexity Theory and Construction Grammar. In some ways, Complexity Theory is also in line with the abovementioned measures, as "the behaviour of complex systems emerges from the interactions of its components" (Larsen-Freeman 1997: 143), and it is not concentrated on a specific component. Rather, complexity theorists are interested in how the parts of a complex system interact (Larsen-Freeman 1997), not merely in the production of a vast taxonomy of individual factors (Larsen-Freeman and Cameron 2008: 206). More recently, Larsen-Freeman (2017) has described Complexity Theory as a metatheory which also requires a theory of language and how it develops. One of these metatheories is Construction Grammar, in which Goldberg (2003: 219) defines constructions as "stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns" and further argues that

> any linguistic pattern is recognised as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognised to exist.

Construction Grammar differs from other grammar descriptions in that it aims to account for the whole of the language. However, no finite typology of all of the possible constructions in the English language has been agreed upon. Therefore, Nelson's (2024: 13) measure seeks to account for "how the diversity of constructions used impacts the

statistical properties of the texts a person produces." The measure is calculated as shown below:

$$C(d) = \frac{1}{N}\sum_{i=1}^{N} D(S_i)\, P(S_i)$$

The complexity of the document $C(d)$ is calculated as the mean diversity $D(S_i)$ and the mean productivity $P(S_i)$ of all of the sentences in the document. First the text is part-of-speech (POS) tagged, then the diversity of each sentence is calculated by partitioning the tags in the sentence into lists of tag pairs. A list of tags is taken and partitioned into pairs at an overlap of one, meaning that the last tag in the pair (T1, T2) is the first in the succeeding pair (T2, T3). For example, the sentence from Mary Shelley's Frankenstein would be converted into tag pairs as shown in (1):

(1) **Sentence**: "There is something at work in my soul, which I do not understand."
Tagged sentence: There_EX is_VBZ something_NN at_IN work_NN in_IN my_PRP$ soul_NN which_WDT I_PRP do_VBP not_RB understand_VB
Tag pairs: ('EX', 'VBZ'), ('VBZ', 'NN'), ('NN', 'IN'), ('IN', 'NN'), ('NN', 'IN'), ('IN', 'PRP$'), ('PRP$', 'NN'), ('NN', ' WDT '), ('WDT', 'PRP'), ('PRP', 'VBP'), ('VBP', 'RB'), ('RB', 'VB')

Next, the 'Shannon entropy' (Shannon 1948) of tag pairs is calculated, the mean of which contributes to the complexity score for the text. The productivity of each sentence is calculated as the entropy of word tag pairings minus the entropy of tags. 1 is added to the productivity calculation to account for situations when the entropy of word tag pairings is 0 or less than 1. The sentence in the Frankenstein example above is comprised of 13 unique pairings of a tag and a word but only ten tags. This is because there are three nouns (*something*, *work*, and *soul*) tagged as 'NN' and two prepositions (*at* and *in*) tagged as 'IN'. This information is used in the productivity calculation. Given pairs of words and their tags (e.g., pairs = {{there, EX}, {is, VBZ}, … {understand, VB}}) which can be represented as two ordered lists (i.e., tags = {EX, VBZ, … VB} and words = {there, is, …understand}), productivity is calculated as the conditional entropy of the words given the tags or H(words | tags) = H(tags, words) - H(tags). The complexity of the sentence is calculated by multiplying diversity and productivity, as they are held to interact. The complexity of a document is taken as the mean of the complexity of all the sentence-level complexity scores in the document.

Nelson (2024) clearly shows how measuring diversity in this way can represent the complexity of constructions using POS graphs. In addition, the sentence in (2), taken from F. Scott Fitzgerald's *The Great Gatsby*, is used as an example to illustrate the need for the productivity element of the measure.

(2) The apartment was on the top floor - a small living-room, a small dining-room, a small bedroom, and a bath.

If the first two occurrences of *small* in the sentence were replaced by *cosy* and *spacious*, the productivity score for the sentence would increase due to the wider range of word tag pairings. This, in turn, would increase the complexity score. Objectively, the sentence containing a wider variety of adjectives would likely be considered more complex by most if not all readers. Calculating complexity in this way is in line with one of Larsen-Freeman and Cameron's (2008: 206) methodological principles to identify collective variables that are characteristic of multiple elements interacting within a system. Although only one complexity score is output for each text, the score represents the interaction of components within the system, as opposed to a vast taxonomy of individual scores that represent individual components in the system. In this sense, the measure could be said to be more in line with Complexity Theory.

In addition to proposing the measure, Nelson (2024) also applies it to several datasets and shows that complexity scores increase 0.015 per month with L1 acquisition data from children (MacWhinney 2000). Furthermore, results from a Bayesian hierarchical model show that an increase in complexity measure scores correlates with the proficiency level of L2 learners in data taken from the *International Corpus Network of Asian Learners of English* (ICNALE; Ishikawa 2023). The measure also shows strong correlations with traditional readability measures. Moreover, when comparing U.S. presidential campaign speeches from 2016, results from a mixed effects model suggest that complexity is not affected by text length.

Although the theory behind Nelson's (2024) construction-based complexity measure has been summarised here, it is highly recommended that readers interested in using the ConPlex engage with Nelson's paper, where a more detailed theoretical background is provided.

3. TOOL CREATION

This section introduces the corpora used to validate the accuracy of ConPlex and describes the technical steps taken to produce the tool.

3.1. Corpora used to validate the tool

To assess the accuracy of the tool output, the complexity scores from the spoken monologues and written essays in ICNALE were compared with Nelson's (2024) results, which were obtained after contacting the researcher. ICNALE includes 4,400 spoken monologues and 5,600 written essays produced by university students in ten countries across Asia. As such, it is one of the largest publicly available learner corpora and includes texts at the A2, B1_1, B1_2, and B2+ CEFR levels. The corpus also includes data produced by native speakers of English who completed the same spoken and written tasks as the L2 users. The transcripts of the monologues and written essays were used in the analysis and no further pre-processing was done to the texts.

To evaluate the tool further, complexity scores were calculated for a new dataset, the *CEFR English Listening Corpus*. This corpus was compiled by the author from listening texts that are freely available for language study online. The first source of texts includes two British Council websites[2] that feature listening texts and videos that have been produced for the website, and *YouTube* videos that are not produced by the British Council. Each text has been assigned a CEFR level by the producers of the website. In some cases, the CEFR level is broad, spanning several levels, such as B1-B2 and B2-C1-C2. In these cases, the lowest CEFR level was counted.

In order to increase the size of the corpus, listening texts from the CEFR-aligned Cambridge exams,[3] which are available online for exam preparation, were added. Although these texts have a different purpose, they were selected for the corpus to include a range of texts aimed at L2 learners of English for practicing and assessing their own listening. The final corpus size was 728 texts and 345,104 words, as can be noticed in Table 1, where more detailed information about the number and length of texts from each source is provided.

[2] https://learnenglish.britishcouncil.org/ and https://learnenglishteens.britishcouncil.org/
[3] https://www.cambridgeenglish.org/learning-english/exam-preparation/

Text source	Texts	Tokens	Text length (tokens)			
			M	*SD*	Min	Max
British Council	563	303,959	540	560	54	4,751
Cambridge Exams	165	41,145	249	200	47	851
Total	728	345,104	474	516	47	4,751

Table 1: Size of the *CEFR English Listening Corpus*

The distribution of texts by CEFR level and text type is shown in Table 2. A limitation of the *CEFR English Listening Corpus* is how imbalanced the dataset is. In particular, only six texts in the corpus are classified as C2 level, all of which were Cambridge listening texts. As the purpose of this part of the study was to investigate whether complexity scores increase across CEFR levels, the C2 level was still included as a separate class, instead of merging it with the C1 level.

Text type	A1	A2	B1	B2	C1	C2	Total
BC Listening	24	41	34	52	22	0	173
BC Videos	15	0	104	15	9	0	143
BC *YouTube*	0	2	34	157	54	0	247
Cambridge	18	32	38	53	18	6	165
Total	57	75	210	277	103	6	728

Table 2: Text types by CEFR level in the *CEFR English Listening Corpus*

In both cases the transcripts produced by the material creators were used. The British Council texts each had a transcript available online, which was presumably produced or at least checked by the creators of the website. For the Cambridge listening exams, transcripts were included in PDF files that were available with the audio files. The *CEFR English Listening Corpus* has not been released due to copyright.

While the desktop version of ConPlex currently supports only English texts, the underlying code is available as a *Python* notebook on GitHub.[4] This enables researchers to adapt the code to investigate complexity in other languages. In the current study, the multilingual *MERLIN* corpus (Boyd *et al.* 2014) was used to evaluate the application of the complexity measure to three different languages. The *MERLIN* corpus features texts written by learners of Czech, German, and Italian that were taken from CEFR-aligned written examinations. CEFR levels containing fewer than ten texts per language were not included in the analysis. The final corpus size is described in terms of texts in Table 3, and in terms of tokens in Table 4. As part of the *MERLIN* corpus project, the CEFR level of each text was re-rated by specially trained testers in what was termed as a 'fair

[4] https://github.com/cooperchris17/ConPlex (accessed 10 March 2025).

rating'. This rating was used in the current study to represent the CEFR level. The plain text versions of the texts were used and no further pre-processing steps were taken.

Language	A1	A2	A2+	B1	B1+	B2	B2+	C1	Total
Czech		76	112	90	75	72			425
German	57	199	107	217	115	219	73	42	1,029
Italian	29	289	92	341	53				804

Table 3: Number of texts per CEFR level in the *MERLIN* corpus used in the current study

Language	Texts	Tokens	Text length (tokens)			
			M	***SD***	**Min**	**Max**
Czech	425	61,013	61,013	61,013	61,013	61,013
German	1,029	126,468	126,468	126,468	126,468	126,468
Italian	804	93,292	93,292	93,292	93,292	93,292
Total	2,258	280,773	474	280,773	280,773	280,773

Table 4: Size of the *MERLIN* corpus used in the current study

It should be noted that a complexity measure that has been designed and validated on the English language will not necessarily behave in the same way when used with other languages. Previous research has shown that there are several differences between the three languages in the *MERLIN* corpus when compared to English. For example, Kettunen (2014) measured the complexity of the European Union constitution written in 21 languages using a morphological complexity measure and two type-token ratio (TTR) metrics. Of the four languages considered in the current study, English and Italian were the least complex for all the measures, German was the most morphologically complex, and Czech had the highest TTR scores. It has also been pointed out that Czech has an overt inflectional morphology, whereas the morphological features of English are predominantly analytic (Hledíková and Ševčíková 2024). Also, when compared with German, the distance between form and meaning is often greater in English (Hawkins 2015). Due to these and other differences, complexity scores will not necessarily be comparable between languages. However, as the grammatical constraints of an individual language impact texts written or spoken in that language in a similar way, the measure should be suitable for at least an initial exploratory investigation of languages other than English.

3.2 Producing the tool

The tool was designed to be accessible to as many researchers as possible. While the code to calculate complexity in Nelson's (2024) paper was written in *Mathematica*, *Python* was chosen for ConPlex. *Python* is one of the most widely used programming languages: it is highly readable and often used for natural language processing tasks. This makes it a suitable choice, as researchers may wish to adapt the code that is released with ConPlex. In addition, while Nelson (2024) worked with POS tagged texts that had been processed before the complexity analysis, ConPlex was designed to accept plain text files and handle POS tagging within the tool. This is simpler for the end user, as they do not need to use a separate tool to tag their texts. Furthermore, it avoids the problem of dealing with output from different taggers, which are likely to follow inconsistent formatting standards. *Stanza* (Qi *et al.* 2020) is used in ConPlex to split the texts into sentences and to tag the texts with treebank-specific POS tags. Next, tokens tagged with the universal POS tag 'PUNCT', meaning all punctuation, are excluded from the analysis. Then, all words are converted to lower case to avoid words that are used more than once in the same sentence being counted as different words when they occur at the beginning of a sentence. *Stanza* was initially chosen to replicate the fact that Nelson (2024) used the *Stanford Tagger* in his analysis.[5] The NLP library *SpaCy* (Honnibal *et al.* 2023) was also trialled, but inspection of the output showed that *Stanza* was more accurate for sentence tokenisation. Diversity, production, and complexity were calculated as described in Section 2 and the entropy function in *SciPy* (Virtanen *et al.* 2020) was used in the calculations. After trialling the code in a *Python* notebook, a downloadable app was created using *PyQt5*.[6] Producing a downloadable app allows researchers who are not familiar with *Python* to use the tool with a graphical user interface.[7] While it is assumed that most researchers will use the downloadable app for the analysis of English texts, a notebook with the *Python* code behind the tool is also available on GitHub. Widgets have been added to the notebook so the functionality is the same as the downloadable tool. Sharing the *Python* code in this way makes it possible for researchers to adapt the code to suit their research goals. Reasons for

[5] https://techfinder.stanford.edu/technology/stanford-part-speech-tagger
[6] https://www.riverbankcomputing.com/software/pyqt/
[7] The *Windows* version of the app is available at https://drive.google.com/file/d/1PljHorFOaXYTIar527GMaDibNAzk6Coo/view (accessed 10 March 2025) and links to the *OSX* and *Linux* versions will be added to the app's GitHub page at https://github.com/cooperchris17/ConPlex (accessed 10 March 2025) when they are ready.

adapting the code might include trying the complexity measure with other languages, as demonstrated in the current study, or changing the tagger to one that is more appropriate for the users' texts. The algorithm used in ConPlex is illustrated in Figure 1.

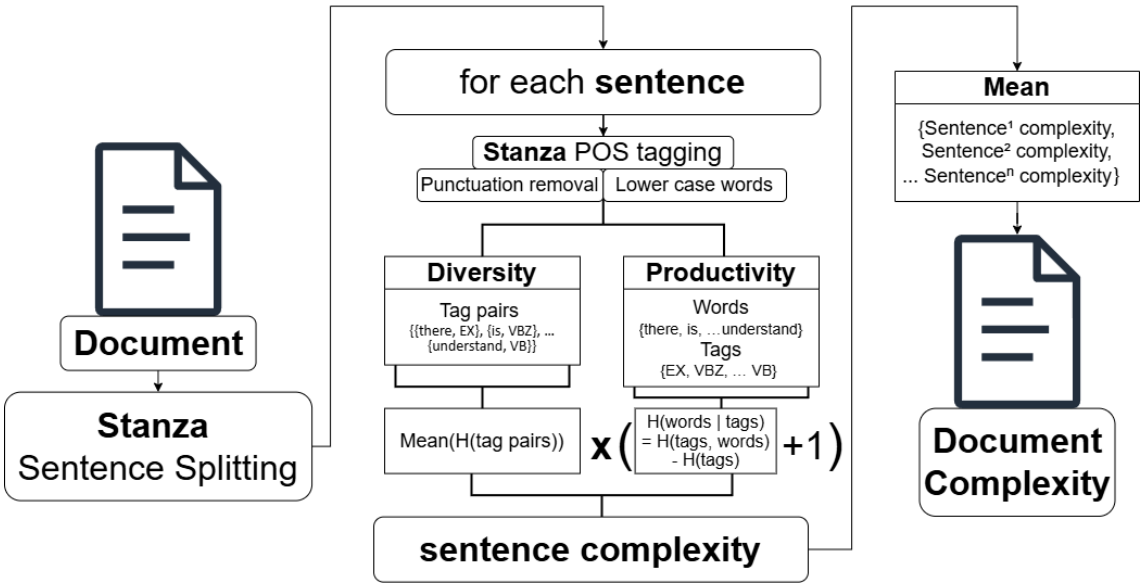

Figure 1: Visualisation of the algorithm used to calculate complexity in ConPlex

4. THE FINISHED TOOL

A screenshot of the downloadable app is shown in Figure 2. The tool has two input methods. The first option is to copy and paste one text into the textbox in the tool's interface, then click 'Process Text Input' to begin processing. For uploading one or multiple plain text files, the user can click 'Upload and Process Files'. For this option, processing begins as soon as the files have been selected. The tool outputs the mean complexity score for each text, along with the mean diversity and mean productivity scores. It is expected that most researchers will only use the complexity scores.

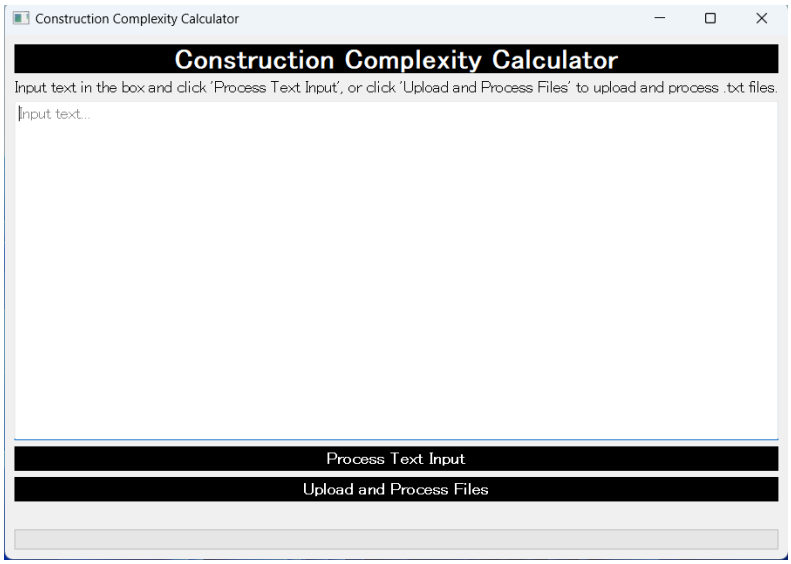

Figure 2: Screenshot of ConPlex in Windows

4.1. Tool validation results

The complexity scores calculated for the ICNALE texts are shown in Figure 3. The first point to note is that there is a progression across CEFR levels in both the present study and Nelson (2024). This is clearly evident in the central tendencies indicated by the boxplots and distributions shown in the violin plots. The progression cannot be described as linear, but this is in line with Complexity Theory, as "learning is not climbing a developmental ladder; it is not unidirectional." (Larsen-Freeman 2017: 27). The two additional points to note about Figure 3 are 1) the difference between scores in the current study and Nelson (2024) and 2) the substantial number of outliers.

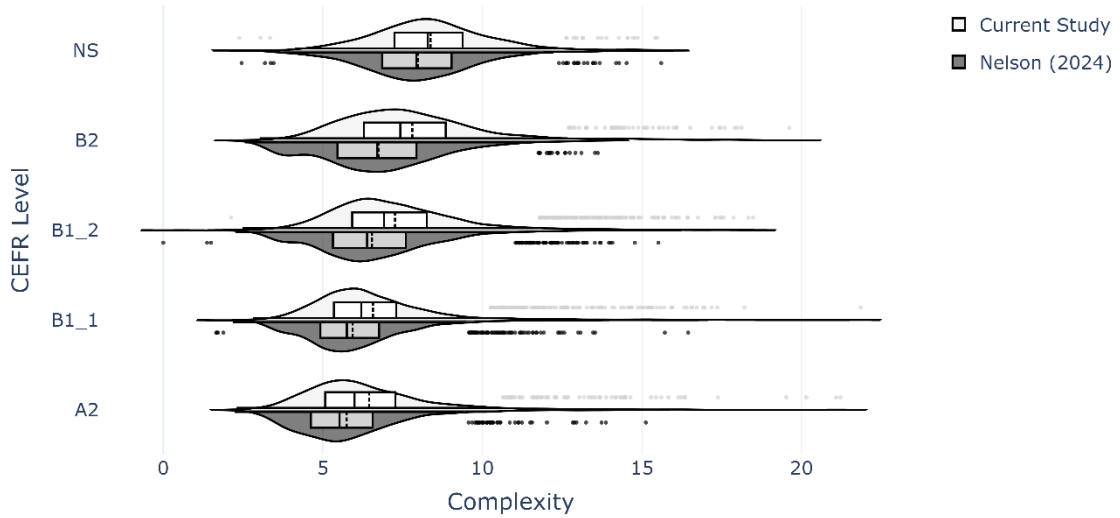

Figure 3: ICNALE complexity score distribution in the present study and Nelson (2024)

For a more fine-grained analysis, the difference between the scores for each text was calculated. Descriptive statistics are shown in Table 5 and the distribution of differences in complexity scores is illustrated in Figure 4. Positive values indicate higher scores in the current study and negative values indicate higher scores in Nelson (2024). In the current study, productivity scores are generally similar but slightly higher, and the diversity scores from Nelson (2024) are generally higher. The complexity scores are also slightly higher for the most part here. Some of the differences are extreme, with the maximum difference being 17.77. Despite these differences, Figure 4 illustrates that the differences for the majority of the texts are close to zero.

Measure	M	SD	Min	Q1	Mdn	Q3	Max
Complexity	0.70	1.99	-5.10	0.09	0.28	0.59	17.77
Diversity	-0.03	0.54	-2.14	-0.16	-0.09	-0.04	3.24
Productivity	0.14	0.27	-0.66	0.03	0.08	0.15	2.31

Table 5: The difference between individual texts in the current study and Nelson (2024)

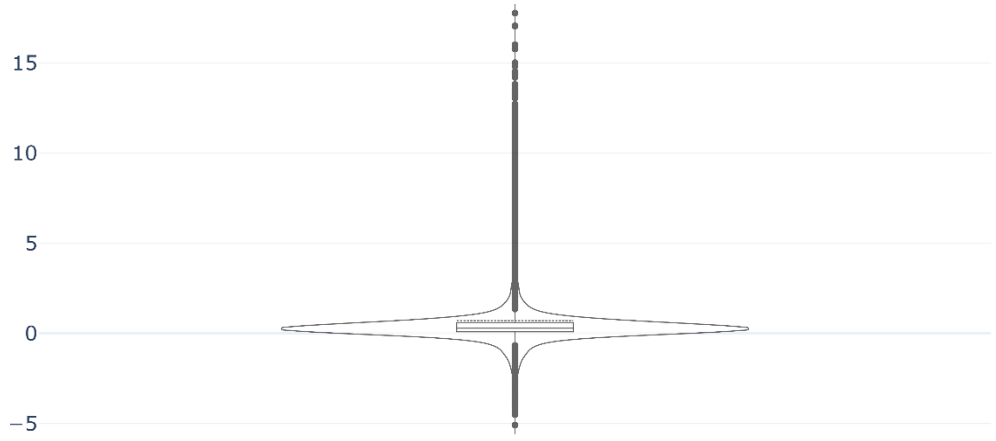

Figure 4: Difference in ICNALE complexity scores in the current study and Nelson (2024)

The reason for the difference in complexity scores was investigated by consulting texts that had large differences in the two studies. It was found that the differences seemed to be largely caused by sentence tokenisation. In the current study, texts were split into sentences using *Stanza*. However, Nelson (2024) used the period POS tag in the *Stanford Tagger* to split sentences at the following punctuation marks: ".", "?", and "!". For example, the sentence "3. must fast show to arrive at the attention in serve." was split into one sentence in this study, but in two in Nelson (2024). In addition, Nelson (2024) attempted to deal with learner texts that did not include accurate sentence

punctuation by splitting texts that featured less than two period POS tags into sentences every ten words using a bespoke function labelled ‘safeBreaks’. When an equivalent function to safeBreaks was added to the ConPlex code, the mean difference in scores reduced slightly to 0.51 (*SD* = 1.42). However, there were still extreme differences (*max* = 13.25) and the median difference was the same (*Mdn* = 0.28). There are some other small differences between the two implementations. The first is the handling of sentence internal punctuation: while ConPlex removes all punctuation based on the universal POS tag ‘PUNCT’ assigned by *Stanza*, Nelson’s (2024) code only seems to remove commas. In addition, there may be differences in the way that entropy is calculated in *Python* when compared with *Mathematica*.

Despite these differences, no changes were made to ConPlex. The motivation for this is that the method of sentence tokenisation in this study seems to represent sentences more accurately than splitting by any occurrence of a period POS tag. In addition, a similar function to safeBreaks was not added, as this is the kind of methodological decision that should be made by individual researchers at the corpus pre-processing stage to match the research questions of the project. The inability to replicate Nelson’s (2024) results precisely is a limitation of the current study. To somewhat overcome this limitation, the *Python* code used in ConPlex is shared on GitHub so that interested researchers can compare it with the *Mathematica* code shared in the supplementary information in Nelson’s (2024) paper.

The results from the *CEFR English Listening Corpus* are shown in Figure 5. The plots represent the distribution of scores throughout the sample. The median is indicated by the solid lines in the boxplots and the mean is indicated by the dotted lines. An incremental increase in complexity scores is evident by examining the boxplots. However, the distributions that are visualised by the violin plots reveal a distinct increase between the B1 and B2 level, and to a lesser extent between the A1 and A2 levels. There are also fewer outliers when compared with the ICNALE data. This could be related to the difference in corpus size, but it might also be related to the more consistent nature of sentence punctuation that exists in transcripts that have been prepared by educational professionals to support learning from listening texts, when compared with written and spoken texts that have been produced by L2 users of the language. Although no previous studies have assessed the complexity of CEFR-aligned listening texts, the results of the current study are somewhat similar to previous research

that investigated the features that discriminated between sentences that were rated from A1 to C2 level. Uchida *et al.* (2024) found that sentences at A and B levels showed lexical and syntactic variation, whereas B- and C-level texts could only be distinguished by lexical aspects. There is greater difference between the A1 and B2 levels in the *CEFR English Listening Corpus*, suggesting that the complexity of constructions is also more variable at the lower CEFR levels.

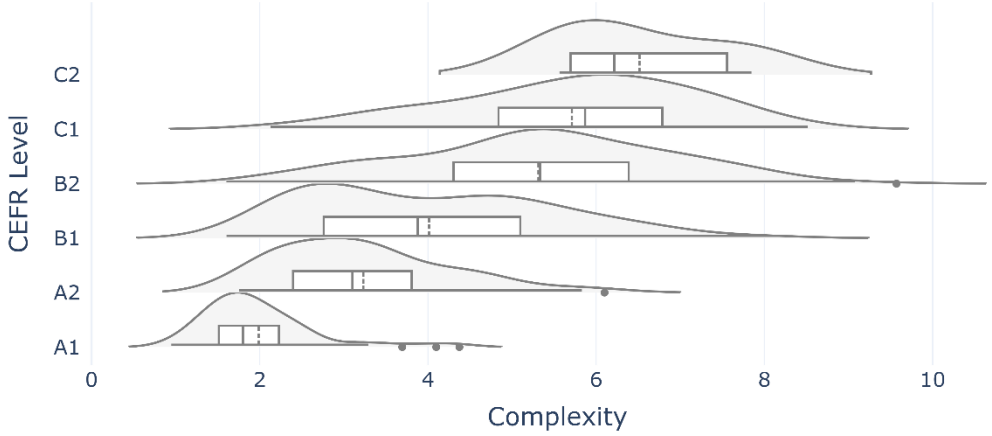

Figure 5: Distribution of complexity scores in the *CEFR English Listening Corpus*

The results from the *MERLIN* corpus are visualised in Figures 6, 7, and 8 for Czech, German, and Italian, respectively. The general trend for German and Italian texts is an increase across CEFR levels. These findings are in line with previous research into L2 German complexity (Weiss and Meurers 2019) that demonstrated accurate text classification from the A2 to B2 level with a selection of 150 complexity features. Classification accuracy was much lower for the A1 and C1/C2 levels, but the general trend aligns with the current study's results, that complexity increases with CEFR level. In L2 Italian, morphological complexity has been shown to be able to distinguish between low- and high-level proficiency learners between the A2 and B2 level (Brezina and Pallotti 2019). However, it was not able to distinguish between the B1 and C2 CEFR levels (Spina 2025). While the current study showed a clear progression, particularly from the A2 to B1+ level, it is not clear whether construction complexity also levels off at the independent to proficient level. Future research could investigate this further with L2 Italian corpora for higher proficiency learners.

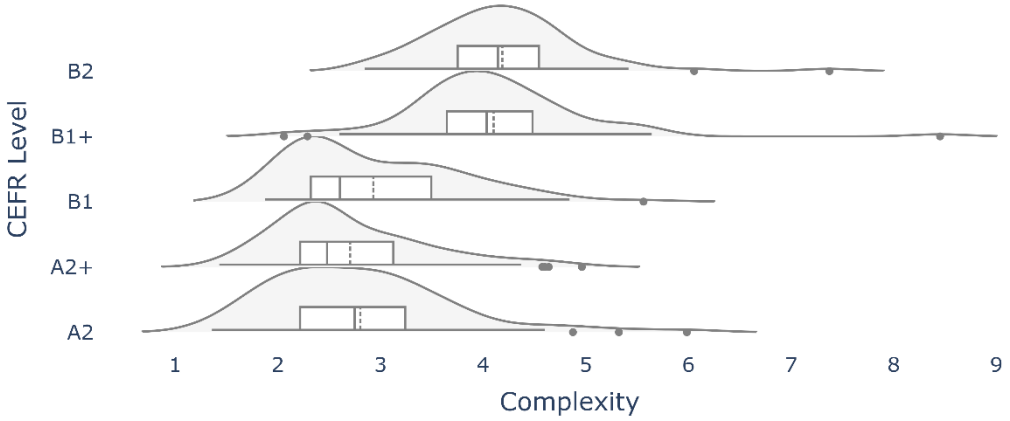

Figure 6: Distribution of complexity scores in the *MERLIN* corpus: Czech

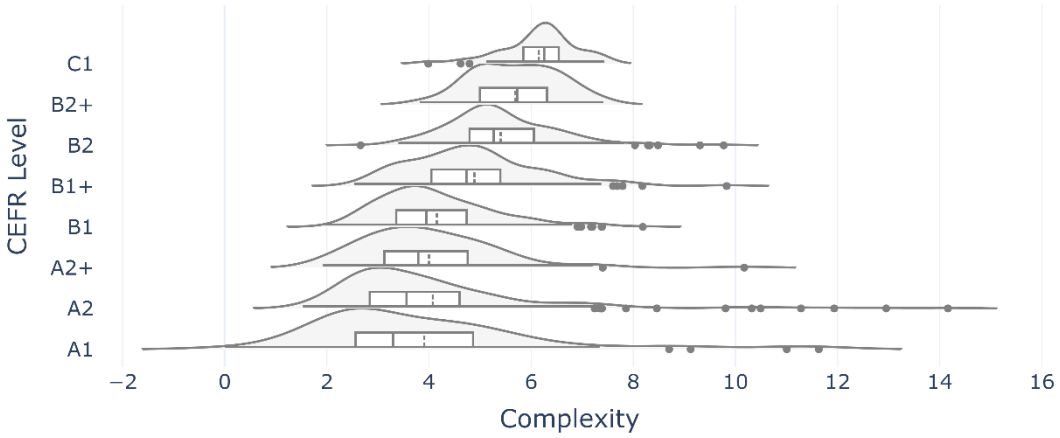

Figure 7: Distribution of complexity scores in the *MERLIN* corpus: German

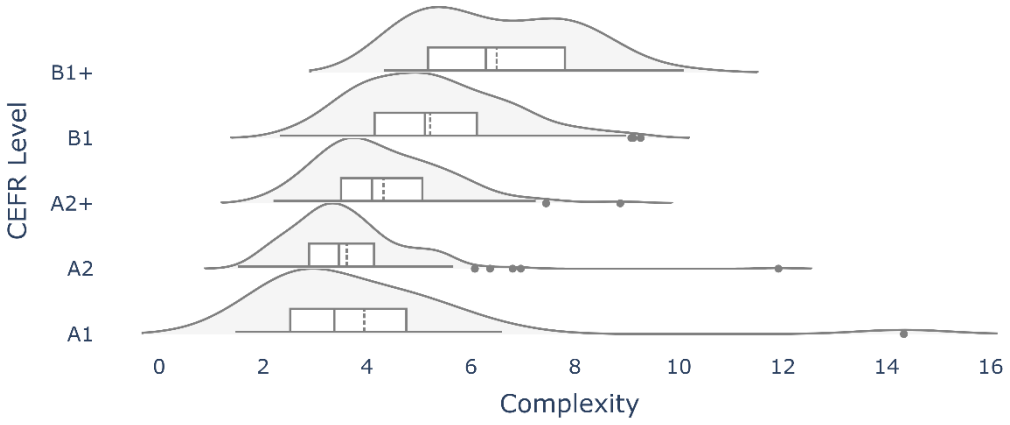

Figure 8: Distribution of complexity scores in the *MERLIN* corpus: Italian

On the other hand, for Czech texts, there seems to be a clear split between complexity scores at the lower levels (A2, A2+, B1) and the higher levels (B1+, B2), whereas the variation within these two groups is more limited. These results somewhat align with previous L2 Czech research (Nogolová *et al.* 2023) that showed a tendency for sentence length, clause length, and number of clauses per sentence to increase from the A1 to B2 level. From the C1 level there was little or no increase in the metrics. Although the researchers pointed out that clause length does not always indicate an increase in syntactic complexity, they argued that clause and sentence length are likely to somewhat correspond with syntactic knowledge. The reason for the difference in CEFR level increase thresholds could be related to the difference in the operationalisation of complexity, or the corpora used in the study. Future research might compare the different complexity measures on the same corpora for further insight into their alignment, or lack thereof.

Overall, the results of the additional datasets analysed in the current study provide further evidence that Nelson's (2024) complexity measure is able to reveal patterns in listening texts in line with the developmental level of the L2 users that they are aimed at. Furthermore, a similar pattern is evident in written texts produced by learners of Czech, German, and Italian.

4.2. Potential uses and limitations

ConPlex has several potential uses, as Nelson's (2024) complexity measure is designed to measure the developmental complexity of language in general, meaning that it is not limited to SLA. It could be used to investigate the complexity of production by first language (L1) and L2 users across developmental levels such as age or CEFR level, as demonstrated by Nelson (2024). In addition, further research could investigate the nature of how construction-based complexity increases across texts that have been produced for L2 reading or listening, as was partially demonstrated in the current study. If a larger corpus was used, a benchmark for each CEFR level could be suggested to measure the complexity of constructions in individual texts. This could provide useful guidelines about the tendencies of text complexity that could be useful for educators and language learners when selecting appropriate texts. The current study also showed that the trend for an increase in construction complexity across CEFR levels extends to languages beyond English. So far, only Czech, German, and Italian have been

considered, but future research could be extended to any of the 70 human languages, at the time of writing, that are supported with pretrained neural models in *Stanza*. It would be worth investigating the constructional complexity of languages in relation to known differences between the languages, such as morphological complexity or word order freedom. In Nelson's (2024) study, the complexity measure was also applied to political speeches. With this in mind, the measure might be of interest to digital humanities researchers if they wish to compare the complexity of constructions used by particular authors, orators, or other language users.

Recently, complexity measures are often integrated into methodologies that aim to assess the readability of texts (Crossley *et al.* 2023), L2 learner writing (Lu 2017), and within the complexity, accuracy, fluency, and lexis framework to measure L2 language performance (Skehan 2009). They can also support the evaluation of interlanguage development over time and provide support in answering other fundamental questions in SLA (Bulté *et al.* 2024). ConPlex could be used to incorporate construction-based complexity into these and other frameworks in the fields of SLA, natural language processing, and beyond.

The main limitation of the tool is its sensitivity to sentence boundaries. How to pre-process texts into sentences is an important methodological consideration that must be made by researchers before using ConPlex. In particular, how spoken texts should be segmented into sentences to represent complexity across utterances is something that should be considered further. Although Nelson's (2024: 23) research showed that the contribution of mode to complexity was small when considering the spoken and written texts in ICNALE, it could be the case that spoken texts have a different complexity threshold to written texts when other corpora are considered. It is possible that the tool could potentially be biased towards measuring complexity in written texts due to its use of the sentence as the unit of measurement. However, these suggestions need to be further investigated in empirical research. Another limitation is the one-dimensional nature of the output of the tool. Depending on the tool's uptake in the research community, further features could be added. For example, complexity scores could be output at the sentence level to allow for more fine-grained analysis and the investigation of complexity across texts. In addition, the tagged sentences and tag pairs for each sentence could be output or visualised, so researchers can gain more insight into the kinds of constructions that are being used across sentences and texts.

5. Conclusion

The current study has fulfilled its aim of producing a tool that adequately replicates Nelson's (2024) construction-based complexity measure. Although there were differences in the ICNALE complexity scores between both studies, the way that sentences were operationalised here, using *Stanza*, allows for more accurate calculation of the measure. The creation and release of ConPlex will allow more researchers to experiment with using this complexity measure to answer a range of research questions. The release of the code along with the tool allows for further modifications to make the tool applicable to other languages, as was demonstrated in the current study with Czech, German, and Italian, and texts that require different taggers. It is hoped that the research community will embrace the tool, adding another dimension to the complexity measure debate.

References

Biber, Douglas, Bethany Gray, Tove Larsson and Shelley Staples. 2024. Grammatical analysis is required to describe grammatical (and "syntactic") complexity: A commentary on "complexity and difficulty in second language acquisition: A theoretical and methodological overview." *Language Learning.* https://doi.org/10.1111/lang.12683

Boyd, Adriane, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik: European Language Resources Association, 1281–1288.

Brezina, Vaclav and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second Language Research* 35/1: 99–119.

Bulté, Bram, Alex Housen and Gabriele Pallotti. 2024. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*. https://doi.org/10.1111/lang.12669

Crossley, Scott, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods* 55/2: 491–507.

Ehret, Katharina, Aleksandrs Berdicevskis, Christian Bentz and Alice Blumenthal-Dramé. 2023. Measuring language complexity: Challenges and opportunities. *Linguistics Vanguard* 9/s1: 1–8.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7/5: 219–224.

Hawkins, John A. 2015. *A Comparative Typology of English and German: Unifying the Contrasts*. Abingdon: Routledge.

Hledíková, Hana and Magda Ševčíková. 2024. Conversion in languages with different morphological structures: A semantic comparison of English and Czech. *Morphology* 34/1: 73–102.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, Adriane Boyd and Henning Peters. 2023. Explosion/spaCy: v3.7.2: Fixes for APIs and requirements. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Ishikawa, Shin'ichiro. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Abingdon: Routledge.

Kettunen, Kimmo. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21/3: 223–245.

Kyle, Kristopher. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-based Indices of Syntactic Sophistication*. Atlanta, GA: Georgia State University dissertation.

Kyle, Kristopher, Scott A. Crossley and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly* 18/2: 154–170.

Larsen-Freeman, Diane. 1997. Chaos/complexity science and second language acquisition. *Applied Linguistics* 18/2: 141–165.

Larsen-Freeman, Diane. 2017. Complexity theory: The lessons continue. In Lourdes Ortega and ZhaoHong Han eds. *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*. Amsterdam: John Benjamins, 11–50.

Larsen-Freeman, Diane and Lynne Cameron. 2008. Research methodology on language development from a complex systems perspective. *The Modern Language Journal* 92/2: 200–213.

Lu, Xiaofei. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing* 34/4: 493–511.

Lu, Xiaofei. 2024. Towards greater conceptual clarity in complexity and difficulty: A commentary on "complexity and difficulty in second language acquisition: A theoretical and methodological overview." *Language Learning* https://doi.org/10.1111/lang.12688

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum Associates.

McNamara, Danielle S., Arthur C. Graesser, Philip M. McCarthy and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.

Mizumoto, Atsushi. 2024. Developing and disseminating data analysis tools for open science. In Luke Plonsky ed. *Open Science in Applied Linguistics*. Online: Applied Linguistics Press, 123–131. https://www.appliedlinguisticspress.org/home/catalog/plonsky_2024

Nelson, Robert. 2024. Using constructions to measure developmental language complexity. *Cognitive Linguistics* 35/4: 481–511.

Nogolová, Michaela, Radek Čech, Michaela Hanušková and Miroslav Kubát. 2023. The development of sentence and clause lengths in Czech L2 texts. *Korpus - gramatika - axiologie* 28: 22–37.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In Asli Celikyilmaz and Tsung-Hsien Wen eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System*

Demonstrations. Online: Association for Computational Linguistics, 101–108. https://aclanthology.org/2020.acl-demos.14.pdf

Shannon, Claude, E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27/3: 379–423.

Skehan, Peter. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30/4: 510–532.

Spina, Stefania. 2025. Complexity and accuracy of verbal morphology in written L2 Italian: The role of proficiency and contingency. *International Journal of Learner Corpus Research* 11/1: 114–144.

Uchida, Satoru, Yuki Arase and Tomoyuki Kajiwara. 2024. Profiling English sentences based on CEFR levels. *ITL - International Journal of Applied Linguistics* 175/1: 103–126.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, and Evgeni Burovski. 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods* 17/3: 261–272.

Weiss, Zarah and Detmar Meurers. 2019. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In Andrea Abel, Aivars Glaznieks, Verena Lyding and Lionel Nicolas eds. *Widening the Scope of Learner Corpus Research: Selected Papers from the Fourth Learner Corpus Research Conference*. Louvain-la-Neuve: Presses Universitaires de Louvain, 419–435.

Corresponding author
Christopher R. Cooper
Rikkyo University
Center for Foreign Language Education and Research
Nishi Ikebukuro 3-34-1, Toshima-ku
Tokyo
171–8501
Japan
Email: cooper@rikkyo.ac.jp

received: January 2025
accepted: March 2025