# Automatic Analysis of Noun Phrase Complexity in Learner spoken Productions

**Saint Joannes Monengue Essengo, Bi Yoman Ghislain Aymard Toa, Thomas Gaillat, Christine Evain, Fabienne Moreau**
Université Rennes 2, LIDILE
{saint-joannes.monengue-essengo, bi.yoman.ghislain.toa}@etudiant.univ-rennes2.fr
{thomas.gaillat, christine.evain, fabienne.moreau}@univ-rennes2.fr

## Abstract

Most tools for measuring syntactic complexity were built for written English, and learner speech, like comparisons across languages, has received far less attention. We present an open pipeline, anchored in Universal Dependencies (UD), that extracts nominal constructions from transcribed speech and applies a single set of metric definitions to every language it processes. We test it on the Corpus InterLangue (CIL), a comparable spoken corpus of learner French and learner English, and on the spoken part of the International Corpus of English – Great Britain (ICE-GB) as a native baseline. As a result ,learner English shows more phrasal complexity than learner French, which leans on appositive and looser structures. Against native speakers the picture changes. Natives pack complexity into the noun phrase, while learners spread it across longer, flatter structures. Against manual gold standards, extraction reaches strict F1 scores between .79 and .95, depending on the corpus, and group differences are tested with Mann–Whitney U tests and Cliff's delta. The tool and its metric definitions are openly available.

## 1 Introduction

### 1.1 Motivation and research gap

Complexity research in second language acquisition (SLA) traditionally relied on global measures such as sentence length, subordination ratios and type–token ratios (Wolfe-Quintero et al., 1998; Norris and Ortega, 2009; Lu, 2010). Such measures are computationally undemanding but insufficiently discriminating. Two texts can share a mean clause length and still build their noun phrases in entirely different ways, and that is where learner and native production tend to diverge (Biber et al., 2011, 2020, 2025). The automatic toolkits that compute these indices, and finer-grained ones, were calibrated on written English (Lu, 2010; Kyle, 2016). This leaves two gaps open.

The first is speech. Speech is not writing with the disfluencies taken out; it packages information differently (Biber et al., 2011), and a metric tuned on academic prose will show different values when applied to a semi-guided interview. Even the unit of analysis is unstable, because transcripts lack the sentence boundaries of writing (Foster et al., 2000). One of our own learner transcripts, for instance, contains an unpunctuated unit of 817 words. Parsers add their share, since those trained on L1 English alone are less accurate on spoken learner English than those that have also seen learner data (Kyle et al., 2022). The second gap is comparability across languages. Asking whether learners of French and learners of English behave alike requires the same construct, measured the same way, on both sides; separate tools and post-hoc alignments do not deliver that. Multilingual UD-based profilers exist (Brunato et al., 2020), but they describe texts broadly, not noun phrases in particular.

This paper presents an analysis based on a tool built for that niche. It is an open-source pipeline[1] that extracts noun phrases from learner speech and, because it is anchored in Universal Dependencies (UD), measures their complexity comparably across languages.

### 1.2 Research questions

Both research questions are questions about the tool, which we test on the Corpus InterLangue (CIL) and on ICE-GB (Section 3).

RQ1. Does the tool accurately capture differences in noun phrase complexity between learners of French and learners of English? To answer this,

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]The tool is openly available on GitLab (https://gitlab.huma-num.fr/lidile/constructions-nominales/-/tree/v0.3.0).



*Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)*
pages 214–228, November 19–20, 2026, University of Tartu Library

we compare learner French and learner English from the CIL, focusing on whether the tool can detect structural differences in how complexity is distributed within noun phrases.

RQ2. Can the tool reliably distinguish between learner and native English in terms of noun phrase complexity? Here, we compare learner English from the CIL with native English from ICE-GB, assessing whether the tool can identify developmental differences in how complexity is packed into noun phrases.

## 2 Theoretical Background

### 2.1 The shift from global to phrasal complexity

Complexity entered SLA as one of three dimensions of performance, alongside accuracy and fluency (Housen and Kuiken, 2009), and for a long time it was measured globally (Norris and Ortega, 2009; Bulté and Housen, 2012). Length- and subordination-based ratios dominated developmental research (Wolfe-Quintero et al., 1998); tools such as the L2 Syntactic Complexity Analyzer automated fourteen of them (Lu, 2010, 2011). Such ratios trace broad developmental trends well, but they cannot locate the point, inside the sentence, where a learner's grammar departs from a native speaker's (Biber et al., 2020).

Norris and Ortega (2009) defended an organic view of CAF and showed that global measures do capture large-scale developmental variation. Bulté and Housen (2012) then took the construct apart, separating absolute from relative complexity, and syntactic complexity (clausal, phrasal, subclausal) from morphological complexity. More recently, Bulté et al. (2025) drew a sharp line between complexity, a structural property of linguistic components and their relations, and difficulty, the cognitive cost of those structures. We work on the structural side of that line.

An important move was brought forward by Biber and Gray. Global indices, they argued, misfire across registers (Biber et al., 2011; Biber and Gray, 2016), because complexity does not sit in the same place everywhere. It is clausal in conversation, whereas in informational writing it migrates into the noun phrase, where phrasal compression prevails over clausal elaboration (Biber and Gray, 2011). Learner writing follows suit. Less proficient writers of English for Academic Purposes (EAP) lean on attributive adjectives, more advanced ones on nouns as premodifiers and on prepositional phrases (Parkinson and Musgrave, 2014). Hence the current call to describe specific structures and their syntactic functions instead of relying on omnibus indices (Biber et al., 2020, 2025), a call that the present paper takes up.

The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) turned these ideas into software (Kyle, 2016; Kyle and Crossley, 2018). Its phrasal indices cross seven noun phrase types, defined by function (nominal subject, direct object, and so on), with ten dependent types (determiners, adjectival modifiers, prepositional phrases and relative clauses, among others), which gives 66 indices in all. On TOEFL essays they predicted writing quality better than global and fine-grained clausal indices (Kyle and Crossley, 2018), which made the phrase, and not only the clause, a unit worth measuring. Section 7.1 returns to TAASSC once our own measures have been introduced.

### 2.2 The noun phrase as the locus of complexity

The noun phrase has become a privileged site for observing complexity in SLA. Global measures average structural differences away; noun phrase measures keep them. Learners, for instance, may pack less into a noun phrase and spread the same content over appositions or separate clauses.

In spoken learner data, this often takes the form of successive reformulations, where each noun phrase remains short and new information is added in a separate, loosely attached phrase instead of being embedded within the head noun phrase. Examples (1a) and (2a) come from CIL interviews, (1a) from a French-speaking learner of English (CIL-EFL) and (2a) from an English-speaking learner of French (CIL-FFL); each is identified by its transcript code in the corresponding sub-corpus. Examples (1b) and (2b) are constructed counterparts.

(1) a. *it's like erh a sausage a big sausage* (CIL_L10_Fr_07_F_15, turn 007)
b. *it's like a big sausage*

(2) a. *une professeur au l'université une professeur de littérature anglophone* (CIL_ENG_JO_86_F_10)
b. *une professeure de littérature anglophone à l'université*

In (1a) and (2a), the speaker produces a first noun

phrase (*a sausage*, *une professeur au l'université*) and then elaborates it through an appositive reformulation (*a big sausage*, *une professeur de littérature anglophone*), so that the information is spread across several short, juxtaposed phrases. In (1b) and (2b), the same content is integrated into a single noun phrase through modification (*big*, *de littérature anglophone*) and embedding (*à l'université*). In Universal Dependencies terms, the first strategy surfaces as `appos` relations between flat phrases, the second as `amod`, `compound` or `nmod` dependents within a deeper phrase. Such reformulations also occur in native speech, so the contrast is one of degree rather than kind; it is precisely this difference in distribution that our metrics are designed to quantify.

Measuring such contrasts automatically requires a syntactic representation that holds across languages, and UD provides one (de Marneffe et al., 2021), although an imperfect one. Several non-canonical nominal phenomena still lack a standard treatment, in English as elsewhere (Schneider and Zeldes, 2021), and every parse-based measure, dependency distance included (Liu, 2008), inherits these gaps. This is why we treat extraction as a step to be evaluated in its own right, not as a by-product of parsing.

## 3 Data

### 3.1 The Corpus InterLangue (CIL) as a testbed

To evaluate the tool, we use the CIL,[2] a comparable spoken corpus of learner French and learner English developed by the LIDILE research unit at Université Rennes 2 (Gaillat et al., 2021). We chose it because both of its sub-corpora follow a single protocol, in which learners of French and learners of English performed the same tasks. The corpus includes 292 recordings (3,014 minutes of speech) from 146 learners, mostly at CEFR B1–B2 levels. The protocol comprises three tasks, namely a semi-guided interview, a read-aloud task and a written task in which learners imagine the continuation of a dialogue. The transcripts are time-aligned in ELAN (Wittenburg et al., 2006), with separate Turn Units (TUs) for interviewers and learners, ensuring clean separation of learner productions. One of us took part in curating the 2025 CIL data (character-encoding review, re-annotation of the transcriptions, anonymisation and pseudonymisation of the recordings). Table 1 gives the composition of the corpus and of the subsets analysed for RQ1 and RQ2, which the tool handles in two different modes (Section 4.1).

| | CIL-FFL (L2 French) | CIL-EFL (L2 English) | Total |
|---|---|---|---|
| *Design* | | | |
| Speakers' L1 | Ar, Cmn, En, Es | French | |
| Proficiency | B1–B2 | A2–B2[a] | |
| Tasks | Semi-guided interview (∼20 min), read-aloud, written continuation of a dialogue | | |
| Language recorded | French | English | |
| *Size* | | | |
| Speakers | 128 | 18 | 146 |
| Recordings | 250 | 36 | 292 |
| Transcripts | 128 | 18 | 146 |
| Duration (min) | 2,702 | 316 | 3,014 |
| *Analysed for RQ1 (Mode A)* | | | |
| Texts | 17 | 17[b] | 34 |
| NPs extracted | 11,927 | 5,019 | 16,946 |
| *Analysed for RQ2 (Mode B)* | | | |
| Texts | — | 18 | |
| Words | — | 28,679 | |
| NPs extracted | — | 6,823 | |

Table 1: Composition of the Corpus InterLangue (CIL) and of the subsets analysed for each research question. [a]Self-assessed; 15 of the 18 CIL-EFL learners are at B1 or B2. [b]Mode A draws on 17 CIL-EFL transcripts, Mode B on all 18.

### 3.2 ICE-GB as a native baseline

The native baseline is the spoken component of ICE-GB, the British component of the International Corpus of English, compiled and fully parsed by the Survey of English Usage at University College London (Greenbaum, 1996; Nelson et al., 2002). It holds 637,562 words of spoken English in 300 texts (Table 2), of which we kept only the 99 texts of the private-dialogue category (S1A), unscripted conversation, which is the register closest to an interview; adding broadcast talk or scripted monologue would have mixed part of the speech–writing contrast, along which nominal complexity varies most (Biber and Gray, 2011), into the learner–native one.

## 4 Design and Implementation of the Tool

### 4.1 Overview of the pipeline

The tool extracts and measures noun phrase complexity in learner productions. It is a modular pipeline. Raw transcripts are cleaned, parsed with UD-based models and turned into nominal con-

[2] https://corpusinterlangue.nakala.fr/

| ICE-GB, spoken component | |
|---|---|
| *Design* | |
| Speakers | Adult L1 British English |
| Texts dated | 1990–1992 |
| Release | 2 (2006), ICECUP 3.1 |
| Annotation | Fully parsed treebank |
| *Size* | |
| Texts | 300 |
| dialogue | 180 (100 private, 80 public) |
| monologue | 120 (70 unscripted, 50 scripted) |
| Words | 637,562 |
| Duration | >70 h |
| Parse trees | 59,640 |
| *Analysed here (S1A)* | |
| Texts | 99 |
| Words | 165,919 |
| NPs extracted | 44,025 |

Table 2: The spoken component of ICE-GB, used as the native baseline.

structions, over which the metrics are computed. Any language with a UD model can in principle go through it.

The pipeline operates in two modes, each tailored to specific comparison tasks. Mode A is designed for cross-linguistic comparisons, such as learner French vs. learner English (Figure 1). It relies on UDPipe 2 (Straka, 2018), queried through its REST service, with models trained on Universal Dependencies 2.17 (Straka, 2025; de Marneffe et al., 2021). To match the spoken nature of our data, we selected models trained on spoken treebanks, namely `french-rhapsodie-ud-2.17`, trained on the Rhapsodie treebank of spoken French (Lacheret et al., 2014; Kahane et al., 2021), and `english-childes-ud-2.17`, trained on UD English-CHILDES, a treebank of spontaneous child and caregiver speech (Liu and Prud'hommeaux, 2023; Yang et al., 2025). Mode B compares learners and native speakers of the same language, here learner and native English. It runs a heavier parser, spaCy's transformer-based `en_core_web_trf` 3.8.0 (Honnibal et al., 2020), built on RoBERTa (Liu et al., 2019), which the denser noun phrases of native spoken English call for.

Both modes share the same six steps, namely input, cleaning, parsing, CoNLL-U structuring, nominal-construction extraction, and mapping to a canonical model. Comparability rests on the last of these. The metric code never reads a parser's raw output, only the canonical records, so that changing

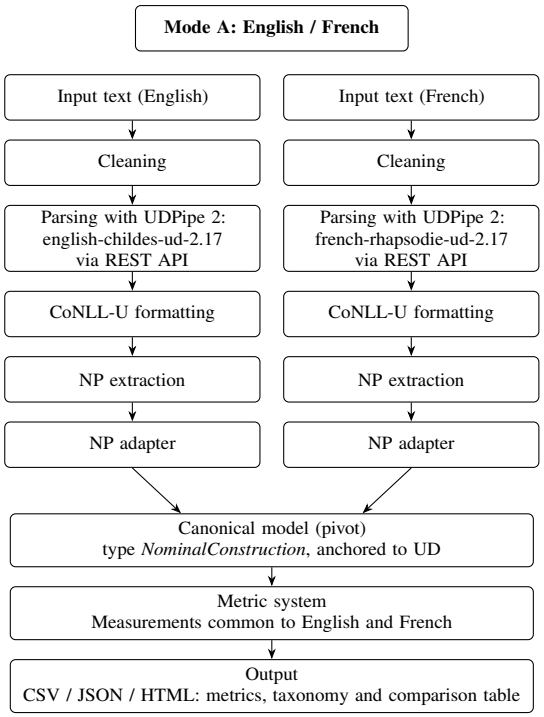



Figure 1: The Mode A pipeline. Each language is cleaned, parsed with UDPipe 2 and converted to CoNLL-U separately; language-specific adapters then map the extracted noun phrases onto a shared canonical model anchored to Universal Dependencies, from which the same metrics are computed for both languages.

the parser or the language leaves the metric definitions untouched. CoNLL-U, the plain-text format of UD treebanks,[3] stores one token per line in ten tab-separated fields (index, form, lemma, universal part of speech, language-specific tag, morphological features, head, dependency relation, enhanced dependencies and a free *MISC* field); every parse reaches the extraction step in that form.

### 4.2 Nominal-construction extraction in Mode B

With English on both sides, Mode B is free to use measures that have no French counterpart and to rely on spaCy's transformer-based English parser (Section 4.1). That does not make the parse trustworthy on spoken learner data, for the reasons given in Sections 1.1 and 2.2. Mode B therefore never reads noun phrases straight off the tree. It extracts them through a constrained procedure and checks the result against a gold standard (Figure 2; Section 5).

Each input is parsed by spaCy and written as en-

[3] https://universaldependencies.org/format.html

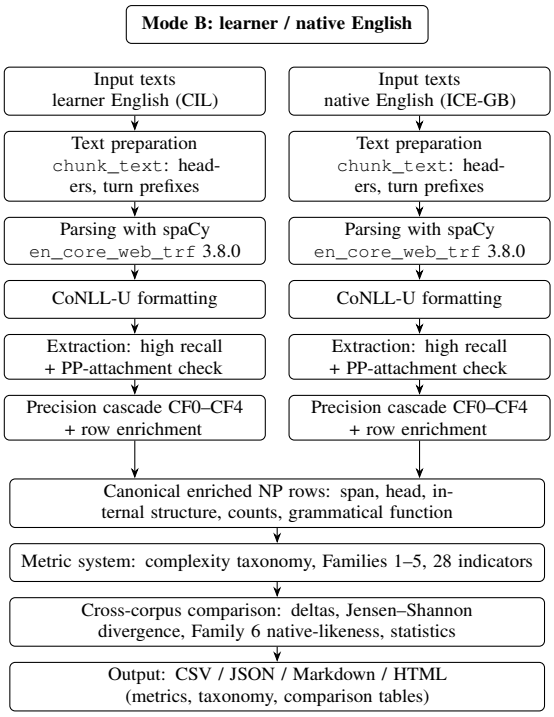



Figure 2: The Mode B pipeline. Learner (CIL) and native (ICE-GB) texts are processed by the same chain of text preparation, parsing with spaCy, CoNLL-U formatting, high-recall extraction and a precision cascade. The resulting noun phrases are stored as canonical records, from which the same metrics and cross-corpus comparisons are computed.

riched CoNLL-U, where spaCy relation labels that have a UD counterpart are mapped onto it in the *MISC* field (*dobj→obj*, *relcl→acl:relcl*). ICE-GB was a special case. We benchmarked spaCy's part-of-speech tags against the hand-corrected ICE-GB tags and found only 85.8% agreement on written and 81.1% on spoken samples, too low to rely on; we therefore kept the ICE-GB tokens and tags and let the parser supply the dependency layer only.

Extraction proper takes three steps. A generator first proposes a span around every nominal head, pronouns and verbless nominal predicates included; at this stage recall comes first. Our error analysis then pointed to one dominant fault, namely right boundaries stretched over a following prepositional phrase. A rule based on the dependency parse now checks whether a prepositional phrase attaches to the head or higher up, and it removed 71% of that error class. Some cases still escape it. In *I'm born in Fougères in France* (CIL, L1), the extractor returns *Fougères in France*, where the annotators marked *Fougères* and *France* as two constructions. Finally, a precision cascade (filters CF0–CF4) drops empty, determiner-only, artefact, idiom and contained spans and collapses duplicate coordinations. Each surviving candidate becomes one canonical record per maximal noun phrase (head, span, modifiers, embedded layers, coordination, grammatical function), and every metric is computed from these records. One rule governed development throughout. A change to the extractor stayed only if it did not lower the strict F1 on the gold standard.

### 4.3 Metrics for noun phrase complexity

The metrics are organized into eleven families, viz. family A, frequency and density; family B, phrasal modification and compression; family C, clausal modification and elaboration; family D, modification scope and syntactic distribution; family E, recursion, embedding, and depth; family F, nominalization and density of nominalized heads; family G, structural and modifier-type diversity; family H, lexical diversity of heads; family I, head type and referential density; family J, the genitive and possessive system; and family K, coordination. They are shared by the two modes, with 18 metrics in Mode A and 17 in Mode B. Table 3 presents the first four with examples. The full metric table[4] gives, for each of the 35 metrics, its formula, unit, comparability between French and English, and counterpart in the other mode.

Family A counts nominal material, per 1,000 words or per finite clause. On its own it says little; it is the baseline against which the other families are read. In Mode A it also breaks modification down by type (epithets, noun premodifiers, postnominal prepositional phrases, relatives, complement clauses, appositions), because these do not occupy the same positions in French and English.

Families B and C carry the distinction drawn by Biber and Gray (2016). Phrasal modifiers (adjectives, nouns, genitives, prepositional phrases, appositives) compress information inside the noun phrase; clausal ones (relative, complement and non-finite clauses) elaborate it. Family B adds informational compression, the number of modifier-borne propositions per finite clause. Read together, the two families ask whether a speaker packs content into the phrase or unfolds it.

Family D looks at breadth and place, that is, how

[4] https://gitlab.huma-num.fr/lidile/constructions-nominales/-/blob/b522959f914773585d810aeca1e7951e21050932/Metrics_Learners_EN_and_FR_FAMILIES_EN__.ods

| Family | What it captures | Example metrics | Examples |
|---|---|---|---|
| **A · Frequency and density** | Normalised frequencies of nominal material, before any complexity | NP density; NPs per finite clause; complement-clause density | *the fact that there is many things to do* (CIL, L1) and *the fact that she's really good* (ICE-GB S1A-005) contain noun complement clauses, counted per 1,000 words; FR *le fait qu'il est parti*. *the accuracy of my results* (ICE-GB S1A-008) counts as one NP. |
| **B · Phrasal modification and compression** | Information packed inside the phrase by adjectives, nouns, genitives, PPs and appositives | Phrasal modifiers per NP; informational compression | Adjective and PP postmodifier in *the best food in the world* (CIL, L13). Noun premodifier and of-phrase in *the response curves of the tape* (ICE-GB S1A-008). FR *une décision sur le budget*. |
| **C · Clausal modification and elaboration** | Propositional elaboration of the NP by relative, complement and non-finite clauses | Clausal modifiers per NP | Relative clauses in *someone who had an Airbnb for us* (CIL, L7) and *the one that Philip's just done* (ICE-GB S1A-008). FR *l'homme qui est arrivé*. |
| **D · Modification scope and distribution** | How many NPs are modified, and where in the sentence complexity sits | Share of modified NPs; NP complexity by function; subject-NP complexity share | Simple subject *my name* (CIL). Complex NP outside the subject slot in *a small town in the South of Brittany* (CIL, L16). Complex, postmodified NP in *the biggest interest in my life* (ICE-GB S1A-003). Heavy subject in FR *[le rapport sur la crise] nous a surpris*. |

Table 3: The first four of the eleven metric families (A–D), with example metrics from both modes. English examples are authentic (CIL learner number or ICE-GB file); French examples are constructed illustrations. Families E–K are described in the text and in the full metric table.

many noun phrases carry any modification at all and where the complex ones end up in the clause. The share of complex subjects is the one we watch most closely.

The remaining families appear in the full table. Family E covers recursion, embedding and depth through recursive depth and two parse-based indicators, mean dependency distance (Liu, 2008) and mean tree depth within the noun phrase. The two do not always agree. Depth grows when modifiers nest inside modifiers, distance when material is strung out linearly, as in coordination, and the disagreement turns out to be informative (Section 6.3). F measures nominalisation; G structural and modifier-type diversity, through a moving-average type–token ratio (Covington and McFall, 2010) over noun phrase patterns and the Shannon entropy (Shannon, 1948) of modifier types; H the lexical diversity of head nouns; I the share of lexical rather than pronominal heads; J the 's- and of-genitives; K coordination inside the noun phrase. Mode B computes 22 scalar indicators and 6 distributional profiles in all; the 17 metrics listed for it in the table are those that pass the source-validation and coverage checks of Section 5.

## 5 Evaluation Methodology

Before any result can be read, two questions must be settled. Is the extraction right, and are the group differences real?

To answer the first question, we built manual gold standards. For Mode A, a single annotator built a gold standard of 200 nominal constructions per language (French and English), so that the tool's output on the same sample could be compared with a human analysis. For Mode B, the gold standard contains 504 span types, 239 from CIL learner English and 265 from ICE-GB spoken English. Mode B annotation followed a written specification, with one span per maximal construction, embedded layers recorded inside it, and explicit rulings on free relatives, idiomatic prepositional objects and disfluent repairs. The specification was revised once, after an audit of the first annotation rounds showed where two careful annotators could legitimately diverge. Two annotators then worked independently; before adjudication, their span-level F1 agreement was .992 on the learner slice and 1.000 on the native one. The Mode B gold standard was then adjudicated by a human annotator and an AI assistant (Claude Opus 4.7, Anthropic) to ensure consistency; the Mode A gold standard, which comes from one annotator, was not adjudicated.

We report precision, recall and F1 under strict matching, in which an extracted span counts as correct only if it matches a gold span exactly. In Mode B, scoring is type-level within each utterance. A span that finds the right construction but places a boundary differently therefore counts as an error; such boundary errors are analysed separately in Section 6.1.

Group differences are tested with the Mann–Whitney U test (Mann and Whitney, 1947), which suits independent samples of unequal size and does not assume normality. Significance alone is not enough, so each test comes with Cliff's delta (Cliff, 1993), read with the thresholds of Romano et al. (2006), namely .147, .33 and .474 for small, medium and large effects. With 22 simultaneous tests, all $p$-values are Holm–Bonferroni corrected (Holm, 1979). The unit of analysis is the text. In Mode B, distributional profiles are also compared with the Jensen–Shannon divergence (Lin, 1991). A second-stage profile then synthesises the scalar metrics. Of its 23 candidate metrics (the 22 tested ones plus the share of core-argument noun phrases, which has no test of its own), it keeps 17. It drops four whose per-text source values do not match the tested values, one whose coverage is insufficient and the untested one. It then clusters the retained metrics correlated at $|r| \geq .80$, orients each standardised metric in the direction of the observed contrast, and averages them by family. Because 18 learner texts face 99 native ones, a balanced bootstrap (Efron and Tibshirani, 1993) with 5,000 resamples sets the learner texts against equally sized native subsamples.

## 6 Results

### 6.1 Extraction accuracy

For Mode A, the tool's output was compared with the annotator's analysis of the same 200 nominal constructions per language. Strict F1 reaches .950 for learner French and .860 for learner English (Table 4). Figure 3 gives the corresponding confusion matrix and explains why precision, recall and F1 coincide in this balanced design.

In Mode B, strict F1 reaches .954 on learner English and .787 on native English, against 504 gold span types (Table 5). The gap has two sources. Native noun phrases are denser and deeper, and ICE-GB is not segmented like CIL. Most residual errors, however, concern the boundary of a construction that was correctly detected. On the full

| Slice | Gold | TP | FP | FN | P | R |
|---|---|---|---|---|---|---|
| Learner EN | 200 | 172 | 28 | 28 | .860 | .860 |
| Learner FR | 200 | 190 | 10 | 10 | .950 | .950 |
| Strict F1 = .860 (learner EN), .950 (learner FR) | | | | | | |

Table 4: Mode A extraction accuracy against the gold standard (strict, type-level matching). Gold = gold nominal constructions; TP, FP, FN = true positives, false positives and missed gold constructions.

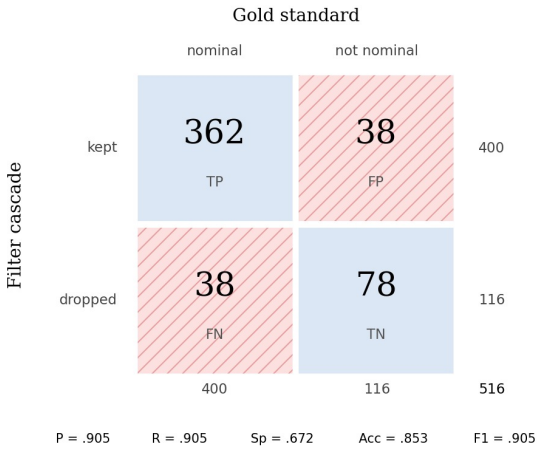



Figure 3: Confusion matrix of the Mode A precision cascade (pooled CIL learner English and learner French gold slices), computed over the 516 candidate span types on which the cascade decides. Because the extractor is constrained to return as many candidates as there are gold constructions (400), false positives and false negatives are equal in number, which yields $P = R = F1$.

gold sample, which also includes an ICE-GB written slice, 117 of the 161 false positives and of the 161 misses form boundary-overlap pairs, leaving 44 residual false positives and 44 residual misses. The extractor also errs on the short side more often than on the long side. In 72 of these pairs the gold span contains the extracted one, in 44 the reverse.

Figure 4 opens up the Mode B scores. Extraction has no natural negative class, so the matrix covers only the 772 span types on which the cascade decides. The cascade keeps 432 genuine noun phrases and 61 spurious ones, and discards 23 genuine and 256 spurious. Its errors lean one way. It rarely throws out a real noun phrase (recall = .949) but lets roughly one spurious candidate in five through (specificity = .808). The 72 end-to-end misses thus come from two places, namely 23 noun phrases the cascade wrongly discarded and 49 that were never proposed at all, a group in which recall can still be

| Slice | Gold | TP | FP | FN | P | R |
|---|---|---|---|---|---|---|
| CIL learner | 239 | 227 | 10 | 12 | .958 | .950 |
| ICE-GB spoken | 265 | 205 | 51 | 60 | .801 | .774 |
| Strict F1 = .954 (CIL), .787 (ICE-GB spoken) | | | | | | |

Table 5: Mode B extraction accuracy against the adjudicated gold standard (strict, type-level matching). Gold = gold span types; TP, FP, FN = true positives, false positives and missed gold spans.

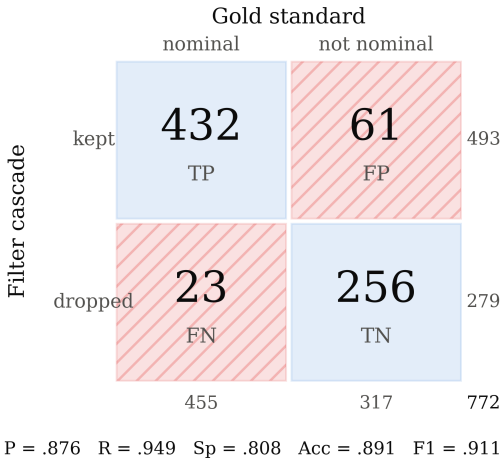



Figure 4: Confusion matrix of the Mode B precision cascade (CIL learner and ICE-GB spoken gold slices), computed over the 772 candidate span types on which the cascade decides. The 49 gold constructions never generated as candidates fall outside this population.

gained.

### 6.2 Learner French vs. learner English

The main finding for RQ1 is that, across phrasal measures, learner English is consistently more complex than learner French. Phrasal modifiers per noun phrase, for instance, were higher in learner English (0.61) than in learner French (0.41; $\delta = -1.00$, $p < .001$), and the share of modified noun phrases was higher as well (50% vs. 39%; $\delta = -.99$, $p < .001$). Phrasal depth, modifier-type entropy and structural diversity also showed significantly higher values for learner English (Figure 5).

However, the tool also revealed that learner French compensates for its lower phrasal complexity by using more appositive and looser structures. The share of appositive structures, in particular, was higher in learner French ($\delta = +.56$, $p < .01$).

Clause-level density measures, such as the number of noun phrases per finite clause, were largely comparable between the two groups, indicating

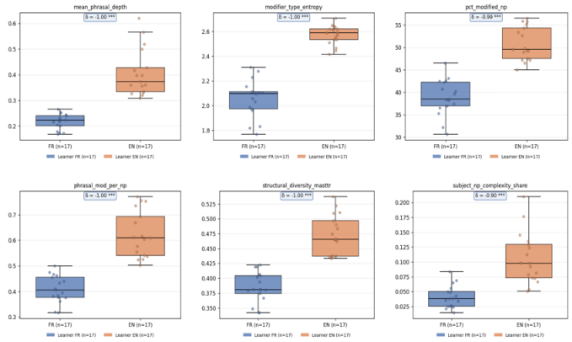


Figure 5: Per-text distributions of the six metrics that separate the two learner corpora most sharply (Learner FR vs. Learner EN, $n = 17$ texts each). Boxes show the median and interquartile range, whiskers the range of the data, and points the individual texts. The top row shows mean phrasal depth, modifier-type entropy, and the percentage of modified NPs; the bottom row shows phrasal modifiers per NP, structural diversity (MATTR), and the share of complexity carried by subject NPs. On all six metrics, the EN texts score higher than the FR texts. Effect sizes are given as Cliff's $\delta$ (negative values indicate FR < EN), ranging from $-0.90$ to $-1.00$, meaning the two distributions barely overlap or do not overlap at all. All contrasts are Holm-corrected and significant at $p < .001$.

that the differences lie primarily in how complexity is distributed rather than in the overall amount of nominal material produced.

### 6.3 Learner English vs. native English

Does the tool separate learner from native English? It does, and in a specific way. Natives pack complexity into the noun phrase, whereas learners spread it over longer, flatter structures. Table 6 gives the full panel of per-text means; all 18 CIL-EFL interviews and 99 ICE-GB private dialogues were retained, and the imbalance is handled as described in Section 5.

The native advantage lies in compression. Natives use more phrasal modifiers per noun phrase (0.091 vs. 0.060; $\delta = -.65$, $p < .001$), and their of-genitive share is three times the learners' (1.65% vs. 0.55%; $\delta = -.71$, $p < .001$), the largest single effect in the panel. Informational compression (0.185 vs. 0.105; $\delta = -.67$, $p < .001$), recursive depth and tree depth (both $\delta = -.54$, $p < .01$) point the same way, as do clausal modifiers per NP ($\delta = -.42$, $p = .044$) and NP density ($\delta = -.46$, $p = .028$).

Figure 6 shows the per-text distributions. The corpora overlap on every metric taken alone; they

| ID | Metric | L | N | $\delta$ | $p_H$ |
|---|---|---|---|---|---|
| A.1 | NPs per 1,000 words | 236.5 | 249.2 | **−.46** | .028 |
| A.2 | NPs per finite clause | 2.95 | 2.51 | **+.66** | <.001 |
| B.2 | Phrasal modifiers/NP | .060 | .091 | **−.65** | <.001 |
| B.3 | Inform. compression | .105 | .185 | **−.67** | <.001 |
| C.2 | Clausal modifiers/NP | .002 | .004 | **−.42** | .044 |
| D.3 | Complex subjects (%) | 8.16 | 7.01 | +.08 | 1 |
| E.3 | Recursive depth | .040 | .056 | **−.54** | .005 |
| E.4 | Tree depth | .195 | .244 | **−.54** | .005 |
| E.5 | Dependency distance | .359 | .274 | **+.62** | <.001 |
| F.2 | Nominalisations/NP | .033 | .035 | +.06 | 1 |
| G.2 | Modifier entropy (bits) | 2.45 | 2.42 | +.16 | 1 |
| G.4 | Pattern MASTTR | .165 | .164 | −.07 | 1 |
| H.2 | Head-noun MATTR | .674 | .648 | +.39 | .082 |
| I.1 | Lexical NPs (%) | 24.0 | 24.5 | −.12 | 1 |
| J.1 | 's-genitive (%) | 1.22 | 1.60 | −.29 | .362 |
| J.2 | of-genitive (%) | 0.55 | 1.65 | **−.71** | <.001 |
| K.1 | Coordinate NPs (%) | 2.80 | 1.33 | **+.54** | .004 |
| *Not retained in the metric table* | | | | | |
| — | NP length (words)† | 1.91 | 1.70 | **+.53** | .005 |
| — | Premodifiers/NP† | .154 | .182 | **−.42** | .047 |
| — | Postmodifiers/NP† | .045 | .058 | **−.46** | .028 |
| — | Complex NPs (%)† | 15.5 | 17.7 | −.37 | .110 |
| — | Phrasal/clausal‡ | 21.0 | 24.3 | −.20 | 1 |

Table 6: Per-text means of the 22 scalar metrics in learner (L, 18 CIL texts) and native (N, 99 ICE-GB texts) English. IDs follow the families of the full metric table (A–K). $\delta$ = Cliff's delta (positive values, learner higher); $p_H$ = Holm-corrected $p$ over the 22 tests; bold = significant. †Failed the second-stage source validation. ‡Computable for 7 learner and 77 native texts only. The full metric table, with standard deviations, is available online (https://qrco.de/bh0mtx).

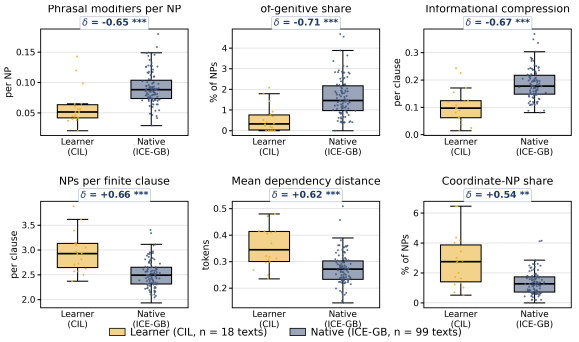



Figure 6: Per-text distributions of the six metrics that separate the two corpora most sharply. Boxes show the median and interquartile range, points the individual texts. The top row holds the three native-higher metrics of nominal compression, the bottom row the three learner-higher metrics of nominal spread. All contrasts are Holm-corrected and significant at $p < .001$, except Coordinate-NP share ($p < .01$).

differ by a shift in central tendency that recurs from one measure to the next, and no single measure provides a clean threshold.

Learners, by contrast, distribute their nominal material over more, and less compact, noun phrases. They produce more noun phrases per finite clause (2.95 vs. 2.51; $\delta = +.66$, $p < .001$), with longer dependency distances inside them (0.359 vs. 0.274; $\delta = +.62$, $p < .001$) and twice as much internal coordination (2.80% vs. 1.33%; $\delta = +.54$, $p < .01$). Their noun phrases are also longer, at 1.91 content words against 1.70 ($\delta = +.53$, $p < .01$). We read this last figure with care, since mean length is one of the four measures flagged by the source validation, although its direction holds across source layers. Length does not equal complexity here. Learners lengthen a noun phrase by adding a conjunct, as in *my mother and my father* (CIL, L2); natives lengthen it by embedding, as in *the response curves of the tape* (ICE-GB S1A-008).

Diversity and external distribution do not separate the groups. No diversity metric survives correction, nor does the share of complex subjects (all $|\delta| \leq .39$, $p \geq .082$), and the distributional profiles are close, with Jensen–Shannon divergences between .001 (coordination, embedding depth) and .024 (head word class). We would not call this equivalence, since with 18 learner texts a null result is inconclusive. On these data, though, learners draw on the same inventory of nominal structures as natives, in similar proportions, and differ in how densely they combine them.

The second-stage profile says the same thing, more strongly. The core profile separates the corpora with $\delta = .87$ ($p < .001$), more than any single metric ($|\delta| = .71$ at best), and does so in every balanced resample (median $\delta = .87$, 95% interval $[.79, .93]$). Grouped by the dimensions the profile uses, the separation sits in the parse-based indicators (.97), frequency (.69) and NP-internal structure (.68), all at $p < .001$, whereas diversity (.16) and external distribution (.08) contribute little. One caveat applies. The profile is oriented by the contrast it measures, so it synthesises the same metrics without validating them independently.

## 7 Discussion

The two comparisons do not ask the same question, but their answers fit together. Learner English carries more phrasal modification than learner French and less than native English, and in each comparison the group with less modification compensates

elsewhere (Sections 6.2 and 6.3). In both cases the difference lies less in how much nominal material speakers produce than in where they put the structure. The spoken data thus echo, on a small scale, what has been described for writing, where nouns and prepositional phrases as modifiers mark more advanced production (Biber et al., 2011; Parkinson and Musgrave, 2014).

Regarding the study's depth, many other aspects could have been tackled, namely a native English vs. native French comparison, that account for interlinguistic differences that are not learner-specific. Or, a monolingual native French vs. learner French study, parallel to the English substudy that shed light to the interlanguage development process in both languages i.e. English and French. But also, one or more similar substudies on other less closely related languages. However, it is worth mentioning that including all these aspects of study involves time and dedicated corpora building to support it. Given that we already embarked in an analysis tool development, we chose not to engage in compiling all the possible analysis in one work. They will be discussed accordingly in future works.

### 7.1 Comparison with TAASSC

TAASSC (Kyle, 2016), whose phrasal indices were introduced in Section 2.1, is the closest existing point of reference for our tool, and much of our work builds on the questions it made measurable. Table 7 sets the two tools side by side, to situate them without ranking them.

**TAASSC in brief.** TAASSC groups its indices into four families (Kyle, 2016). They are the fourteen large-grained indices of the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010), such as mean length of T-unit (MLT), dependent clauses per clause (DC/C) and complex nominals per clause (CN/C); fine-grained clausal indices, which count the dependents of each clause by type; the fine-grained phrasal indices described in Section 2.1; and usage-based indices of syntactic sophistication, which score main verbs, verb argument constructions (VACs) and verb–VAC combinations by their frequency and strength of association in the written sections of the Corpus of Contemporary American English (COCA; Davies, 2009) (see Kyle and Crossley, 2017; Kyle et al., 2021). The original release relied on Stanford CoreNLP and Stanford typed dependencies (de Marneffe and Manning, 2008); Kyle et al. (2021) identify VACs with the Stanford neural dependency parser (Chen and Manning, 2014), and version 2.0 of the tool parses with spaCy, as our Mode B does.

| | TAASSC | Mode A | Mode B |
|---|---|---|---|
| Data | Written English essays | Spoken learner FR and EN | Spoken learner and native EN |
| Parser | Stanford; spaCy (v2.0) | UDPipe 2, spoken UD models | spaCy `trf`; ICE-GB tags |
| Unit | Dependents of NP heads | Nominal construction, gold-evaluated | Maximal NP span, gold-evaluated |
| Indices | Means, SDs by NP × dependent type | 18 metrics (A–K) | 22 scalar, 6 distrib.; 17 validated |
| Embedding | — | Phrasal depth | Depth, dependency distance |
| Frame | Dependent types | Phrasal vs. clausal; UD function | Phrasal vs. clausal; *'s* vs. *of*; compression |
| French | — | Adj. position, *de*/*à*/*en*, relatives, determiner | — |
| Diversity | SDs of counts | Entropy | MATTR, entropy, JSD |
| Statistics | Indices only | MWU, Cliff's $\delta$, Holm | Same + JSD, profile |
| Beyond scope | — | Clausal; sophistication | Clausal; sophistication |

Table 7: TAASSC (Kyle, 2016; Kyle and Crossley, 2018) compared with the two modes of our tool.

**Common ground.** Both tools count the modifiers of nominal heads and relate them to the function of the phrase. Our phrasal modifiers per noun phrase are conceptually close to TAASSC's dependents per nominal, and our share of complex subjects to its indices for nominal subjects. Both also report their values per text. TAASSC returns one row of indices per text, and the text is our unit of analysis as well (Section 5), which makes a text-by-text comparison of the two tools possible. The correspondences between the two sets of measures nevertheless remain conceptual, since the definition of the noun phrase, the parser, the dependency scheme and the normalisation differ. Our tool's contribution lies mainly in these differences, which we take in turn.

**(1) French and English on a common basis.** TAASSC is designed for English, and its sophistication indices rely on an English reference corpus. Our tool, anchored in Universal Dependencies (de Marneffe et al., 2021), describes a second lan-

guage with the same categories as the first. In Mode A, the French and English parses produced by UDPipe 2 (Straka, 2018) meet in the canonical model of Section 4.1, where modifiers are coded by their UD relation whatever their position. The French postnominal adjective and the English prenominal adjective are thus both counted under *amod*. Mode A also adds measures for properties of the French noun phrase, such as adjective position, prepositional phrases in *de*, *à* and *en*, the type of relative pronoun and the omission of the determiner. The comparison of learner French with learner English (Section 6.2) depends on this design, which is the main extension our tool offers.

**(2) The noun phrase as a delimited unit.** TAASSC counts the dependents that a parser attaches to a head, without delimiting the noun phrase itself. Our tool delimits each noun phrase as a unit and evaluates that unit, span by span, against a gold standard (Section 5; Tables 4 and 5), so that a reader can check what was counted. Reporting accuracy on the unit of analysis follows a concern that the TAASSC studies themselves raise. Kyle et al. (2021) report that 97.9% of main verbs but 81.25% of verb–VAC combinations were correctly identified in their learner corpus, and note that such errors carry over into any downstream analysis.

**(3) Embedding and dependency distance.** Treating the noun phrase as a delimited unit also makes its internal structure measurable, through the depth and distance indicators of family E (Section 4.3); Liu (2008) relates the second of these to processing difficulty. A count of dependents does not in itself distinguish dependents attached to the same head from dependents embedded in one another. In our data, depth and dependency distance separate learners and natives more sharply than the other dimensions (Section 6.3), which suggests that this aspect of the noun phrase is worth measuring alongside dependent counts. Because spans depend on segmentation, the dependency-distance half of this result carries the caveat set out in the Limitations section.

**(4) A developmental reading.** The phrasal indices of TAASSC describe how many dependents of each type occur. Families B and C operationalise instead the distinction between phrasal and clausal modification that Biber et al. (2011) place along a hypothesised developmental sequence, and that Biber and Gray (2016) describe as compression against elaboration. The longitudinal study of Kyle et al. (2021) gives one reason why this reading may be useful. Over two years, their adolescent learners of English grew in DC/C and MLT but not in CN/C, and the relation between time and MLT ($r = .443$) became negligible once DC/C was controlled for ($r = -.046$); the authors interpret this in light of Biber et al. (2011), with dependent clauses developing first and noun phrase features possibly later. A measure that sets phrasal against clausal modification may help to follow that shift, which an overall count of complex nominals does not isolate. Whether it does so over time is a question that longitudinal spoken data would need to answer.

**(5) Functional contrasts.** TAASSC separates nominal dependents by type. Our tool adds functional contrasts, namely phrasal against clausal modification, the *'s*-genitive against the *of*-genitive (family J), and informational compression (family B). In our data these contrasts produce the strongest effects (Table 6), and they are robust to the segmentation problem discussed in the Limitations section.

**(6) Diversity.** In TAASSC, variability is expressed through the standard deviation of dependent counts. Our tool adds measures of diversity proper (families G and H, Section 4.3) and, in Mode B, the Jensen–Shannon divergence between distributional profiles. A standard deviation describes how much counts vary across nominals, whereas these measures describe how varied the repertoire of structures is and how close two distributions are.

**(7) Transcribed learner speech.** TAASSC was developed and validated on written texts, mainly TOEFL essays (Kyle and Crossley, 2017, 2018) and learner essays (Kyle et al., 2021). Our tool was developed for transcribed speech from learners and natives. Mode A relies on the spoken-treebank models presented in Section 4.1, Mode B on the hand-corrected ICE-GB tagging (Section 4.2), and both modes filter disfluencies and interactional forms and handle mass and institutionalised bare nouns. Earlier findings suggest that such an adaptation is worthwhile. With the companion tool TAALES, Kyle and Crossley (2016) found that the lexical sophistication features related to writing scores differed between independent and source-based essays, so that indices do not necessarily behave alike across tasks, and complexity itself is

distributed differently in conversation and in writing (Section 2.1).

**(8) Normalisation.** TAASSC normalises mainly per clause or per nominal. Our tool reports measures per 1,000 words, per finite clause and per noun phrase, so that the unit can be chosen according to the question, whether it concerns how much nominal material a speaker produces (family A) or how much each phrase or clause carries (families B to D).

**(9) Statistical analysis.** TAASSC exports indices, and the statistical modelling is carried out separately, as in the mixed-effects analyses of Kyle et al. (2021). Our tool builds the group comparison of Section 5 into the pipeline, in both modes. Since the unit of analysis is the text, values for each speaker remain available next to group results, in line with the caution, recalled by Kyle et al. (2021) after Lowie and Verspoor (2019), that group means may not represent individual trajectories.

**What TAASSC covers and our tool does not.** TAASSC covers several dimensions that our tool does not attempt. It includes clausal complexity, through the L2SCA indices and its fine-grained clausal counts, whereas we measure only compression per clause. It includes usage-based sophistication, its largest family; in Kyle et al. (2021), main verb frequency and DC/C together predicted the time point of an essay ($R^2_{\text{adj}} = .583$). Such indices presuppose a suitable reference corpus, which a comparison of learner speech in two languages would have to supply on both sides. TAASSC also crosses each nominal function with each dependent type; our family D reports complexity by function and the share of complex subjects, but not by dependent type. The two tools are therefore best seen as complementary. TAASSC offers broad coverage of written English, ours a more detailed description of the noun phrase, comparable across French and English and adapted to learner speech.

### 7.2 Implications for SLA and CALL research

For SLA and CALL research, the results argue for phrasal measures over global ones when the data are spoken. The noun phrase is sensitive to developmental differences, and it can now be measured automatically. Because it rests on Universal Dependencies, the tool also serves work on cross-linguistic influence between L1s, where comparable measurement across languages has long been a sticking point. The learner results also warn against reading complexity as a single scale. Learners build more noun phrases per clause, and coordinate them more often, whereas natives build fewer but more heavily modified ones, so that a longer noun phrase is not necessarily a more complex one.

For CALL systems, the metrics could be used for two purposes. First, they could be used as features in proficiency prediction models. Second, depending on their statistical significance, they could be exploited to tailor specific feedback messages giving details on what type of nominal complexity (if any) learners employ. Such fine-grained, structure-specific features would contribute to providing actionable language learning analytics. In class, the tool could also support data-driven learning, with learners comparing their own noun phrases to native ones. The tool is open, so these claims can be checked, extended or overturned by others.

## 8 Conclusion and Future Work

This paper has presented an open, modular tool that measures noun phrase complexity in learner speech through Universal Dependencies, together with an extraction procedure validated against gold standards, with strict F1 between .79 and .95 depending on the corpus. The empirical picture that emerges does not point to missing structures. Learners draw on the native inventory of nominal structures but assemble it into flatter, longer and more coordinated noun phrases, which carry less modifier-borne information per clause than native ones (Table 6, B.3). Because every language-specific step happens before the canonical model, a single metric code serves both the cross-linguistic and the learner–native comparison.

Resegmenting both corpora into AS-units (Analysis of Speech units), that is, units made of an independent clause or sub-clausal unit together with its subordinate clauses (Foster et al., 2000), is a natural next step and, in our view, the least costly improvement available. Adapting the parser to spoken learner English, for instance with the treebank of Kyle et al. (2022), would reduce the parsing errors that the cascade only partly absorbs. Calibration against other measurements would also help, in two ways. Since learner English is measured by both modes, Mode A and Mode B could be run on the same CIL-EFL texts to see which measures agree. The English branch could likewise be compared with TAASSC. Running both tools on

the same transcripts and correlating the measures they share (Table 7), once normalisation is aligned, would provide a test of convergent validity, and any divergence could be traced to the differences discussed in Section 7.1.

## Limitations

The samples are small, with 17 speakers per language for RQ1 and 18 learners against 99 native texts for RQ2, and more proficiency levels and more first languages are needed before any finding is generalised. Mode A needs inter-annotator agreement, since its gold standard comes from a single annotator, and both modes need firmer segmentation guidelines. The analysis covers semi-guided interviews and read-aloud tasks in French and English only. Other registers, such as academic speech and spontaneous conversation, and other language pairs, such as Spanish or German, remain untested.

Segmentation is the main weakness of Mode B. Neither corpus carries sentence punctuation, so the units are long and uneven, with 53.8 words on average in the learner transcripts and 19.7 in the native ones, and maxima of 817 and 344. Every extracted noun phrase is bounded by its unit, and a unit that runs across a turn drags the right boundary with it. In ICE-GB S1A-008, a relative clause on the pronoun *one* swallows the following utterance and yields a 97-word "noun phrase". In CIL_L3, a construction that should stop at *my gym* runs through a turn marker into the next turn. Such cases are rare. Nominals of 15 words or more make up 0.16% of learner records and 0.25% of native ones. The measures behind our main finding, namely *of*-genitive share, phrasal modifiers and informational compression, do not depend on span length. Mean dependency distance does, which is why we trust it least, and it qualifies the result on depth and dependency distance discussed in Section 7.1. The four boundary-sensitive metrics that failed the source validation are reported with a flag (Table 6).

## Ethics Statement

The CIL transcripts analysed here come from the pseudonymised 2025 release of the corpus (Section 3); learners are referred to by codes only, and no personal information is reported. ICE-GB is used as released by the Survey of English Usage (University College London). The use of an AI assistant in the adjudication of the Mode B gold standard is reported in Section 5.

## AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

**Details:** [Describe the AI tools used and their purpose, or state: "No AI tools were used in preparing this work."] "'

## References


Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15(2):223–250.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press, Cambridge.

Douglas Biber, Bethany Gray, Tove Larsson, and Shelley Staples. 2025. 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*, 75(2):575–581.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1):5–35.

Douglas Biber, Bethany Gray, Shelley Staples, and Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46:100869.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar.

Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.

Thomas Gaillat, Leonardo Contreras Roa, and Juvénal Attoumbre. 2021. A data repository for the management of dynamic linguistic datasets. In *CLARIN Annual Conference 2021*, Madrid, Spain (online).

Sidney Greenbaum, editor. 1996. *Comparing English Worldwide: The International Corpus of English*. Clarendon Press, Oxford.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Alex Housen and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Kristopher Kyle, Scott Crossley, and Marjolijn Verspoor. 2021. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 43(4):781–812.

Kristopher Kyle and Scott A. Crossley. 2016. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34:12–24.

Kristopher Kyle and Scott A. Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle and Scott A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45. Association for Computational Linguistics.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

Wander M. Lowie and Marjolijn H. Verspoor. 2019. Individual differences and the ergodicity problem. *Language Learning*, 69(S1):184–206.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Gerald Nelson, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. John Benjamins, Amsterdam.

John M. Norris and Lourdes Ortega. 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*, 30(4):555–578.

Jean Parkinson and Jill Musgrave. 2014. Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes*, 14:48–59.

Jeanine Romano, Jeffrey D. Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*, Cocoa Beach, Florida.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172. Association for Computational Linguistics.

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

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2025. Universal dependencies 2.17 models for UDPipe 2 (2025-11-25). LINDAT/CLARIAH-CZ digital library, Institute of Formal and Applied Linguistics (ÚFAL), Charles University.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Number 17 in Technical Report. University of Hawai'i, Second Language Teaching and Curriculum Center, Honolulu.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.