ExpLay: A new Corpus Resource for the Research on Expertise as an Influential Factor on Language Production

Carmen Schacht
Ruhr-University Bochum, Germany
Faculty of Philology
Department of Linguistics
carmen.schacht@rub.de

Renate Delucchi Danhier
TU Dortmund University
Department of Cultural Studies
Institute for Diversity Studies
renate.delucchi@tu-dortmund.de

Abstract

This paper introduces the ExpLay-Pipeline, a novel semi-automated processing tool designed for the analysis of language production data from experts in comparison to the language production of a control group of laypeople. The pipeline combines manual annotation and curation with state-of-the-art machine learning and rule-based methods, following a silver standard approach. It integrates various analysis modules specifically for the syntactic and lexical evaluation of parsed linguistic data. While implemented initially for the creation of the ExpLay-Corpus, it is designed for the processing of linguistic data in general. The paper details the design and implementation of this pipeline. To demonstrate the pipeline's capabilities and explore linguistic markers of expertise, we present the initial release of the ExpLay-Corpus. This corpus comprises German oral descriptions of urban landscapes elicited from architectural students (characterized as a semi-expert population) and a group of matching laypersons. Using the ExpLay-Pipeline, preliminary analyses of syntactic and lexical complexity between these two groups were conducted. While the primary focus of this work lies on the architecture of the pipeline and its annotation methodology, these preliminary findings serve to showcase the pipeline's functionality and establish ExpLay as an accessible resource for future research on linguistic markers of expertise.

1 Introduction

This research is grounded in three core assumptions concerning the influence of expertise on cognition and language production.

First, it draws on the principle of *linguistic relativity* (Whorf, 1956; Slobin, 1996), which postulates that language plays a role in shaping thought, attention allocation, and cognition in general. Empirical support for linguistic relativity has been documented across various cognitive domains, including color perception (Winawer et al., 2007; Roberson et al., 2000), the conceptualization of motion events (Slobin, 1996; Papafragou et al., 2008) and the use of spatial frames of reference (Levinson, 2003; Majid et al., 2004).

Second, effects similar to *linguistic relativity* are observed beyond language: Expertise, whether professional or personal, can shape cognition in a manner analogous to language. For instance, a neuro-imaging study (Maguire et al., 2000) found structural alterations in the posterior hippocampus of taxi drivers compared to non-drivers, suggesting that its expansion results from extensive navigational experience. Other findings reveal a significant improvement in reaction time for e-sports players (Ersin et al., 2022) as well as decision making and dexterity (Jiang et al., 2020) for non-professional gamers (semi-experts), compared to laypeople. Effects of domain-specific expertise on attention and cognition have also been documented, for example in the field of architecture. In a previous eye-tracking study using stimuli similar to those in the present research, Mertins et al. (2020) found that architects and laypeople differ systematically in how they allocate visual attention. While laypeople focused more on human figures in indoor scenes; architects attending to outdoor scenes concentrated longer on architectural elements, particularly upper-level features like roofs, whereas laypeople remained focused on elements at eye level.

Third, rational communication aims to maintain linguistic code maximally efficient and to this end adapts dynamically to situational and communicative demands. Just as language influences cognition, expertise influences language production. This is reflected in domain-specific, conventionalized linguistic codes (Teich et al., 2021), which facilitate both perception and communication within specialized fields. Such patterns are evident in domain-related language use and mirror the cogni-

tive effects of linguistic relativity discussed earlier. This phenomenon has been observed across various domains, including literary discourse (Degaetano-Ortlieb and Piper, 2019), the physical sciences (Halliday, 1988/2004), and diachronic shifts in scientific English (Degaetano-Ortlieb and Teich, 2022, 2018; Biber et al., 2011; Biber and Gray, 2016; Juzek et al., 2020) as well as scientific German (Jakobi et al., 2024). Domain-specific features also emerge in the use of linguistic structures such as compounding (Gamboa et al., 2025) and metaphor usage (Halliday, 1988/2004; Webster, 2018) in scientific and technical texts.

Despite growing interest in the cognitive effects of expertise, little is known about how architectural expertise influences spatial cognition and its linguistic encoding. This study addresses this gap by analyzing how architects describe urban and natural landscapes. To investigate the linguistic manifestations of expertise in architecture, a dedicated corpus resource was curated and subjected to a preliminary linguistic analysis.

As an initial exploratory step, the study focused on syntactic and lexical complexity as indicators of domain-specific language use, comparing the speech production of semi-expert participants (students of architecture) with that of non-expert controls (students of German language and literature). The metrics selected incorporate a range of syntactic and lexical measures, thereby capturing a broad variety of structural linguistic features that may exhibit domain-specific variation across the two groups. Given the central role of communicative efficiency, we decided to focus on linguistic complexity as a suitable entry point for exploration of experts' language use. A higher communicative efficiency is often associated with denser, more complex structures (compared to more linear constructions), suggesting the hypothesis that expert language production may exhibit greater structural complexity than that of non-experts.

This preliminary analysis primarily serves to demonstrate the capabilities of the parsing and evaluation pipeline presented in this paper. It is not intended as an exhaustive account of architectural expertise in language use.

2 Previous work

Most existing studies on complexity measures such as dependency length so far focus on dependency processing (Juzek et al., 2020; Futrell et al., 2015) rather than on dependency production. Moreover, they tend to treat expertise as a factor either in the processing of other expert's data (Jakobi et al., 2024) or in written expert language such as scientific discourse (Banks, 2003; Biber et al., 2011). Studies applying the Universal Dependencies (UD) framework (de Marneffe et al., 2021) to spoken data usually focus on the creation of spoken language treebanks (Dobrovoljc, 2022; Dobrovoljc and Nivre, 2016) rather than addressing differences between the groups of speakers who produced the linguistic data for those treebanks in the first place.

While Dobrovoljc and Nivre (2016) at least address some particularities of oral data during the annotation process of the resource, in general very little attention is given to the characteristics of the speakers who produced the linguistic material and possible differences among groups (such as experts vs. non-experts). To address the gap between these two areas, the present study curates experimentally elicited spoken data from both expert and non-expert participants. In doing so, it offers a novel corpus resource to facilitate further investigation into how expertise shapes linguistic structure in spoken language.

This approach is motivated in particular by the eye-tracking findings of Mertins et al. (2020), which revealed systematic domain-dependent differences in visual attention patterns between architects and non-architects. These findings suggest domain-specific cognitive processing, and, by extension, the possibility of domain-specific linguistic realizations of such cognitive behaviors, consistent with the study's core assumptions. So we aspire to use a corpus-based and computational linguistic approach to analyze verbalizations in a similar experimental set-up as the one used in the eye-tracking study.

To conduct an initial exploratory analysis of potential syntactical and lexical differences between expert and non-expert verbalizations in addition to the curation of the resource itself, this study draws on established (syntactic and lexical) complexity metrics. These include dependency distance (Gibson, 1998; Futrell et al., 2015), dependency and constituent-tree tree height (Yngve, 1960), dependency-based clause count (Biber, 1988; Lu, 2011), and constituency-based phrase count (Lu, 2011) as well as word class (Shi and Lei, 2021). Additionally, following the methodology of Park (2024) we apply Principal Component Analysis (PCA) to generate a combined syntactic complex-

ity score, using the PC-loadings to determine the weightings of individual metrics contributing to the combined score.

3 ExpPlay release

The initial release of the ExpLay-Resource comprises the raw (unparsed) data, the parsing and evaluation pipeline, as well as the parsed corpus of experimentally elicited spoken language produced by experts and non-experts in the field of architecture. Following the *silver standard* approach described in (Rebholz-Schuhmann et al., 2010), the dataset was manually pre-processed, automatically parsed, and partially curated across multiple linguistic levels using the ExpLay-Pipeline introduced in this paper. This pipeline integrates several state-of-the-art tools for natural language processing, linguistic annotation, and the evaluation of linguistic structures. The full resource including the pipeline and corpus is made freely available on Gitlab.[1] The entire dataset can be accessed under a CC BY 4.0 license on OSF[2], to support open-access initiatives and facilitate accessible future research in linguistics.

3.1 Data collection

A controlled, online language production experiment was conducted via Zoom, in which participants were asked to orally describe a series of images depicting urban and natural environments (Figure 1). The images were presented one at a time in randomized order, and participants were given unlimited time to respond. The participants were instructed to describe each scene as if speaking to an artist who had never seen it and would need to recreate it through drawing. This task design intentionally avoided priming architecture students to adopt an expert-oriented communicative register, thereby ensuring that both groups (experts and non-experts) shared a common baseline assumption about their audience. As a result, any observed effects in the expert group's descriptions can be interpreted as reflecting general language processing and cognitive-linguistic tendencies influenced by the presence or absence of architectural expertise of the respective participant group, rather than from professional communication demands.

All descriptions were produced in German, which was the native language of all participants. Afterwards, a second group of laypeople with no architectural background completed the same task under identical conditions. for the present study, an initial sample of 13 participants per group was selected from a larger pool of participants. The control group was deliberately selected to closely match the architect group in gender, age, and multilingual status, thereby controlling for potential confounding variables while isolating the influence of domain-specific expertise. This study design allows to compare language use between participants with and without architectural training, while keeping other demographic and linguistic factors constant.

Because the expert sample in this study consists of architecture students rather than practicing architects with extensive professional experience, the level of domain-specific expertise must be interpreted with some caution and can thus be more appropriately characterized as a semi-expert group. Nevertheless, we still anticipate some measurable differences between students with architectural training and those without, reflecting varying degrees of architectural knowledge.

The resulting initial sample for the ExpLay-Resource comprises 13 participants per group: Among the experts, 5 were male and 8 female; among the laypeople, 4 were male and 9 female. All participants were between 19 and 32 years old. Each group included 12 monolingual and 1 bilingual speaker. All oral descriptions were recorded, transcribed, and subsequently analyzed.

Figure 1: Experimental set-up in the verbalization experiment showing the used visual stimuli.

For this initial release of the ExpLay corpus, only the urban environment stimuli (images B1 to B5) were selected, as these are more likely to elicit domain-specific differences between expert and

[1]Gitlab: https://gitlab.ruhr-uni-bochum.de/schaccmr/explay-resource.git.

[2]The entire dataset is made freely available on OSF: https://osf.io/ky87h/?view_only=4a0c7ae6a07c4fe89bc8632787742616.

Dummy Token	Function	Category
%	Grammatical correction	1
&	Insertion of ellipsis (oral structure)	2
$	Insertion of ellipsis (stylistic structure)	2
§	Nominalization	3
@	Substantivized determiner/quantifier	3

Table 1: Overview of dummy tokens used to mark different types of insertions in the data.

non-expert participants due to their closer thematic alignment with architectural expertise. The natural environment stimuli will be included in a future release. Each participant contributed five text productions, resulting in a total of 130 descriptions in the current version of the resource.

3.2 Data curation

To prepare the transcripts for annotation, the oral productions were first extracted and cleaned according to a strict protocol aimed at ensuring comparability while preserving the integrity of the original data.

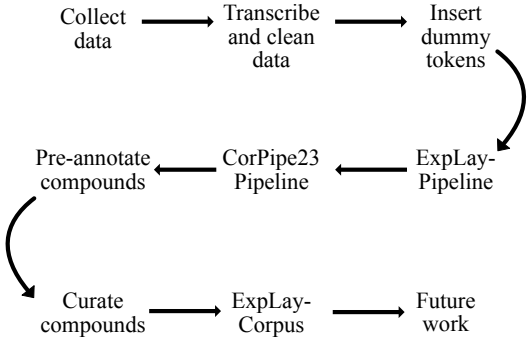

Figure 2: Workflow of ExpLay's curation process.

Cleaning involved the removal of filler particles and inaudible segments, which are excluded from the current release. Subsequently, the cleaned transcripts were manually annotated. Different dummy tokens (see Table 1 for details) were inserted to flag (1) ungrammatical structures that do not impede comprehension, (2) elliptical constructions typical of spontaneous speech or used for stylistic effect, and (3) elliptical references, such as nominalized adjectives. Category 3 tokens include the inferred original token in parentheses. Deleted structures are indicated with pipe symbols marking the start and end of the omitted span. Incomprehensible sentence parts (those severely ungrammatical to the point of impeding interpretation) were also marked. Although excluded from the parsed versions used for analysis, these segments are preserved in the unparsed data to support potential future research. Insertions are encoded using special characters that indicate the type of dummy-token (see Table 1). In the case of category 3 dummy-tokens, the original token is added in parenthesis after each insertion. Section 3.3 will show in more detail how those dummy-token insertions are handled in the pipeline, and Section 3.4 will show the different versions of the parse.

After annotating dummy tokens and incomprehensible structures, the transcripts are fed into the ExpLay-Pipeline described in Section 3.3. This pipeline performs automatic parsing and multi-level linguistic evaluation and is included as part of the ExpLay-Resource release. Subsequently, compound words were pre-annotated using a modified version of the Tuggener *compound-split* compound splitter (Tuggener, 2016) and then manually curated. In the final step, coreference annotation was conducted using the CorPipe23 system (Straka, 2023). An overview of the complete annotation and curation workflow of the ExpLay-Resource is illustrated in Figure 2, and Section 3.4 summarizes the resulting parsed data versions.

3.3 ExpLay-Pipeline

The ExpLay-Pipeline was implemented for the creation of the ExpLay-Corpus specifically and for the processing of expert-language data in general and is available in the repository. It is implemented in Python (Van Rossum and Drake, 2009), an untyped open-access programming language, and incorporates several state-of-the-art natural language processing systems (see Figure 4 for a depiction of the pipeline's architecture).

The ExpLay-Pipeline processes `.txt` files located in a designated directory, each containing curated transcripts that have undergone dummy-token annotation and the removal of ungrammatical structures (see 3.1). Meta-data of partici-

```
# sent_id = 0
# text = Zu sehen ist der Blick vom Bürgersteig aus auf eine Kreuzung .
# ['TOP', ['SINV', ['PP', ['IN', 'Zu'], ['NN', 'sehen']], ['VP', ['VBZ', 'ist']], ['NP', ['NP', ['DT', 'der'], ['NN', 'Blick']], ['PP', ['IN', 'von'], ['NP', ['NNP', 'dem'], ['NNP'
# ['TOP', ['SINV', ['PP', ['PART', 'Zu'], ['VERB', 'sehen']], ['VP', ['AUX', 'ist']], ['NP', ['NP', ['DET', 'der'], ['NOUN', 'Blick']], ['PP', ['ADP', 'von'], ['NP', ['DET', 'dem']
1       Zu      zu      PART    PTKZU   _       2       mark    _       -|start_char=0|end_char=2
2       sehen   sehen   VERB    VVINF   VerbForm=Inf    0       root    _       -|start_char=3|end_char=8
3       ist     sein    AUX     VAFIN   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   2       aux:pass        _       -|start_char=9|end_char=12
4       der     der     DET     ART     Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art      5       det     _       -|start_char=13|end_char=16
5       Blick   Blick   NOUN    NN      Case=Nom|Gender=Masc|Number=Sing        2       nsubj:pass      _       -|start_char=17|end_char=22
6-7     vom     _       _       _       _       _       _       -|start_char=23|end_char=26
6       von     von     ADP     APPR    _       8       case    _       ----
7       dem     der     DET     ART     Case=Dat|Definite=Def|Gender=Masc|Number=Sing|PronType=Art      8       det     _       ----|Entity=(c1--2
8       Bürgersteig     Bürgersteig     NOUN    NN      Case=Dat|Gender=Masc|Number=Sing        2       obl     _       -|end_char=38|Entity=c1)|start_char=27
9       aus     aus     ADP     APZR    _       8       fixed   _       -|start_char=39|end_char=42
10      auf     auf     ADP     APPR    _       12      case    _       -|start_char=43|end_char=46
11      eine    ein     DET     ART     Case=Acc|Definite=Ind|Gender=Fem|NumType=Card|Number=Sing|PronType=Art  12      det     _       -|end_char=51|Entity=(c2--2|start_char=47
12      Kreuzung        Kreuzung        NOUN    NN      Case=Acc|Gender=Fem|Number=Sing 2       obl     _       -|end_char=60|Entity=c2)|start_char=52
13      .       .       PUNCT   $.      _       2       punct   _       -|start_char=60|end_char=61
```

Figure 3: Exemplary parse of a sentence from participant P002/ stimulus B1. Note that the linear representation of the constituent trees was truncated for the illustration.

pants must be encoded in the filename in a fixed order using the format: participant-ID, gender, expert-status, stimulus-ID and language status (e.g. `P001_F_L_B1_M_.txt`). Each `.txt`-file in the directory is parsed individually, returning both individual and aggregate output statistics. During pre-processing, three versions of each transcript are created from each original `.txt` file: (1) A *raw-version* with all ungrammatical structures and dummy-tokens removed, (2) a *cleaned-corrected version*, which mirrors the raw-version but retaining the correction dummy-tokens and (3) an *all-dummy-version*, containing all dummy-tokens but excluding ungrammatical structures. To ensure compatibility with parsing tools, the pipeline removes the special character markers from the text-string and stores them as a separate object. Therefore, the original text production transcript itself cannot contain any of the special characters used to mark the dummy-tokens, as the pipeline would interpret those as dummy-token markers.

All three versions are then parsed using the stanza pipeline (Qi et al., 2020) applying the following processors: `tokenize, POS, lemma` and `depparse`. Stanza is an NLP toolkit that provides models for several different languages and a range of NLP tasks. The `POS` processor returns part-of-speech (POS) annotations and the `depparse` processor generates dependency annotations – both following the Universal Dependencies (UD) framework, which aims to standardize the format of various annotations, such as dependencies and POS-tags. The stanza pipeline returns the parsed data in the standardized `.conllu` format(Universal Dependencies Consortium, n.d.a), which is broadly supported by NLP tools. After parsing with stanza, the ExpLay-Pipeline re-introduces the dummy-token markers into the `.conllu` formatted parse by inserting the marker into the `MISC` column of the respective tokens in the `.conllu` file. This ensures that inserted tokens remain traceable for subsequent analysis.

Next, the parsed data is fed into the `Berkeley Neural Parser` (Kitaev and Klein, 2018), an NLP library providing state-of-the-art self-attentive language models for parsing various linguistic structures such as constituencies, which it returns in the form of an NLTK-tree object from the NLTK library (Bird et al., 2009). The parser uses the revised Penntreebank (PTB) tag-set of the English News Text Treebank (Bies et al., 2015) for the constituency nodes and the POS-tags. The multilingual model `benepar_en3` is used, as it is more robust than the German model and can also handle German data. After parsing the constituency structure of each version of a single production, the ExpLay-Pipeline creates a duplicate of the constituency tree and exchanges the revised PTB POS-tags for the upos-tags from the stanza-parse. This way, two trees are parsed, containing both sets of POS-tags. The trees are then stored as commentary lines between the sentence-ID and the parse in the `.conllu` format. Those are exported as `.conllu` files as single parses and added to a collective `.json` file containing the entire dataset of each version organized by the meta-data encoded in the filenames for easy access. For an exemplary parse of a sentence see Figure 3.

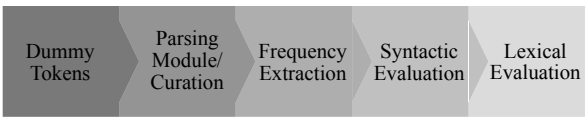

Figure 4: Architecture of the ExpLay-Pipeline.

Subsequently, the rawfile-version is passed to the frequency-extraction module of the pipeline, which collects various linguistic frequency measures both into single and collective `.csv` files. It collects simple surface measures such as word- and sentence-count and the usage of all POS-tags,

but also more linguistically complex structural measures from the constituency and dependency frameworks based on previous findings regarding the influence of those metrics on syntactic complexity. These structural measures include dependency distance (Gibson, 1998; Futrell et al., 2015), dependency and constituent-tree tree height (Yngve, 1960), dependency-based clauses count (Biber, 1988; Lu, 2011), and constituency-based phrase count (Lu, 2011). It should be noted, that due to the spontaneous, oral nature of the linguistic data, sentence boundaries, although defined as precisely as possible during transcription, should ultimately be regarded as approximations.

The extracted frequency data is then exported as both individual and aggregated files for further analysis. The aggregated data from the raw-version is then processed through the syntactic and lexical analysis modules of the pipeline, which utilize the libraries Pandas (pandas development team, 2020), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020) and Sklearn (Pedregosa et al., 2011). The syntactic module first assesses the normality of the data distribution using the Shapiro-Wilk test (Shapiro and Wilk, 1965) (see Equation 1). Depending on the outcome, statistical significance is evaluated using either a t-test for normally distributed data (Student, 1908) (see Equation 2 and 3) or the Mann-Whitney-U test (Mann and Whitney, 1947) (see Equation 4) for non-normally distributed data. It simultaneously tests for effect size using Cohen's delta (Cohen, 1988) (see Equation 5) if the data is distributed normally or a Rank-Biserial correlation (Cureton, 1956) (see Equation 7) for non-normal distributions.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{2}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{3}$$

Following these calculations, all metrics showing significant group differences are collected and normalized using Z-score standardization (see Equation 8), centering the data around a mean of 0 and a standard deviation of 1 while preserving the general shape of the distribution. Principal Component Analysis (PCA) (Jolliffe, 2002) (see Equations 9 and 10) is then performed, following the approach outlined in Park (2024) to assess the contribution of each metric to overall group variance.

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \tag{4}$$

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{5}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{6}$$

$$r_{rb} = 1 - \frac{2U}{n_1 n_2} \tag{7}$$

Principal Component loadings from the PCA, that represent linear combinations of the original metrics, are used to derive weights for a combined syntactic complexity score, which is likewise realized as a linear combination of the significant metrics.

$$Z = \frac{X - \mu}{\sigma} \tag{8}$$

$$Z = XW \tag{9}$$

$$PC_k = \sum_{i=1}^{n} w_i^{(k)} X_i \tag{10}$$

Then the module calculates a combined syntactic complexity score as a weighted sum of all the significant metrics normalized with min-max-normalization (see Equations 11 and 12) into a final dataset for a last test of normality, significance and effect-size as well as Pearson's r (see Equation 13) for a correlation between the PCA results and the combined syntactic complexity score.

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{11}$$

$$C = \sum_{i=1}^{m} w_i \cdot X_i \tag{12}$$

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{13}$$

In the final step, the lexical module of the ExpLay-Pipeline estimates lexical complexity by computing the frequency of open and closed word classes, following the approach of Shi and Lei

(2021), who (among other factors) investigated lexical complexity on the basis of word class in the context of social class differences — a framework also applicable to the study of expertise as a factor influencing language. The classification is based on the upos-tags from the Stanza parse, following the Universal Dependencies (UD) project (Universal Dependencies Consortium, n.d.b):

- **Open class or lexical words:** ADJ, ADV, INTJ, NOUN, PROPN, VERB
- **Closed class or grammatical words:** ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ
- **Other:** PUNCT, SYM, X

Mirroring the process of the syntactic module, the lexical module applies the same statistical procedures as the syntactic module to assess distribution (test for normality), significance, and effect size. Both modules export the results as `.csv` files to a results folder in the directory. In addition, the modules also generate various plots visualizing the significance tests outcomes and the PCA results. The graphics are exported to a plot folder inside the results folder using the Python libraries Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) for visualization.

3.4 ExpLay-Corpus

The resulting ExpLay-Corpus consists of three parsed versions per transcribed verbalization, corresponding to the three previously mentioned versions: The raw-version, the cleaned-corrected-version, and the all-dummy-version. These versions are stored in `.conllu` files, along with additional collective `.json` files containing the entire dataset. The results amount to three parses of the 130 texts and three files of the complete parse. Each individual parse consists of 11778 parsed tokens, derived from the raw-file version. Each of the three versions are enriched with two iterations of the constituency trees generated from the Benepar module, which are added before each sentence. The raw-file version was chosen for the evaluation modules as it best preserves the original text and includes minimal alternations, therefore providing a reliable basis for text-level comparison between the two groups. This choice can be manually adjusted should the application of the pipeline on future corpora require the evaluation of a different parse version.


```
{
    "id": [
        8
    ],
    "text": "Bürgersteig",
    "lemma": "Bürgersteig",
    "upos": "NOUN",
    "xpos": "NN",
    "feats": "Case=Dat|Gender=Masc|Number=Sing",
    "head": 2,
    "deprel": "obl",
    "start_char": 27,
    "end_char": 38,
    "misc": [
        "-",
        "[['Bürger', 'tail', '-'], ['Steig', 'head', 'Bürger']]"
    ]
},
```


Figure 5: Exemplary .json file entry from the production of P002/ stimulus B1 including the compound parse of the noun 'Bürgersteig' (Engl. sidewalk).

After parsing and evaluation with the ExpLay-Pipeline, the all-dummy-version was fed into the CorPipe23 (Straka, 2023) module for coreference parsing and pre-parsed using a derivation of the Tuggener (2016) *compound-split* compound splitter for compound words. The rationale behind this choice of parse iteration was that curation costs should be kept minimal, therefore only one of the parses should be annotated and curated for compound words. The all-dummy version was selected for compound word annotation to minimize curation efforts, as it can be easily mapped back to the raw-file version. The compound parse was then manually curated and stored in the `MISC` column of the respective token in the `.json` parse, using the format `'NoC'` for non-compound words or the pattern `'compound': [('first constituent', 'tail', '-'), ('second constituent', 'head', 'remaining part of compound', 'linking element')]` for compound words with two constituents (see Figure 5). This representation uses the maximum split approach and does not account for the branching direction in multi-constituent compounds.

4 Preliminary analysis of syntactic and lexical complexity

In a preliminary evaluation of the newly created corpus, the syntactic and lexical evaluation modules of the ExpLay-Pipeline were applied to the rawfile-parse of the corpus. This served two purposes: running a field test on the pipeline and the evaluation modules, as well as providing an initial exploration of the new resource.

4.1 Syntactic metrics

The previously described syntactical metrics evaluated in the pipeline include dependency distance, dependency and constituent tree height, dependency-based clause count, and constituency-based phrase count. Additionally, the pipeline also calculate surface measures such as sentence count and average tokens per sentence, but – as stated earlier – the annotated sentence boundaries should be considered with some reservations. For a complete display of the descriptive measures calculated for the ExpLay Corpus see Table 4 in Section A. The module then tests the data for normality, significance and effect size using the previously mentioned tests. Significant individual metrics are then combined into a combined syntactic complexity score. PCA is conducted on the chosen individual metrics and the resulting principal component loadings are used as weights for the combined score. Finally, a second round of normality, significance, and effect size tests is applied to the combined metric scores.

Metric	p-value	Cohen's d	RB
sent-count	0.75	0	0.03
tok-per-sent	0.25	0	-0.11
dep-dist	0.41	0.14	0
num-clauses	0.18	0	-0.14
dep-tree-height	0.31	0	-0.10
con-tree-height	**0.05**	0	**-0.02**
num-phrases	0.22	0	-0.13

Table 2: p-values, Cohen's d and Rank-Biserial correlation values of the single syntactic metrics before running the PCA.

4.2 Lexical metric

To calculate the lexical metric, the pipeline first calculates the count of open and closed word classes per text by adding up the counts of the single POS-tags per text according to the categorization of the UD-project. Then the same statistical tests as in the syntactic module are applied to those measures to test for normality, significance and effect size.

4.3 Results

Of the syntactic metrics analyzed for the 13 speakers per group reported in this paper, only constituent tree height showed a statistically significant group difference ($p < .05$), with a moderately small effect size. Experts exhibited slightly higher average tree heights than laypeople (see Table 4 in Section A), suggesting a tendency toward more deeply embedded, hierarchically complex sentence structures, in opposition to the laypeople's use of a slightly flatter syntax.

In contrast, surface-level syntactic features (e.g., sentence length, tokens per sentence) and lexical measures (e.g., distribution of word classes) did not differ significantly between groups, as can be seen in Table 2 for the significance values of the syntactic metrics, as well as in Table 3 for the evaluation of the lexical measures. Not only do the experts produce longer descriptions in general, they also display a slightly elevated use of open word classes compared to the laypeople, even though the differences did not turn out to be significant.

For a graphical visualization of the distribution of word classes among the two groups as well as for an exemplary output of the visualization module of the pipeline see also Figure 6. These features, however, are less sensitive to hierarchical syntactic depth as constituent tree height. The elevated tree height in expert speech points to denser phrasal layering, potentially reflecting more domain-specific and information-dense language use, in line with prior findings on expert discourse such as scientific writing. Laypeople on the other hand seem to use more shallow and linear constructions.

As no other syntactic measures reached significance, the combined syntactic complexity metric is identical to constituent tree height and is thus not reported separately.

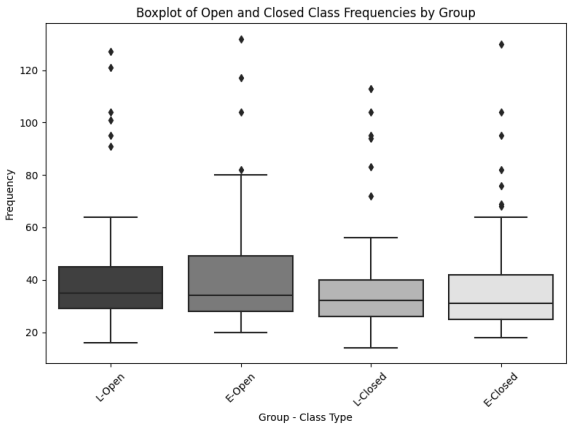

Figure 6: Boxplots of the descriptive values of the lexical metric.

4.4 Conclusion

The application of the ExpLay-Pipeline on datasets exports both individual and composite met-

Group	mean	sd	min	max	median	p-value	Cohen's d	RB
L-open	41.86	23.81	16.0	127.0	35.0	0.73	0	-0.04
L-closed	36.98	20.64	14.0	113.0	32.0	0.66	0.08	0
E-open	42.97	22.80	20.0,	132.0	34.0	**0.04**	0	**-0.02**
E-closed	38.62	21.99	18.0	130.0	31.0	0.12	0	-0.16

Table 3: Evaluation of the lexical metric.

rics, accompanied by normality assessments, significance tests, effect sizes, and Pearson correlations to assess group differences. The current paper's goal is primarily to showcase the range of syntactic and lexical measures the ExpLay-Pipeline can generate. We anticipate that increasing the participant number to at least 40 speakers per group in the future would enhance statistical power and reveal more differences between experts and laypeople.

These preliminary findings suggest that while both experts and non-experts use similar syntactic elements, they differ in the degree of syntactic complexity, with constituent tree height capturing features of structural depth possibly not reflected in other metrics. Given the exploratory nature of this initial analysis and the current limited number of speakers as well as the limitation to verbalizations of half of the described images, these results are not to be considered definitive. Future inclusion of the remaining parsed stimuli as well as more speakers will provide a more comprehensive basis for analysis.

However, this first evaluation offers initial evidence of domain-specific linguistic patterns in expert discourse in the domain of architecture in addition to the primary objective of this study: showcasing the functionality of the new pipeline. The observed increase in structural complexity (despite similar lexical and surface-level syntactic measures) raises the hypothesis of more complex linguistic structures in the expert population compared to the more linear constructions in the control group and consequently of a higher information density in expert language. This, in turn, opens up promising directions for future research, including semantic analyses and computational approaches of machine learning, to explore whether such structural differences persist across additional linguistic features.

Limitations

This study is limited in both its disciplinary scope and linguistic coverage: the data was collected for the specific domain of architectural expertise and in the German language, which may constrain the generalizability of the findings to other domains or languages. The current dataset includes speech from 26 participants (13 architects and 13 non-architects), each describing five stimuli. This relatively small sample size, along with the limited number of stimuli, restricts the statistical power and robustness of the analyses. Therefore, statistically significant results were not anticipated at this early stage. In addition to the limited sample size, the reduced level of expertise within the tested expert sample (that is more accurately characterized as a semi-expert group) must be taken into account. Future investigations may benefit from a follow-up study involving professional architects with greater practical experience. We expect that the effects observed in the preliminary present evaluation would be more pronounced with participants exhibiting a higher degree of domain-specific expertise.

While the manual pre-processing and annotation of the data were conducted with care, inter-annotator agreement was not assessed, which may introduce some degree of variability. Additionally, the annotation decisions, mirrored in the code and detailed documentation provided, rely on a specific theoretical framework, which may not align with all linguistic traditions. Future work will aim to expand the dataset substantially and to incorporate reliability measures to strengthen the generalizability and replicability of the findings.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments. This research is partially funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102. Figure 1 includes elements generated using DALL·E 3, which were subsequently modified by the authors.

References

D. Banks. 2003. The evolution of grammatical metaphor in scientific writing. *Amsterdam Studies in the Theory and History of Linguistic Science Series*, 4:127–148.

D. Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

D. Biber and B. Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press.

D. Biber, B. Gray, and K. Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *TESOL Quarterly*, 45(1):5–35.

A. Bies, J. Mott, and C. Warner. 2015. *English News Text Treebank: Penn Treebank Revised LDC2015T13. Web Download*. Linguistic Data Consortium, Philadelphia.

S. Bird, E. Klein, and E. Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences. 2. Auflage*. Lawrence Erlbaum Associates, Hillsdale.

E. E. Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

S. Degaetano-Ortlieb and A. Piper. 2019. The scientization of literary study. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 18–28, Minneapolis, USA. Association for Computational Linguistics.

S. Degaetano-Ortlieb and E. Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

S. Degaetano-Ortlieb and E. Teich. 2022. Toward an optimal code for communication: The case of scientific english. *Corpus Ling.. Ling.. Theory*, 18(1):175–207.

K. Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

K. Dobrovoljc and J. Nivre. 2016. The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

A. Ersin, H. Ceren Tezeren, N. Ozunlu Pekyavas, B. Asal, A. Atabey, A. Diri, and İ Gonen. 2022. The relationship between reaction time and gaming time in e-sports players. *Kinesiology*, 54(1):36–42. Doi:10.26582/k.54.1.4.

R. Futrell, K. Mahowald, and E. Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

J. Gamboa, K. Braun, J. Järvikivi, and S. E. M. Allen. 2025. The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis. *Corpus Linguistics and Linguistic Theory*, 21(1):137–171.

E. Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

M. A. K. Halliday. 1988/2004. On the language of physical science. In Jonathan J. Webster, editor, *The Collected Works of M. A. K. Halliday (Vol. 5)*, pages 140–158. Continuum, London and New York.

C. R. Harris, K. Jarrod Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, and 7 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

D. N. Jakobi, T. Kern, D. R. Reich, P. Haller, and L. A. Jäger. 2024. Potec: A german naturalistic eye-tracking-while-reading corpus. *Preprint*, arXiv:2403.00506.

C. Jiang, A. Kundu, and M. Claypool. 2020. Game player response times versus task dexterity and decision complexity. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '20, page 277–281, New York, NY, USA. Association for Computing Machinery.

I. T. Jolliffe. 2002. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York. Doi:10.1007/b98835.

T. S. Juzek, M.-P. Krielke, and E. Teich. 2020. Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119, Barcelona, Spain (Online). Association for Computational Linguistics.

N. Kitaev and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.

S. C. Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge University Press.

X. 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.

E. A. Maguire, D. G. Gadian, I. S. Johnsrude, C. D. Good, J. Ashburner, R. S. J. Frackowiak, and C. D. Frith. 2000. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8):4398–4403.

A. Majid, M. Bowerman, S. Kita, D. B. Haun, and S. C. Levinson. 2004. Can language restructure cognition? the case for space. *Trends in Cognitive Sciences*, 8(3):108–114.

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

H. Mertins, R. Delucchi Danhier, B. Mertins, A. Schulz, and B. Schulz. 2020. The role of expertise in the perception of architectural space. In C. Leopold, C. Robeller, and U. (Hrsg. Weber, editors, *Research Culture in Architecture*, pages 279–288. Birkhäuser, Basel.

The pandas development team. 2020. pandas-dev/pandas: Pandas.

A. Papafragou, J. Hulbert, and J. Trueswell. 2008. Does language guide event perception? evidence from eye movements. *Cognition*, 108(1):155–184.

S. Park. 2024. Identifying key linguistic variables of second language speaking proficiency using principal component analysis. *Forum for Linguistic Studies*, 6(6):623–633.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

D. Rebholz-Schuhmann, A. J. Jimeno-Yepes, E. M. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, and U. Hahn. 2010. The calbc silver standard corpus for biomedical named entities- a study in harmonizing the contributions from four independent named entity taggers. In *Nicoletta Calzolari (Conference Chair), et al*, Valletta, Malta. European Language Resources Association (ELRA. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

D. Roberson, I. Davies, and J. Davidoff. 2000. Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3):369–398.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611.

Y. Shi and L. Lei. 2021. Lexical use and social class: A study on lexical richness, word length, and word class in spoken english. *Lingua*, 262:103155.

D. I. Slobin. 1996. From "thought and language" to "thinking for speaking". In J. J. Gumperz and S. C. Levinson, editors, *Rethinking linguistic relativity*, pages 70–96. Cambridge University Press.

M. Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

E. Teich, P. Fankhauser, S. Degaetano-Ortlieb, and Y. Bizzoni. 2021. Less is more/more diverse: On the communicative utility of linguistic conventionalization. *Frontiers in Communication*, 5.

D. Tuggener. 2016. *Incremental Coreference Resolution for German*. Phd thesis, University of Zürich, Zürich, Switzerland.

Universal Dependencies Consortium. n.d.a. Universal dependencies documentation: Format. `https://universaldependencies.org/format.html`. Accessed: 2025-04-06.

Universal Dependencies Consortium. n.d.b. Universal dependencies: Part-of-speech tags. `https://universaldependencies.org/u/pos/index.html`. Accessed: 2025-04-06.

G. Van Rossum and F. L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett,

J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

M. L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

J. J. Webster. 2018. *18. The Language Of Science – A Systemicfunctional Perspective*, pages 345–363. De Gruyter Mouton, Berlin, Boston.

B. L. Whorf. 1956. *Language, Thought, and Reality*. Cambridge, Ma.

J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

V. H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

A Appendix

Metric	mean	sd	min	max	median
L-sent-count	7.58	3.96	3.0	29.0	7.0
L-tok-per-sent	11.82	2.49	7.17	18.33	11.57
L-dep-dist	2.74	0.36	1.89	3.62	2.77
L-num-clauses	0.26	0.21	0.0	1.0	0.25
L-dep-tree-height	2.62	0.46	2.0	4.67	2.6
L-con-tree-height	7.34	0.66	6.17	10.0	7.17
L-num-phrases	20.97	4.23	13.0	32.33	20.57
E-sent-count	7.48	3.71	3.0	24.0	7.0
E-tok-per-sent	12.53	2.92	7.88	20.0	12.2
E-dep-dist	2.79	0.36	2.1	3.73	2.72
E-num-clauses	0.33	0.26	0.0	1.0	0.27
E-dep-tree-height	2.66	0.35	2.13	4.0	2.6
E-con-tree-height	7.56	0.71	6.38	9.25	7.5
E-num-phrases	22.25	5.08	14.5	36.33	21.25

Table 4: Descriptive values of the syntactic metrics.