IDS
OPEN ONLINE-ONLY PUBLIKATIONEN
DES LEIBNIZ-INSTITUTS FÜR DEUTSCHE SPRACHE

The logical architecture of CoMParS and its XML implementation

Piotr Bański

Abstract CoMParS (Collection of Multilingual Parallel Sequences) is a project under way at the Leibniz Institute for the German Language (IDS) in Mannheim, Germany. CoMParS runs in the context of the long-term project GDE (Grammatik des Deutschen im europäischen Vergleich; German Grammar in European Comparison). The goal of the GDE is to create a novel contrastive grammar of German vis-à-vis other European languages. Alongside German, which is the central focus, the core languages for comparison are English, French, Hungarian and Polish, representing different typological classes.

The primary practical aim of CoMParS is to serve as an electronic extension of printed contrastive grammars created as deliverables of GDE. A more far-reaching goal is to lay the foundation for a flexible multilingual exploratory resource, capable of supplying new contrastive generalizations.

The present submission looks at the logical architecture of CoMParS and at selected details of the XML implementation.

Keywords parallel corpus, treebank, GDE-V, contrastive studies, CoMParS

Inhalt

1. Introduction, goals and assumptions[1]

CoMParS (Collection of Multilingual Parallel Sequences) is a project under way at the Leibniz Institute for the German Language (IDS) in Mannheim, Germany. CoMParS runs in the context of the long-term project GDE (Grammatik des Deutschen im europäischen Vergleich; German Grammar in European Comparison). The goal of the GDE is to create a novel contrastive grammar of German against the background of other European languages. The first phase of the project (GDE-N), running from 2001 to 2013, was devoted to the nominal domain. In the second phase (GDE-V), which began in 2013, the verbal domain is the subject of investigation. Alongside German, which is the central focus, the core languages for comparison are English, French, Hungarian and Polish, representing different typological classes.

The primary practical aim of CoMParS is to serve as an electronic extension of printed contrastive grammars created as deliverables of GDE. Thanks to CoMParS, the target users of GDE grammars will be able to access a larger number of examples than can be fit into the physical limits of a printed work and to access the counterparts of these examples in selected languages (ideally, all the languages in the scope of GDE, but minimally the languages targeted by the particular book publication). A more far-reaching goal for CoMParS is to lay the foundation for a flexible multilingual exploratory resource, capable of supplying new contrastive generalizations.

In order to achieve both ends as well as to ensure the longevity and interoperability of the data, and – on a more political note – to secure a measure of recognizability in the field, it makes a lot of sense to base CoMParS on the existing annotation standards and best practices, including those in the process of attaining formal certification at recognized standardization bodies, in our case ISO. For this reason, CoMParS follows the current and upcoming recommendations of ISO Technical Committee 37, Subcommittee 4 (ISO TC37 SC4) "Language resource management", in particular what is commonly known as the "ISO-LAF family of standards", i. e., standards based on the Linguistic Annotation Framework (ISO 24612:2012).[2] A firm basis for interoperability is also provided by the Text Encoding Initiative (TEI) Guidelines (TEI Consortium 2017), especially the proposals advanced in the Special Interest Group "TEI for Linguists". Details of the implementation will be presented in sections that follow.

CoMParS cannot be described with a single label commonly used for coarse identification of the nature of language resources. It can be viewed as a federation of monolingual corpora, which nevertheless is able to function as a parallel corpus. It can be viewed both as a corpus in the fairly traditional sense (roughly, an assembly of sentences selected according to predefined criteria), but it can also be viewed as a treebank, that is, a collection of syntactic trees or dependency graphs. For this purpose, the terms "collection" and "sequences" have been used in the name of this resource, in order to encompass all the aspects of its multivariate nature. While, in most cases, the basic textual unit stored in CoMParS is a sentence, we use the term "sequence" in order to cover cases of one-to-many or many-to-many relationships between sentences in the parallelized data from several languages.

1 The present report describes the CoMParS project as of April 2017. The project has, naturally, evolved since then – see Trawiński/Schlotthauer/Bański (2021) for a glimpse at its current state. I would like to warmly thank an anonymous reviewer for very valuable remarks that led to an improvement of this submission.

2 For a publicly accessible description of the standard, see Ide/Suderman (2014). For a comparison with other existing standards, see Eckart de Castilho et al. (2017).

CoMParS does not aim at creating its own set of accompanying tools for searching and visualisation. Instead, the basic but powerful search functionality will be offered by a native XML database once the initial dataset has been established, and once the details of encoding and implementation have been reasonably fully determined. At this point, CoMParS has a set of tools in the form of XSLT scripts that convert between its native format and the format of designated search tools.

The sections that follow look at the logical architecture of CoMParS (Section 2) and present details of the XML implementation (Section 3). Section 4 looks at the tool system, and Section 5 wraps up the presentation and outlines the intended further development.

2. Logical structure of CoMParS

The fundamental division in the architecture of CoMParS splits it into the monolingual part, comprising all the language data separated according to the language, and the aligned part, which encodes the correspondences and is the basis for applications in contrastive studies.

This is illustrated in the diagram below, where the difference between the formal and functional sections of the aligned part has been slightly exaggerated for the purpose of presentation and is described in greater detail in Section 2.2.

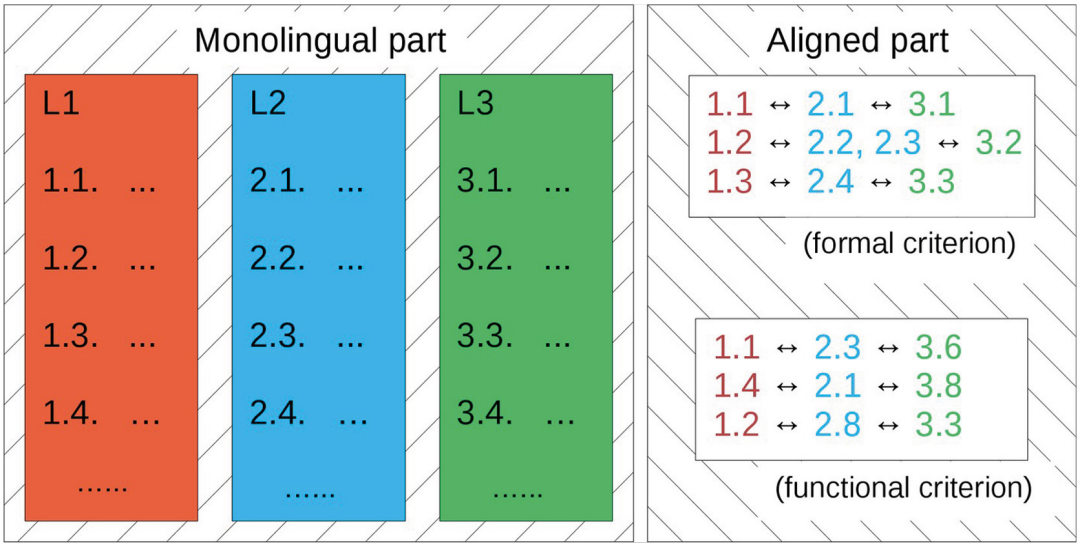


Figure 1: Basic divisions in the data architecture of CoMParS

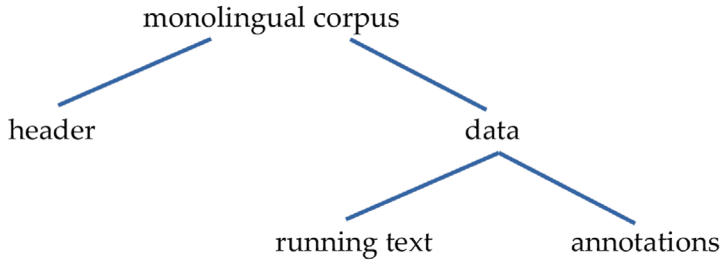
Other major components of CoMParS are the header system, which stores most of the metadata concerning, among others, the data sources, persons responsible for the creation of various parts of the resource (including translations and data sources), and the list of modifications. The extended CoMParS architecture also includes export and import tools, documentation, and feature structure declaration components. See also Figure 5 for a glimpse of the implementation of aforementioned divisions.

The remainder of this section looks at the monolingual and aligned parts in turn, before we move to examine some details of the XML implementation in Section 3.

2.1 Monolingual part

The monolingual part of CoMParS currently consists of five subcorpora, bound into a single system both "physically" (by means of XML inclusions) and logically, by keeping the major data parts aligned from the moment they are input into the system.

A single corpus has the hierarchical structure diagrammed below.



The header is the place for all formal corpus metadata including, among others, the change log, the sources for each data sequence, and the record of corpus queries used for deriving the data (if available). The data part stores the individual sequences as running text, and annotations that describe these sequences. The annotations consist of:

- (token- and chunk-level) segmentation,
- identification of word-forms (syntactic terminals) and their description (lemmatization, part-of-speech labels, morphosyntactic features),
- syntactic annotation (any number of layers, each of which references the word-form layer),
- semantic, functional, and topological annotation (any number of layers referencing the word-form layer and/or one of the syntactic annotation layers),
- "utility" annotation (glosses, grammatical judgements, blanks – for indicating parts of sequences that are retained from the original data and which form the context of the entire utterance, but are not important for the purpose of illustrating the grammatical phenomena described in the given GDE grammar); this kind of annotation can reference both the segmentation layer and the word-form layer.

We will now briefly look at each layer of annotation in turn, starting with segmentation, which is expected to list all text segments that are used by other annotation layers. This commonly means text tokens, that is sequences of characters divided by whitespaces and punctuation marks, but it is also possible to list morphemes (understood as character sequences below the intuitive word level, as in, e. g., "ver", "lieb", "t" being a possible series of segments next to the token "verliebt"), multi-word expressions or even phrases. This practice of listing all possible textual segments of any grade inside a single layer follows the practice of the XStandoff approach (Stührenberg/Jettka 2009). This approach allows for handling many cases of potential disagreements among annotation systems concerning the coarseness of tokenization (cf. also Chiarcos/Ritz/Stede 2012).

The layer of word-forms or syntactic terminals may be perceived as the most complex among the current CoMParS annotations, because of the amount of information packaged into the individual data components. This is because of the pivotal nature of this part: it is the foundation of most of the higher-level annotation layers and at the same time the storage place for much of the information presented to the putative user of the online extension of GDE grammars. Apart from the usual information needed for the purpose of

constructing, for example, a syntactic tree over a sequence of terminals, namely the lemma, part-of-speech (henceforth POS) and possibly other morphosyntactic information such as the identification of the grammatical case, gender, person, etc., this layer also stores user-friendly grammatical labels, potentially in all CoMParS languages, and minimally in German.

Recasting the view onto the perspective of a speaker of English understood as a target user of a GDE grammar, let us consider the pair of sentences below.

(1) a. Ich habe mich im Winter in dich verliebt
b. I fell in love with you last winter

An average English target user need not be interested in the fact that a typical set of POS symbols used for the description of German, STTS (Stuttgart-Tübingen-Tagset),[3] provides labels such as PPER for *Ich* and *mich*, VAFIN for *habe*, and APPRART for *im* – these labels are necessary for syntactic parsers of German to construct a constituency tree or a dependency graph over the example sentence, but are not user-friendly in any way, and arguably not even informative for someone who is not a computational linguist. The example English-speaking user would rather expect labels such as "Pronoun", "Auxiliary" and "Preposition" (or possibly "Prep+Det", in the case at hand) for the initial words of (1a), and similar labels for the initial words of (1b) rather than those provided by the CLAWS tagset (a common counterpart of STTS used in English corpus linguistics).[4] Furthermore, in keeping with the current goal of applying a maximally uniform syntactic description to all the languages represented in CoMParS, by means of Universal Dependencies (henceforth UD, see Nivre et al. 2016), terminals for (1a) have to be equipped with (i) STTS labels, (ii) UD labels, (iii) user-friendly labels, potentially in more than one language. All this information has to be packaged in such a way that a grammatical parser of German constituency or dependency grammar sees the STTS labels alone, a UD parser sees the UD labels, and the visualisation system for the target user is capable of presenting both the user-friendly labels (in the language of choice) as well as the labels used by the particular grammatical parsing system, in case the user chooses to see the corresponding syntactic annotation. A technical implementation of this will be exemplified in Section 3.

The next variety of grammatical description envisioned in CoMParS is that of syntactic analysis. An assumption from the initial stage of the project was that each language would receive its own layer of automatic syntactic annotation performed by tools specific to that language, possibly the most popular tool freely available, and allowing the resulting annotations to be shared with no legal restrictions or with only minimal restrictions. This, however, may become untenable and difficult to maintain, especially when the number of languages served by CoMParS grows. For this reason, since the early 2017, plans for CoMParS involve a simple but potentially uniform solution (or, more precisely, a solution involving a large degree of uniformity and minimal language-specific modifications), namely a dependency grammar system known as Universal Dependencies.

Universal Dependencies can be considered a common sense, pragmatic attempt at postulating a grammatical description system that is general enough to embrace multiple languages inside the same formalism (allowing for restricted and well-defined language-specific exten-

3 STTS Guidelines and supporting documents are available from www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/ (last accessed 26-3-2025).

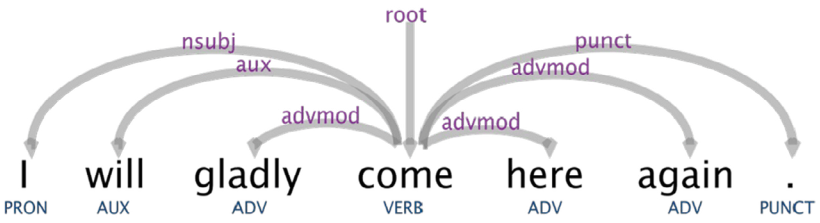
4 The relevant symbols for the initial three words in (1b) would be, respectively, "PNP", "VVD", and "PRP" in the simple tagset known as CLAWS-5, used for the annotation of the British National Corpus, and "PPIS1", "VVD", and "II" in the currently more popular tagset known as CLAWS-7. See http://ucrel.lancs.ac.uk/claws/ (last accessed 26-3-2025) for more details.

sions) and at the same time complex enough to offer contrastive-linguistic insights. As the creators of the system state (http://universaldependencies.org/introduction.html, last accessed 26-3-2025):

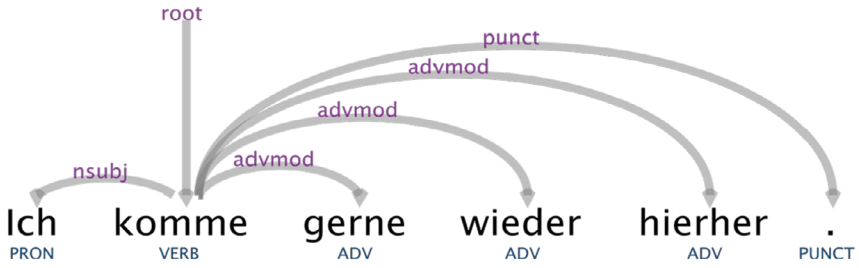
1. UD needs to be satisfactory on linguistic analysis grounds for individual languages.
2. UD needs to be good for linguistic typology, i. e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent annotation by a human annotator.
4. UD must be suitable for computer parsing with high accuracy.
5. UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a *habitable* design, and it leads us to favor traditional grammar notions and terminology.
6. UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …).

As a dependency formalism, UD may require some getting used to, especially for researchers customarily using constituency approaches. The following two examples illustrate relatively typical, simple dependency structures in English and German, parsed with UDPipe (Straka/Hajič/Straková 2016) and visualised via the Tündra system.

(2) a.



b.



As can be seen in the examples, UD aims for simple, intuitive functional labels from a restricted repertoire of values.[5]

Apart from lexical and syntactic description, each sequence in CoMParS may be accompanied by any number of layers of semantic, functional, and topological description. While we await a crystallization of the full typology of Functional Domains and their parameters (see Kutscher 2014), we are considering annotating CoMParS with German Framenet, for the purpose of marking the cross-linguistic similarities and differences while rooting the system in German as a point of departure. An alternative would be to attempt to annotate

5 A thorough description of the values of the POS, feature, and relational labels used in UD can be accessed at http://universaldependencies.org (last accessed 26-3-2025). The visualisation has been exported from TüNDRA (Tübingen aNnotated Data Retrieval Application Web tool for treebank research), available at https://weblicht.sfs.uni-tuebingen.de/Tundra/ (last accessed 26-3-2025).

each monolingual corpus with the matching variety of Framenet. This, however, raises several issues: firstly, the process of annotation of each corpus separately would require more manpower than is currently available for this purpose. Secondly, the resulting Framenets would have to be aligned within CoMParS, which demands expertise in all of them and would essentially turn the project in question into a cross-Framenet study, and finally, given the GDE's goals, German is the pivot language for contrastive studies, and assuming the German Framenet throughout will address the issues of potential contrasts *directly* and will allow them to be stated within the same formalism. Other cross-linguistic semantic annotation frameworks will also be considered.

A basic example of Functional Domain annotation by (ab)using the Framenet format is shown in Figure 2, which is a screenshot of example (1b) exported from an early version of CoMParS (currently dubbed "vanilla" and exemplified in Section 3) into the SALSA format (Erk/Padó 2004) and visualised in the Salto tool (Burchardt et al. 2006).

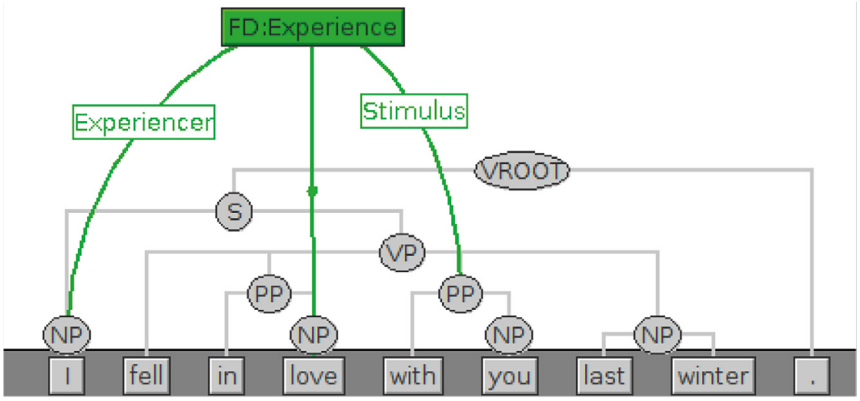


Figure 2: Experimental Functional Domain annotation example visualised in Salto. *Love* is treated as the trigger of the FD frame, with the subject *I* fulfilling the role of the Experiencer and the phrase *with you* being the Stimulus.

The last kind of annotation that may potentially appear here can be referred to as "utility annotation": assuming that this kind of annotation only targets the segmentation information, it can provide additional glosses to be displayed under the segments, it can target parts of examples that should be hidden from the display (or greyed out, as less relevant in the given context), and finally, it can encode grammaticality judgements.

2.2 Alignment information in CoMParS

In the monolingual part of CoMParS, the sequences in the individual corpora carry no information on how they are aligned with other sequences – they merely expose targets for alignment. This way, the maintenance and extension of alignment information can be located in a single place in the architecture, where bundles of pointers to various parts of the monolingual description are stored. If another monolingual corpus is added or if a set of new multilingual examples is added to the collection, the alignment part is the only place that needs to be updated or extended.

Unlike traditional contrastive grammars available for German, which usually cover language pairs, namely German and one another language, and are based on the classical parts of speech and grammatical categories, the GDE grammar is developed in the spirit of functionalist typology. This implies that, instead of formal criteria, cognitively motivated functional domains in terms of Givón (1984) are used as *tertia comparationis* (cf. Kutscher

2014; Trawiński 2016). In order to reflect this theoretical stance, CoMParS provides two kinds of alignment sets: the traditional formal alignment among individual sequences (potentially down to word-level alignment, which is not implemented at this time and low on the priority list) as well as onomasiological alignment, ultimately based on the content of Functional Domain annotations. This can be illustrated with a recourse to the already adduced examples from early stages of CoMParS, visualised in the Salto tool, where examples (1a) and (1b), together with their corresponding Polish sequence *Zakochałam się w tobie zeszłej zimy* have been annotated with the then-current state of Functional Domains, where the domain Experience with the sub-parameter LOVE is identified. Next to the English example (1b) diagrammed in Figure 2 above, the following correspond to the German (1a) and the Polish counterpart:

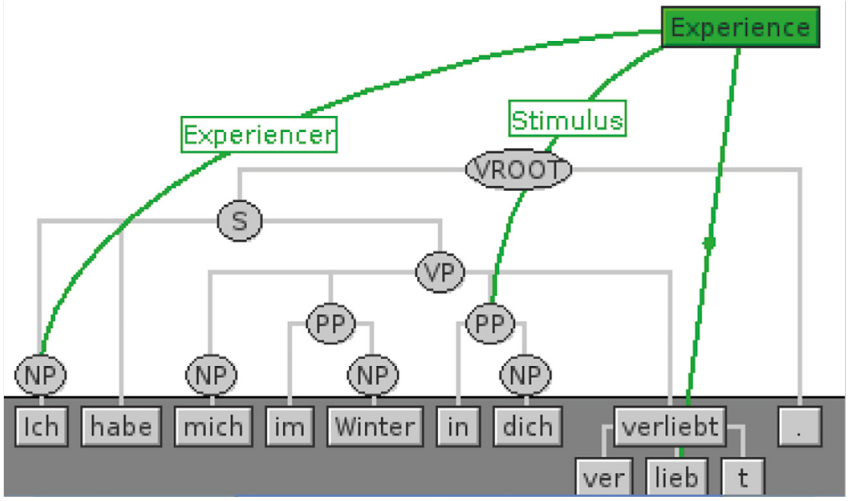


Figure 3: Salto visualisation of example (1a) with Functional Domain annotation. The participle *verliebt* is further analysed and the root *lieb* is annotated as the actual trigger of the frame.

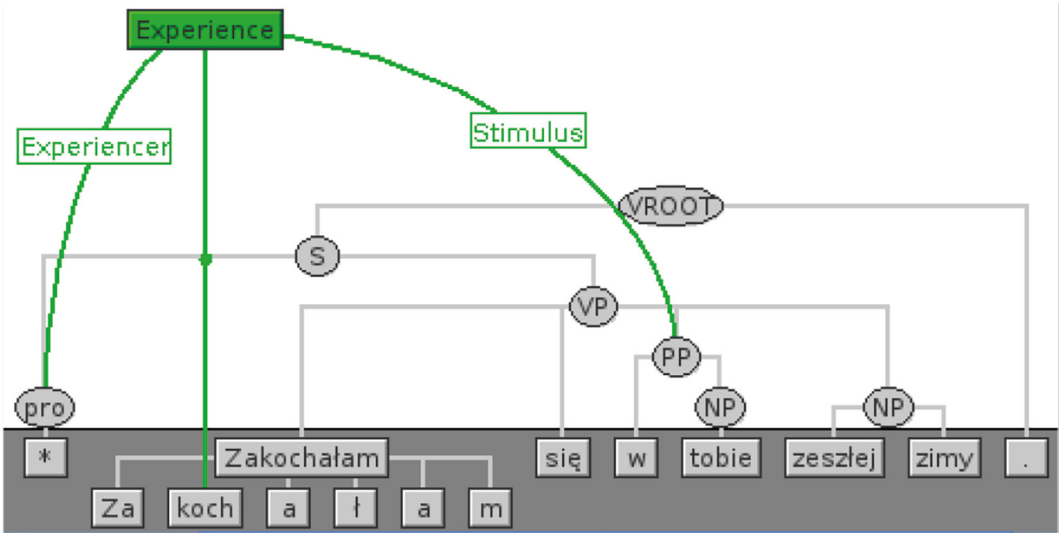


Figure 4: Salto visualisation of the Polish counterpart to examples in (1), with prototype Functional Domain annotation. The past-tense *zakochałam* (*się*) is further analysed and the root *koch* is annotated as the actual trigger of the frame. The optional subject marked as *pro* does not signal a theoretical commitment but rather a visualisation strategy.

The above-mentioned formal alignment section targets the entire sequences and may attain greater granularity by aligning the individual phrases as well as terminals. The con-

cept (onomasiological) alignment section looks at the Functional Domain annotation and groups all sequences that, at the highest degree of abstraction, match, in each language, the domain "Experience" and within this group those that are triggered by a root paired with the concept LOVE, and so on, across all the parameters defined for the domain in question. Combinations of the parameters chosen can then be obtained by the usual set-theoretic operations.

3. XML implementation of the CoMParS data model and architecture

CoMParS is implemented as a collection of documents stored in an IDS-internal Subversion (SVN) repository, which allows for tracking changes to the individual files as well as the authorship of these changes. A snapshot of the production area of the SVN repository is presented below and will be briefly described in what follows.

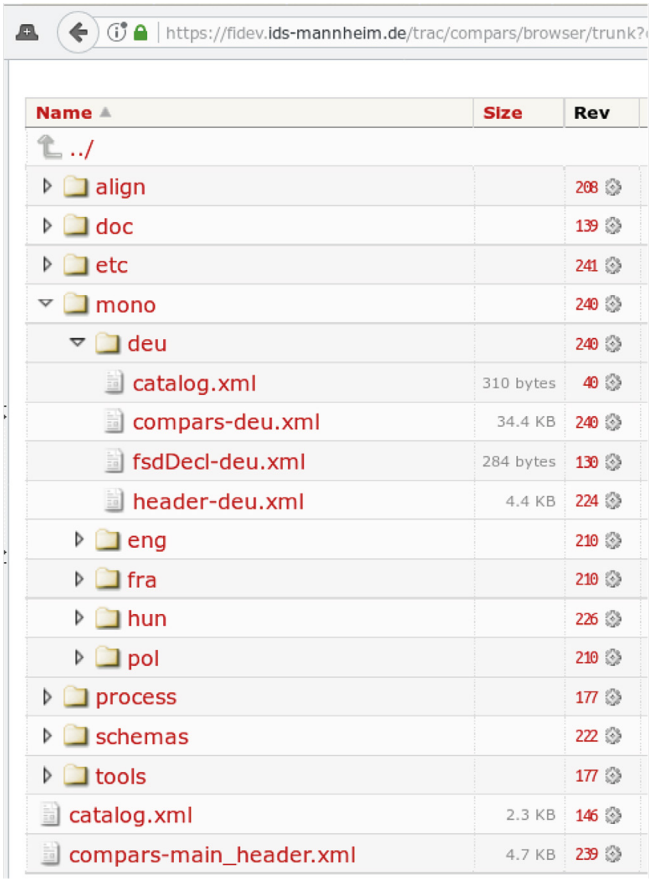


Figure 5: A snapshot of the CoMParS SVN repository

Files in the root directory are the main header, containing metadata relevant to the entire resource, and an "XML catalog" which acts as a table of contents for various tools that CoMParS is processed with: whenever a tool *fails* to locate a necessary document (for example, a schema that is necessary for validating the document grammar of a monolingual corpus file) but sees a `catalog.xml` file instead, it consults this file in order to look up the location of the schema. If the operation takes place in the directory `mono/deu/` (open in Figure 5), the tool first encounters a statement

```
<nextCatalog catalog="../../catalog.xml"/>
```

which brings it to another instance of `catalog.xml` located two levels up ("`../../`"), that is, in the main corpus directory. In the main `catalog.xml` file, the schema reference is located in the line:

```
<systemSuffix systemIdSuffix="compars-mono.rng"
              uri="schemas/compars-mono.rng"/>
```

which instructs the tool to search for the schema (`compars-mono.rng`) in the subdirectory `schemas/`.

Looking from the top of Figure 5, the subparts of CoMParS are as follows:

- the `align/` directory stores the alignment information in the form of two XML documents, `compars-align-concept.xml` for onomasiological alignments, and `compars-align-form.xml` for formal alignments (recall that these are currently restricted to the coarse alignment among the particular sequences);
- the `doc/` directory stores low-level corpus documentation (currently a `README` file);
- the `etc/` directory contains various utility documents, such as example annotations used as tests for the corpus tools, stored Weblicht pipeline definitions, and, crucially, a set of ODD documents that define both the document grammar of a TEI XML file as well as provide extensive documentation on the TEI XML vocabulary that can be used in the corpus;
- the `mono/` directory stores the individual monolingual corpora, each in a subdirectory marked by a 3-letter ISO 639-3 language code; inside the `deu/` directory, the `compars-deu.xml` file is one that holds all the German-language primary data and their annotations, the `header-deu.xml` file contains metadata pertaining to this monolingual corpus alone, and the `fsdDecl-deu.xml` file is a placeholder for Feature Structure Declaration, written in the ISO 24610-2:2011 schema-definition language that constrains Feature Structure Descriptions (exemplified later in this section); because the feature-structure description of CoMParS still evolves, codifying it inside a `fsdDecl.xml` document makes no sense at this stage;
- the `tools/` directory contains XSLT scripts that transform data from the CoMParS format into a format understood by various external tools, or conversely, transform the output from external tools into the CoMParS format; these intermediate files are located in the `process/` directory;
- the `process/` directory is an interface between CoMParS and various tools into which CoMParS data can be exported, and whose output data can be imported into CoMParS.

CoMParS is encoded in a variant of XML defined according to the Text Encoding Initiative (TEI) Guidelines (TEI Consortium 2017). The name "TEI" has evolved from a name of a project set up in the late 80's, aiming to create a well-designed and documented set of guidelines for literary and linguistic encoding, into a consortium of mostly universities and libraries and into a broad community of users, who together maintain a set of open-source specifications and tools that make it possible to adequately encode practically any kind of text (including ancient inscriptions and modern speech transcripts) in such a way as to describe its structure (including potential variant interpretations thereof) and expose its content for the purpose of analysis, cross-linking and document interchange. The TEI community is organised, among others, into Special Interest Groups (SIGs), the so-called

'LingSIG' also called "TEI for Linguists"[6] has worked since 2010 on making the TEI attractive and usable in the area of modern corpus linguistics, e-lexicography, and, generally, language-resource modelling. CoMParS annotation stems from some of that work and also serves as a reference implementation of a subset of standards codified at ISO TC37 SC4.[7]

The TEI LingSIG follows the ISO proposals in moving towards standoff annotation, i. e., a kind of annotation where the source text is kept maximally "pristine" (with preferably no annotations intervening in the original text stream, and instead being kept separate, while addressing the original text by means of character offsets and/or ID attributes). Within the TEI, after several years of research and debate (see e. g., Bański 2010; Pose/Lopez/Romary 2014; Bański et al. 2016) the approach began to crystallize into a concise proposal, currently hosted at https://github.com/laurentromary/stdfSpec (last accessed 26-3-2025). Since December 2016, CoMParS has moved towards this style of annotation as well, in this way becoming a reference resource for a new LingSIG-related proposal to the TEI Technical Council (the elected body that decides about the development of the TEI Guidelines), and at the same time a reference implementation of a new work item at ISO TC37 SC4, namely the 3rd part of a multi-part standard known as "Syntactic Annotation Framework".

The pre-standoff version of CoMParS, preserved under the tag "vanilla" in the Subversion repository, has been built as a proof-of-concept showing that complex syntactic annotation can, with minimal amount of compromise regarding the interpretation of TEI Guidelines, be fit into a regular TEI document, i. e., be placed under the <text> element.[8] A reduced version of that representation (in fact underlying, e. g., the Salto visualisation of example (1a)), is presented below.

```
<body>
  <div type="sequence" xml:id="deu-s_001">
    <div type="graph" xml:id="deu-s_001-g1" n="1">
      <ab type="terminals">
        <seg type="t" xml:id="deu-s_001_01">
          <fs type="t">
            <f name="word">Ich</f>
            <f name="lemma">ich</f>
            <f name="pos">PPER</f>
            <f name="morph">1.Sg.Nom</f>
          </fs>
        </seg>  (…)
      </ab>
      <ab type="nonterminals">
        <seg type="nt" xml:id="deu-s_001_502">
          <fs type="nt">
            <f name="cat">NP</f>
          </fs>
          <ref target="#deu-s_001_01" type="edge">
            <fs type="edge">
               <f name="label">NK</f>
            </fs>
          </ref>
```

6 More information on the LingSIG can be accessed at https://github.com/LingSIG (last accessed 26-3-2025).

7 See www.iso.org/committee/297592/x/catalogue/ for the current repertoire of specifications offered by TC37SC4 (last accessed 26-3-2025).

8 This is not to say that pre-2017 CoMParS was the first such application ever. The annotation of the National Corpus of Polish, which around the year 2010 was probably the most robust linguistic resource encoded in the TEI (cf. e. g., Bański/Przepiórkowski 2009; Przepiórkowski 2009, and the samples at http://nlp.ipipan.waw.pl/TEI4NKJP/, last accessed 26-3-2025), also used the standard TEI textual structure, with a massive amount of ISO/TEI Feature Structures (ISO 24610-1:2006). However, CoMParS was most probably the first application of "vanilla TEI" that mimicked syntactic annotation originating in Tiger XML (Mengel/Lezius 2000) and later encoded in the ISO SynAF family of specifications.

```
        </seg>    (…)
      </ab>
    </div>
    <div type="functional">
      <ab type="frames">
        <seg type="fd" xml:id="deu-s_001_f1">
          <label>Experience</label>
          <ref type="trigger" target="#deu-s_001_08_s1">
            <fs type="trigger">
               <f name="sem">LOVE</f>
            </fs>
          </ref>
          <ref type="fe" xml:id="deu-s_001_f1_e1" target="#deu-s_001_08">
            <label>Experiencer</label>
          </ref>
          <ref type="fe" xml:id="deu-s_001_f1_e2" target="#deu-s_001_501">
            <label>Stimulus</label>
          </ref>
        </seg>

  (…)
```

Figure 6: A "vanilla TEI" version of CoMParS annotation

The above early version of CoMParS annotation, realised in "vanilla TEI", shows some of the features presented or visualised earlier in the present report, such as the morphosyntactic annotation of terminal elements or the functional domain identification. Note that this reduced listing encodes a graph consisting of a list of leaf (terminal) nodes and a list of non-terminal nodes, each of which defines an arc (branch) terminating in either another non-terminal node or, as in the listing above, pointing at a terminal. The pointing is performed by means of @xml:id attributes (e. g., xml:id="deu-s_001_01") that encode the endpoint for a reference, and @target attributes (e. g., target="#deu-s_001_01") that mark the start of the reference (the XML element which they modify, in this case <ref target="#deu-s_001_01" type="edge">).

Recall from Section 2 that the monolingual corpora do not provide alignment information but merely expose targets for alignment, which are then referenced in the aligned part of CoMParS. These targets for alignment are identified exactly by the by now familiar @xml:id attributes, which are present on practically each element of monolingual annotation.

The "vanilla TEI" representation, while available practically out-of-the-box, suffers nevertheless from certain drawbacks, both practical and theoretical, that have been eliminated with the move to standoff representation. In the remainder of this section, selected features of the new representation will be exemplified and described.

The reduced and trimmed listing below presents the top part of a typical monolingual corpus inside CoMParS.

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="compars-main_header.xml"/>
 <TEI>
   <xi:include href="header-deu.xml"/>
   <xi:include href="fsdDecl-deu.xml"/>
   <text xml:lang="de">
     <body>
       <ab xml:id="deu-ab1" n="1">
         <seg xml:id="deu-ab1seg1">Ich habe mich im Winter in dich verliebt.</seg>
       </ab>

         (…)
     </body>
   </text>
```

Figure 7: Top part of a monolingual corpus, current CoMParS annotation

Looking from the top down, the <xi:include> elements have the effect of being replaced by the documents referenced by the @href attribute. In this way, a single monolingual corpus includes the main corpus header (`compars-main_header.xml`), the monolingual header (`header-deu.xml`), and the placeholder for Feature Structure Declarations. The <text> element, now in full accordance with its TEI definition, contains "pristine" pieces of uninterrupted text of the corpus. This is now the place where new material enters the corpus. From here, it is operated on by the human annotator or it is sent to the Weblicht linguistic annotation pipeline (Hinrichs/Hinrichs/Zastrow 2010) and returns, after re-import, in the form of standoff annotations, presented below.

Below the <text> element, the standoff area begins, as shown in the reduced fragment below.

```
<standOff xmlns="http://standoff.proposal" xmlns:tei="http://www.tei-c.org/ns/1.0">
  <listAnnotation n="1" corresp="#deu-ab1" type="sequence">
    <listAnnotation type="segmentation" xml:id="deu-ab1tok">
      <tei:seg from="0" to="3" xml:id="deu-ab1tok1">Ich</tei:seg>
      <tei:seg from="4" to="8" xml:id="deu-ab1tok2">habe</tei:seg>
      <tei:seg from="9" to="13" xml:id="deu-ab1tok3">mich</tei:seg>

          (…)
  </listAnnotation>
  <listAnnotation type="terminals" xml:id="deu-ab1t">
      <tei:seg corresp="#deu-ab1tok1" type="t" xml:id="deu-ab1t1" >
        <tei:fs>  <!-- later on, we should use 'type' here -->
          <tei:f name="word">Ich</tei:f>
          <tei:f name="lemma">ich</tei:f>
          <tei:f name="pos" xml:lang="de">Subst</tei:f>
          <tei:f name="morph" xml:lang="de">1.Sg.Nom</tei:f>
          <tei:f name="UD">
            <tei:fs type="UD">
             <tei:f name="pos">PRON</tei:f>
             <tei:f name="feats">PronType:Prs|Person:1|Number:Sing|Case:Nom</tei:f>
           </tei:fs>
          </tei:f>
          <tei:f name="STTS">
            <tei:fs type="STTS">
              <tei:f name="pos">PPER</tei:f>
              <tei:f name="feats">1.Sg.Nom</tei:f>
            </tei:fs>
          </tei:f>
        </tei:fs>
      </tei:seg>

          (…)
  </listAnnotation>
          (…)
```

Figure 8: Fragment of the standOff annotation area, current CoMParS annotation

Unlike the robust <text> element, the <standOff> element offers a very modest number of internal building pieces, and annotations are encoded essentially as list of lists. The fragment above corresponds to the earlier example of "vanilla" annotation, with certain important enhancements.

Firstly, tokenization has been delegated to a separate <listAnnotation> element, and, in the reduced listing above, addresses the first three words of the sequence identified as "deu-ab1". This is done by means of @from and @to attributes, which contain numeric references to inter-character points, starting with the value "0" (this type of addressing is recommended by the ISO LAF family of specifications). Note that the first segment is identified by means of xml:id="deu-ab1tok1", which is referenced by the @corresp attribute in the next <listAnnotation>, which contains terminals. Each terminal element in the new version of CoMParS is equipped with a massive amount of information, as discussed in Section 2.

The feature structures identified as “STTS” carry the standard German lexical annotation values, familiar from the previous listings and figures. Those identified as “UD” carry values specific to Universal Dependencies.

The feature “word” has as its value the orthographic shape of the word, including capitalization (this is a redundancy that serves as a means of consistency checking across annotations). The feature “lemma” carries the standard lemmatization information (which may be overridden by specific tagsets and tools, which is not the case in the example at hand). The top-level feature named “pos”, carrying additional language identification (xml:lang=”de”) is the human-readable part-of-speech name meant to be displayed to the German-language users. The same is true of the feature named “morph”, in this case identical with the STTS value.

The annotations adduced above are followed by <listAnnotation> elements containing annotations analogous to those shown in the “vanilla” version, with the addition of dependency annotations according to the guidelines of Universal Dependencies (still in the process of being added).

4. Tools

CoMParS does not aim at producing its own visualisation or analysis tools. The technical aim of the project is to annotate linguistic data in such a way as to make it possible for relatively simple transformation scripts to create formats that will be accepted by standard annotation, analysis and visualisation tools, such as TrEd (Pajas/Štěpánek 2010) or CQPWeb.[9] Automatic annotation is delegated to the WebLicht pipeline (due to the hand-crafted nature of CoMParS, WebLicht output is examined and hand-corrected before inclusion into monolingual corpora).

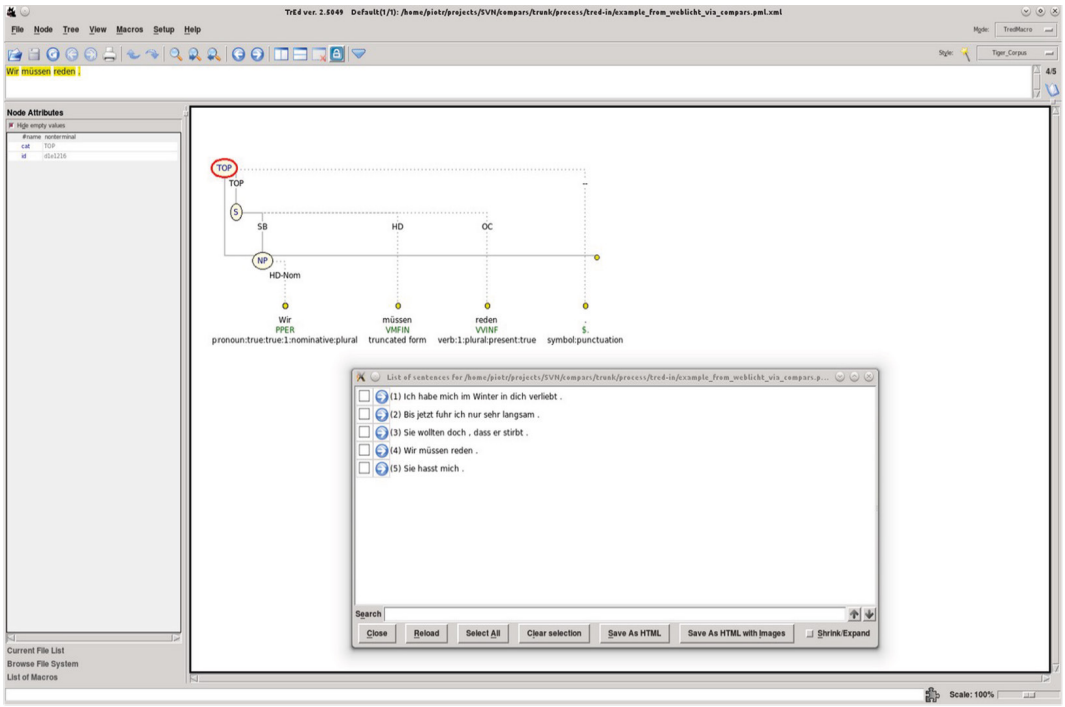


Figure 9: A screenshot of the editing window of TrEd, which is going to be used for manual correction of CoMParS annotations

9 TrEd is available from http://ufal.mff.cuni.cz/tred/ (last accessed 26-3-2025), and CQPWeb from http://cwb.sourceforge.net/cqpweb.php (last accessed 26-3-2025).

The "vanilla" version of CoMParS has a transformation tool suite that makes it possible to convert data in and out of Salto (as shown in Section 2), and into TrEd. While we plan to abandon Salto, which has not been maintained for years, the tools for TrEd will be modified to serve the current version of CoMParS, after the various remaining wrinkles of the new format have been smoothed out. A screenshot from the early stage of using TrEd with the "vanilla" CoMParS is shown in Figure 9.

A major obstacle in the transformation process to and from TrEd used to be the fact that, natively, TrEd only allows its plugins to use XSLT 1.0, which would be far too limiting in writing the plugins for CoMParS. Thanks to the assistance offered by its co-creator and maintainer, Jan Štěpánek, processing data with XSLT 2 and 3 is now also an option.

5. Summary and further steps

We have presented and discussed the basic assumptions underlying the architecture and implementation of CoMParS, sketching the history of its development until April 2017, and illustrated the discussion with selected listings and screenshots, for the purpose of both reporting on the current state of the project, and in order to lay the basis for more detailed discussion and planning in the GDE project of the Grammatik department of IDS Mannheim.

The next steps in the development of CoMParS are the upgrade of corpus tools after the format change and creation of new tools, simplifying the steps needed to fully utilize the Weblicht annotation pipeline, and adding UD and (potentially) Framenet annotations to all the sequences in the monolingual corpora.

References

Bański, Piotr (2010): Why TEI stand-off annotation doesn't quite work: And why you might want to use it nevertheless. In: Proceedings of Balisage: The Markup Conference 2010, Montréal, Canada, August 3–6, 2010. (= Balisage Series on Markup Technologies 5). DOI: 10.4242/BalisageVol5.Banski01.

Bański, Piotr/Przepiórkowski, Adam (2009): Stand-off TEI annotation: The case of the National Corpus of Polish. In: Stede, Manfred/Huang, Chu-Ren/Ide, Nancy/Meyers, Adam (eds.): Proceedings of the 3rd Linguistic Annotation Workshop (LAW III), Suntec, Singapore, 6–7 August 2009. Suntec: Association for Computational Linguistics, pp. 64–67.

Bański, Piotr/Gaiffe, Bertrand/Lopez, Patrice/Meoni, Simon/Romary, Laurent/Schmidt, Thomas/Stadler, Peter/Witt, Andreas (2016): Wake up, standOff! TEI Conference 2016, Sep 2016, Vienna, Austria. https://hal.inria.fr/hal-01374102 (last accessed 26-3-2025).

Burchardt, Aljoscha/Erk, Katrin/Frank, Anette/Kowalski, Andrea/Padó, Sebastian (2006): SALTO – A versatile multi-level annotation tool. In: Calzolari, Nicoletta/Choukri, Khalid/Gangemi, Aldo/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Tapias, Daniel (eds.): Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06). Genoa: European Language Resources Association (ELRA), pp. 517–520.

Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Goggi, Sara/Grobelnik, Marko/Maegaard, Bente/Mariani, Joseph/Mazo, Helene/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (eds.) (2016): Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16). Portorož: European Language Resources Association (ELRA).

Chiarcos, Christian/Ritz, Julia/Stede, Manfred (2012): "By all these lovely tokens..." Merging conflicting tokenizations. In: Lang Resources & Evaluation 46, pp. 53–74. https://doi.org/10.1007/s10579-011-9161-0 (last accessed 26-3-2025).

Eckart de Castilho, Richard/Ide, Nancy/Lapponi, Emanuele/Oepen, Stephan/Suderman, Keith/Velldal, Erik/Verhagen, Marc (2017): Representation and interchange of linguistic annotation. An in-depth, side-by-side comparison of three designs. In: Schneider, Nathan/Xue, Nianwen (eds.): Proceedings of the 11th Linguistic Annotation Workshop. Valenzia: Association for Computational Linguistics, pp. 67–75. www.aclweb.org/anthology/W17-0808 (last accessed 26-3-2025).

Erk, Katrin/Padó, Sebastian (2004): A powerful and versatile XML format for representing role-semantic annotation. In: Lino, Maria T./Xavier, Maria F./Ferreira, Fátima/Costa, Rute/Silva, Raquel (eds.): Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04). Lisbon: European Language Resources Association (ELRA), pp. 799–802.

Givón, Talmy (1984): Syntax. A functional-typological introduction. Vol. 1. Amsterdam/Philadelphia: Benjamins.

Hinrichs, Erhard/Hinrichs, Marie/Zastrow, Thomas (2010): WebLicht: Web-based LRT Services for German. In: Kübler, Sandra (ed.): Proceedings of the ACL 2010 System Demonstrations (ACLDemos '10), Stroudsburg, PA, USA. Uppsala: Association for Computational Linguistics, pp. 25–29.

Ide, Nancy/Suderman, Keith (2014): The linguistic annotation framework: A standard for annotation interchange and merging. In: Lang Resources & Evaluation 48, pp. 395–418. https://doi.org/10.1007/s10579-014-9268-1 (last accessed 26-3-2025).

Kutscher, Silvia (2014): Entwurf einer Makrostruktur zum Bereich der Sachverhaltsversprach lichung im Rahmen des Projekts „Grammatik des Deutschen im europäischen Vergleich (GDE-V)". (= Arbeitspapiere der Abteilung Grammatik 1). Mannheim: Institut für Deutsche Sprache.

Mengel, Andreas/Lezius, Wolfgang (2000): An XML-based representation format for syntactically annotated corpora. In: Gavrilidou, Maria/Carayannis, George/Markantonatou, Stella/Piperidis, Stelios/Stainhauer, Gregory (eds.): Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC '00). Athen: European Language Resources Association (ELRA), pp. 1–6.

Nivre, Joakim/de Marneffe, Marie-Catherine/Ginter, Filip/Goldberg, Yoav/Hajič, Jan/Manning, Christopher D./McDonald, Ryan/Petrov, Slav/Pyysalo, Sampo/Silveira, Natalia/Tsarfaty, Reut/Zeman, Daniel (2016): Universal dependencies v1: A multilingual treebank collection. In: Calzolari et al. (eds.), pp. 1659–1666. www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf (last accessed 26-3-2025).

Pajas, Peter/Štěpánek, Jan (2008): Recent advances in a feature-rich framework for treebank annotation. In: Scott, Donia/Uszkoreit, Hans (eds.): Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008). Manchester: COLING 2008 Organizing Committee, pp. 673–680.

Pose, Javier/Lopez, Patrice/Romary, Laurent (2014): A generic formalism for encoding stand-off annotations in TEI. https://hal.inria.fr/hal-01061548 (last accessed 26-3-2025).

Przepiórkowski, Adam (2009): TEI P5 as an XML standard for treebank encoding. In: Passarotti, Marco/Przepiórkowski, Adam/Raynaud, Savina/Van Eynde, Frank (eds.): Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT8), 4–5 December 2009, Milan, Italy. Milan: Universitario dell'Università Cattolica, pp. 149–160.

Straka, Milan/Hajič, Jan/Straková, Jana (2016): UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In: Calzolari et al. (eds.), pp. 4290–4297.

Stührenberg, Maik/Jettka, Daniel (2009): . In: Proceedings of Balisage: The Markup ConferencA toolkit for multi-dimensional markup: The development of SGF to XStandoffe 2009, Montréal, Canada, August 11–14, 2009. (= Balisage Series on Markup Technologies 3). DOI: 10.4242/BalisageVol3.Stuhrenberg01.

TEI Consortium (2017): TEI P5: Guidelines for electronic text encoding and interchange, ver. 3.0.0. www.tei-c.org/Guidelines/P5/ (last accessed 26-3-2025).

Trawiński, Beata (2016): Linguistic data in contrastive studies. Addressing the need for a multilingual parallel resource annotated with semantic-functional information. In: Domínguez Vázquez, María J./Kutscher, Silvia (eds.): Interacción entre gramática, didáctica y lexicografía. Estudios contrastivos y multicontrastivos. Berlin/Boston: De Gruyter, pp. 85–98.

Trawiński, Beata/Schlotthauer, Susan/Bański, Piotr (2021): CoMParS: Eine Sammlung von multilingualen Parallelsequenzen des Deutschen und anderer europäischer Sprachen. In: Lobin, Henning/Witt, Andreas/Wöllstein, Angelika (eds.): Deutsch in Europa. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston: De Gruyter.

Bibliographic information

Information on the citation of this publication:

Bański, Piotr (2025): The logical architecture of CoMParS and its XML implementation. (= IDS*open* 12). Mannheim: IDS-Verlag.

DOI https://doi.org/10.21248/idsopen.12.2025.53

Contact information

Piotr Bański
Leibniz-Institut für Deutsche Sprache
R 5, 6–10
68161 Mannheim
E-Mail: banski@ids-mannheim.de

Imprint

Bibliographic information of the German National Library
The German National Library lists this publication in the German National Bibliography; detailed bibliographic data is available online at
http://dnb.dnb.de.

IDS-Verlag · Leibniz-Institut für Deutsche Sprache
R 5, 6–13 · 68161 Mannheim
www.ids-mannheim.de

IDS-Verlag

Publication series: IDS*open*: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache
Series editors: Norman Fiedler, Katrin Hein, Siegwalt Lindenfelser, Beata Trawiński
Copy editor: Melanie Kraus
Typesetters: Joachim Hohwieler, Melissa Manara

ISBN: 978-3-948831-77-6 (PDF)

ISSN: 2749-9855