

TypeCraft Software Release Notes

Dorothee Beermann, Tony Gjendahl, Tormod Haugland, Lars Hellan

NTNU, Trondheim, Norway

Pavel Mihaylov

Ontotext, Sofia, Bulgaria

VERSION: [2.2](#)

Release date: [20.08.2015](#)

Note: [The release notes cover the major release v.2.0 and two subsequent minor releases](#)

Introduction

TypeCraft is a linguistic infrastructure project at the Norwegian University of Science and Technology. Its aim is to support researchers in the Humanities by providing an easy-to-control online access to language data. The TypeCraft tool is an open-ended interlinear glossed text repository and linguistic data management system. Its online workspace allows for the comfortable handling of small-to-medium-size corpora created from fieldwork, or imported from other online resources.

The TypeCraft Importer is a light-weight, web-based client, specially designed for the import of Toolbox (ELAN) data and TypeCraft XML. The latter functionality supports the reuse of already structured data for further annotation. In addition our importer features a POS tagger for Norwegian data.

The TypeCraft Editor is an efficient tool for the manual annotation of linguistic data to be used in linguistic research.. Users work with several predefined annotation layers and controlled vocabularies, which increases the likelihood that their data is consistently interpreted and easier to integrate with other linguistic data across public resources and archives.

TypeCraft Language Portal allows at present public access to annotated data from more than 140 languages. The data can be searched conveniently online by multi-tier search. A corpus query language is not required.

The TypeCraft Exporter facilitates the export from selected or all annotation tiers in formats that allow an easy import to other applications including standard text editors. This makes your life

easier when you want to migrate your data to other applications, or publish it, using standard publishing channels.

The TypeCraft Wiki is maintained by TypeCraft users. Articles discuss mostly less-described languages and address linguistic questions. The TypeCraft wiki allows for the automatic inclusion of annotated TypeCraft data, which makes it a great place to pre-publish data-sets as part of linguistic research.

Minor Release v2.2

Features

- MediaWiki
 - The TypeCraft-MediaWiki has been redesigned completely with a new user-friendly design.
 - The MediaWiki has been updated to version 1.24.2.
 - Added extension “Semantic MediaWiki”.
 - Added a new Special page “Portal of Languages”.
 - The page is automatically updated when the database changes. It contains overview information of the count of texts of specific languages in the TypeCraft database.
 - Each language has its own page containing an overview over all public texts of that language.
 - Updated all extensions.
- Exporter
 - Exportation of phrases to HTML (for usage in MS Word or OpenOffice) now opens an options-menu where the user can decide which tiers to export. (Flexible export)

Open Issues version 2.2

- **Import** of *Annotation Pro* data (propagation of time-alignment information to the Editor)
- **Export**
 - Improved word list export: proper progress bar that reflects the real state of the task for search and editor
 - improved export to LaTeX
- **Search**
 - Filter chains, storing of results
 - Search and replace in collections of own texts

Wiki

- Book printing is functional but has some bugs and design flaws. These are planned to be addressed and fixed with v2.3

Minor Release v2.1

Features

- Importer (**Resolved Issues relating to the unstable importer**)
 - The importer has been rebuilt from scratch to conform with the new layout.
 - A progress-counter is shown for each upload.
 - Any error, warning or relevant info is shown in the *Importation Status* bar.
 - A “Goto text”-button is shown for each completed upload. Clicking this opens the newly imported text in the TypeCraft Editor.
 - The importer now has a “drop-zone” where files can be dragged and dropped.
 - The importer now imports any file **asynchronously**, allowing users to leave the page after an import has started without harming the import-process.
 - Asynchronicity allows the importer to run multiple import-processes simultaneously, even though they are started at different points in time.
 - The importer now automatically detects the import type of added files depending on file type.
 - .xml maps to TypeCraft XML
 - .txt maps to Toolbox (ELAN)
 - The importer can now validate XML-files before starting the importation process. The validation picks up common errors in XML files and prints them to the user in a constructive manner.
- System
 - The database has been optimized on several points making loading of phrases and loading of texts significantly quicker. This change will also make the system generally quicker on any task related to texts, phrases, words and morphemes.
- API
 - New endpoint GET *conversion/{id}/status*
 - Fetches the conversion/importation status of the import specified by the id *{id}*.
 - New endpoint GET *get_typecraft_xsd*
 - Returns the typecraft.xsd schema. The request is equivalent to calling GET <http://typecraft.org/typecraft.xsd>.
 - New endpoint POST *validate_xml*
 - Validates an XML file by checking both the XMLI-syntax and checking validity against the typecraft.xsd schema.

Bugs

- Fixed a bug where certain TC-XML imported files would have their phrases classified as “Lost phrases”.
- Fixed a bug where any Toolbox- or Smart Tagger-imported file would have its phrases classified as “Lost phrases”.
- Fixed a bug where any imported text with empty morphemes would be incorrectly stored in the database.

- Fixed a bug where the transliteration of Pinyin phrases would crash for phrases without a morpheme tier.

Public database contents

- Made available 200 000 annotated German phrases from the Leipzig Wortschatz corpus.
- Made available 420 Akan sentences with English translations.

Major Release v2.0

Features

- A completely redesigned editor with proper support for texts
 - Visual rich text editor
 - Annotated phrases are now linked to the original text
 - Texts are now related to metadata
 - Introduction of *Senses* which are elements of text and allow the annotation of discourse related information.
 - Improved lazy annotation mode
 - Added support for annotating punctuation
 - Added second translation tier
 - Added a way to hide unnecessary tiers
 - A new cleaner and themeable look
- Better search
 - Search for text metadata
 - Search for owner and date
 - Search for *Senses*
 - Search results are organized in pages.
 - Highlighting of the search terms in the results
 - Display of the number of matching instances of search terms in results
- TypeCraft Importer: Converts annotated text from Toolbox and WORD-based IGT to TypeCraft XML and imports it.
- Export Wordlists from texts as .txt and html from the search page and from the editor.
- Implemented an API
- more consistently organised, more accessible TCwiki

Open Issues version 2.0

- Import of ELAN data
- Export
 - Export of marked tiers (user chooses tiers for export)
 - improved export to LaTeX
- Search
 - Filter chains, storing of results
 - Search and replace in collections of own texts