Software Release and Progress Report

Start date: 2018-10-23 End date: 2019-03-18

Short Summary

Scientific output:

- Maarten van Gompel and Iris Hendrickx submitted an extended abstract on LaMachine for the CLARIN conference, which Iris presented in Pisa in poster-form. We also worked on a full paper which was recently accepted for the proceedings of that conference.
- Posterpresentatie op CLIN 2019, Groningen on the evaluation of the Frog modules

After months of development, FoLiA v2.0 has been released (for details see the release notes later in this report).

Ongoing unreleased work by Ko van der Sloot:

- libfolia
  - een processor class geïmplementeerd om FoLiA streaming te verwerken
- ucto:
  - gebruikt nu de streaming class van libfolia
  - verschillende issues met folia verwerking gerepareerd
  - grote schoonmaak gehouden in de code
- frog:
- herimplementatie met een simpelere interne datastructuur
- voor FoliA verwerking gebruik makend van de libfolia processor class daardoor kleinere memory footprint.
- kleine bug fixes
- ticcltools:
  - nieuwe features gegevoegd tbv TICCL/PICCL/TICCLAT etc

Henk van den Heuvel worked on:

- data curation e-WGD: retrieving placenames and kloekecodes where missing
- eWGD: prelim. version website
- audio presentation of fragments in OpenSoNaR+ (topic for discussion: continue?)
- Outstanding: metadata linking CGN-SONAR (Nelleke wil start in April)
- Hosted CLARIN-DELAD workshop Utrecht 28-30 January on sensitive data: See https://www.clarin.eu/blog/clarin-workshop-delad-database-enterprise-language-and-speech-disorders
- CLARIAH-paper about DELAD issues in prep for CLARIN AC Leipzig: GDPR & data use cases sensitive data
- SSHOC-project started: sensitive data & interview data in CLARIN workflow
- G2P webservice for Dutch provided (v1.0): https://webservices-lst.science.ru.nl/g2pservice

Iris Hendrickx worked on:

- Frog evaluation of Ucto, NER, POS, LEM modules
- student projects with Frog in Master course Text and Multimedia Mining
- internship, enhancing Frog NER with more fine-grained categories
- Frog within case study with KB
- Ucto documentation, released 30 Oct 2018
- documentation of Frog in new online format, https://frognlp.readthedocs.io/en/latest/ (work in progress)

CLAM

Project/task/deliverable references: CLARIAH-PLUS WP3 T142; CLARIAH-CORE WP3 T21 [D2.8 (software); D2.9 (doc)]

clam v2.4.4

Bugfix release: fix for forwarders in web interface #69

(Released by Maarten van Gompel on 2019-02-11) https://github.com/proycon/clam/releases/tag/v2.4.4

clam v2.4.3

- Allow CLAMClient to use HTTP Basic Auth instead of Digest
- check whether the proper authorization scheme is used

(Released by Maarten van Gompel on 2018-12-12) https://github.com/proycon/clam/releases/tag/v2.4.3

clam v2.4.2

- Bugfix: Boolean paramters (checkboxes) with default True did not get posted #73
- Allow HTTP digest authentication to be disabled (e.g. for allowing only HTTP basic authentication)

(Released by Maarten van Gompel on 2018-12-11) https://github.com/proycon/clam/releases/tag/v2.4.2

clam v2.4.1

Minor bugfix release:

- Fix in CLAMFile.read()

(Released by Maarten van Gompel on 2018-11-21) https://github.com/proycon/clam/releases/tag/v2.4.1

clam v2.4.0

- Ported documentation to sphinx and rewrote/restructured various sections (#72), the old documentation is now obsolete.
- added system details to the CLAM footer (so it's not just a CLAM footer)
- Added Forwarders (#69) and ForwardViewer
- Implemented CLAMData.get() (akin to dict.get()) #50
- parameter default fix (#71)

(Released by Maarten van Gompel on 2018-11-19) https://github.com/proycon/clam/releases/tag/v2.4.0

clam v2.3.6

- Further updates and fixes to clamnewproject to generate a new project with host-specific external configurations #70

(Released by Maarten van Gompel on 2018-11-06) https://github.com/proycon/clam/releases/tag/v2.3.6

clam v2.3.5

- Updated clamnewproject to generate projects with a setup.py #70

(Released by Maarten van Gompel on 2018-11-02) https://github.com/proycon/clam/releases/tag/v2.3.5

FLAT

Project/task/deliverable references: CLARIAH-PLUS WP3 T062 & T063; CLARIAH-CORE WP3 T24 [D2.10 (software); D2.11 (docs)]

(no releases this period)

FoLiA

Project/task/deliverable references: CLARIAH-PLUS WP3 T108; CLARIAH-CORE WP3 T71 [D2.5 (libs); D2.6 (docs); D2.7 (tools)]

folia 2.0.0

This is a major new release of FoLiA, which includes some breaking changes such as renamed elements. Nevertheless, the FoLiA libraries retain backward compatibility and can read FoLiA v1 (and v0) documents and upgrade them. Points of general interest:

- Completely revised the FoLiA documentation, turned into more formal specification; automatically drawn from the official specification; with automatically validated examples. Now available as a webpage hosted on https://folia.readthedocs.io (PDF still available too) #43
  - The documentation includes some guidelines on good FoLiA practises (arose from #70 and others)
- Added proper support for provenance logging in FoLiA #46
- Renamed alignment annotation to relation annotation #59
- Ensured most examples are "sensible" #9
  - Extended tests using these examples, all examples are automatically tested now
- The FoLiA tools are now split from the central FoLiA repository into a separate project at https://github.com/proycon/foliatools #55
  - Cleaner output without stack traces from FoLiA validator #44
- Implemented the ability to add inline annotations on multi-word spans (group annotations) and solved related multi-word issues. These were previously reserved only for use with structural elements. #51
- Revised the structure annotation hierarchy (i.e. which structural elements are allowed under which parents) on certain points #42
- Implemented a hidden words annotation type, allowed a layer of implicit/empty/ghost words that can be referenced from span annotation. Needed e.g. for syntactic movement annotation. #58
- Allow encoding of soft word breaks / hyphenation #66 More technical points:
- Add support for provenance in FQL #60
- Annotation declaration overhaul and handle missing set attribute in declarations #54
- Explicitly forbid and prevent forward wrefs from span annotation #41
- Apply space attribute more generically to multiple structure elements #61
- Added a new property in the specification to detect tags that may be (or MUST be) used as Wrefs #63
- Added a new property to distinguish folia:id (IDREF) from xml:id (ID) #64
- Alias attribute does not propagate to RelaxNG schema yet #65 A new FoLiA library has been released (replacing the previous one in PyNLPl): https://github.com/proycon/foliapy/releases/tag/2.0.0 A new version of FoLiA tools has also been released: https://github.com/proycon/foliatools/releases/tag/2.0.0 You may also consult the FoLiA release plan (#68) for more information on upgrading and compatibility.

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/folia/releases/tag/2.0.0

foliapy 2.0.1

Allow (limited!) serialisation of older FoLiA versions if `keepversion` is set.

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/foliapy/releases/tag/2.0.1

foliapy 2.0.0

Major new release; the former `pynlpl.formats.folia` module in PyNLPl has now been migrated to this new standalone package, which constitutes the new FoLiA library for Python. Included is the FoLiA library, FQL library and Set Definition library. This release implements the new FoLiA v2.0, consult the release notes there for a comprehensive list of changes. Library documentation is now hosted on https://foliapy.readthedocs.io/en/latest/. Upgrading from pynlpl is usually as easy as doing a `pip install folia` and replace within your software: `from pynlpl.formats import folia` with `import folia.main as folia`. The library is backward compatible.

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/foliapy/releases/tag/2.0.0

foliatools 2.0.1

Enable keepversion for certain tools, so they don't upgrade to FoliA v2 if FoLiA v1 input is given, ensuring better backward compatibility.

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/foliatools/releases/tag/2.0.1

foliatools v2.0.0

This is the FoLiA v2 version of the FoLiA-tools, now split from the central FoLiA repository at https://github.com/proycon/folia and building on the new FoLiApy library. New documented resides at https://folia-tools.readthedocs.io/

- There is a new `foliaupgrade` tool to upgrade from FoLiA v1 (or v0) to v2 (although this is usually done implicitly by any of the other tools, or the FoLiAPy library itself, anyway)
- Refactored the entire foliavalidator
- Cleaner output without stack traces from FoLiA validator (proycon/folia#44)
- Added automatic injection of provenance information to various tools

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/foliatools/releases/tag/v2.0.0

foliautils v0.10

[Ko vd Sloot]

- fixed icu:namespace issues
- added FoLiA-abby, an ABBY to FoLiA convertor
- src/FoLiA-abby.cxx, src/FoLiA-page.cxx, src/FoLiA-pm.cxx:
  - Allow 'none' value for --prefix
- src/FoLiA-page.cxx, src/FoLiA-hocr.cxx: fixed Alignment info
- src/FoLiA-correct.cxx:
  - fixed a problem with correction of the last word of a trigram.
  - fix correction of paragraphs with only deeper text
  - The --rank option accepts more flavors of files
- src/FoLiA-stats.cxx:
  - added a --detokenize option
- several minor fixes, refactorings etc.
- updated tests

(Released by Ko van der Sloot on 2018-11-29) https://github.com/LanguageMachines/foliautils/releases/tag/v0.10

libfolia v1.15

- added (still experimental) code for a FoLiA Builder, Processor and TextProcessor class. Use with care. The API may change unannounced!
- a foliadiff script (using folialint) is installed now
- several refactorings, to make the code more clear.
- the 'ref' attribute was not serialized for TextContent
- several smaller small bug fixes
- the .so version is bumped to 9 because of a lot of API/ABI changes

(Released by Ko van der Sloot on 2018-11-29) https://github.com/LanguageMachines/libfolia/releases/tag/v1.15

pynlpl v1.2.9

The FoLiA library is being migrated out of PyNLPl to a new standalone project: https://github.com/proycon/foliapy (`pip install folia`), which is backward compatible with the one in pynlpl. Retaining the old library for a transition period, but implemented warnings and notices.

(Released by Maarten van Gompel on 2019-03-13) https://github.com/proycon/pynlpl/releases/tag/v1.2.9

pynlpl v1.2.8

- Enable proper confusion matrix in case of a dissimilarity between goals and observations #43

(Released by Maarten van Gompel on 2018-11-12) https://github.com/proycon/pynlpl/releases/tag/v1.2.8

Frog

Project/task/deliverable references: CLARIAH-PLUS WP3 T139; CLARIAH-CORE WP3 T22 [D2.1 (software); D2.2 (doc)]; T23 [D2.3 (froggen software), D2.4 (froggen docs)]

LaMachine

Project/task/deliverable references: CLARIAH-PLUS WP3 T098; CLARIAH CORE WP3: LaMachine v2 plan in scope of CLARIAH WP3 VRE

codemetapy v0.2.1.1

(Minor rerelease without changes just to trigger a DOI on Zenodo)

(Released by Maarten van Gompel on 2019-01-16) https://github.com/proycon/codemetapy/releases/tag/v0.2.1.1

labirinto v0.2.4.1

(minor rerelease just to trigger DOI update on Zenodo)

(Released by Maarten van Gompel on 2019-01-16) https://github.com/proycon/labirinto/releases/tag/v0.2.4.1

LaMachine v2.5.0

Adds support for FoLiA 2.0, see also the release plan in proycon/folia#68

- Added new FoLiApy
- FoLiA-tools is now split from the central FoLiA repository and LaMachine is adapted accordingly. The central repository containing examples and specifications is still supplied with LaMachine, but usually not needed for most end-users.

(Released by Maarten van Gompel on 2019-03-14) https://github.com/proycon/LaMachine/releases/tag/v2.5.0

LaMachine v2.4.11

- Added FLAIR to pytorch package
- Added phonetisaurus and g2pservice (limited availability), includes openfst independent of kaldi

(Released by Maarten van Gompel on 2019-03-05) https://github.com/proycon/LaMachine/releases/tag/v2.4.11

LaMachine v2.4.10

- implemented initial support for LXC/LXD containers #134
- fix for macOS installation of uwsgi
- added Go
- added a -h/--help flag to lamachine-update
- added --only option for lamachine-update to more selectively update

(Released by Maarten van Gompel on 2019-02-06) https://github.com/proycon/LaMachine/releases/tag/v2.4.10

LaMachine v2.4.9

- Added GLEM webservice
- listen on all IPs for Jupyter Lab (fixes #129)
- fixes for continuous integration on travis
- fixed for installation on Ubuntu 18.04

(Released by Maarten van Gompel on 2019-01-16) https://github.com/proycon/LaMachine/releases/tag/v2.4.9

LaMachine v2.4.8

- Fixes lamachine update error in docker/VM #127

(Released by Maarten van Gompel on 2018-12-19) https://github.com/proycon/LaMachine/releases/tag/v2.4.8

LaMachine v2.4.7

- Set default path of docker container to home dir and create convenience symlinks to $DATA_PATH and $LAMACHINE_PATH (arose from #125)
- explicitly install python3-setuptools (debian/ubuntu)
- Travis-CI integration tests are finally on ubuntu 16.04 instead of 14.04
- Use pip instead of `python setup.py install` whenever possible, with fallbacks
- Added GLEM to LaMachine
- Improved sanity check and error feedback for possibly outdated/malfunctioning virtualenv python
- Improved Ansible log output (colors and less cruft)

(Released by Maarten van Gompel on 2018-12-08) https://github.com/proycon/LaMachine/releases/tag/v2.4.7

LaMachine v2.4.6

Minor bugfix release:

- Fix: PICCL didn't get registered suddenly?

(Released by Maarten van Gompel on 2018-11-21) https://github.com/proycon/LaMachine/releases/tag/v2.4.6

Miscellaneous

Project/task/deliverable references: Dependants of others

ticcutils v0.20

- PrettyPrint: added printing of pairs
- several small bug fixes
- added more tests to 'make check'
- fixed icu::namespace issues

(Released by Ko van der Sloot on 2018-11-27) https://github.com/LanguageMachines/ticcutils/releases/tag/v0.20

PICCL & TICCL

Project/task/deliverable references: CLARIAH-PLUS WP3 ???; CLARIAH-CORE WP3 T26 [D1.3 (PICCL software); T1.4 (PICCL docs)]

PICCL v0.7.6

- Another fix for plain text input and no ocr AND no ticcl scenario (addressed in #43)
- Clean up in the wrapper script (it's becoming too convoluted)

(Released by Maarten van Gompel on 2019-03-05) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.6

PICCL v0.7.5.1

Minor correction for missing version update only, no functional changes

(Released by Maarten van Gompel on 2019-02-28) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.5.1

PICCL v0.7.5

- Fix for plaintext input and no ocr scenario (previous fix was still wrong)

(Released by Maarten van Gompel on 2019-02-27) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.5

PICCL v0.7.4

Added Autosearch forwarder

(Released by Maarten van Gompel on 2019-02-11) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.4

PICCL v0.7.3.1

(Minor rerelease without changes to trigger a DOI update on Zenodo)

(Released by Maarten van Gompel on 2019-01-16) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.3.1

PICCL v0.7.3

At least for some PDFs, the PDF to image file convertor in PICCL, i.e. PDFimage, created spurious image files. These sometimes resulted in 'pages' of garbage. Also, when we started building this pipeline, PDFimages did not yet convert straight into tiff-format. So we also used 'convert'. Both have now been replaced by pdftoppm, which seems to produce exactly the same amount of output tiff-files as regular PDF viewers report.

(Released by martinreynaert on 2018-12-12) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.3

PICCL v0.7.2

- fixed --help flag for ocr and ticcl (was broken)
- fix for text input when skipping TICCL
- Use 300dpi instead of the default 72dpi when converting bitmaps to TIF, should reduce garbage output #45
- Minor logging improvements: output tesseract version to standard output (#45) + feedback on why frog is enabled

(Released by Maarten van Gompel on 2018-12-11) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.2

PICCL v0.7.1

- Implemented FLAT viewers (allows to delegate FoLiA output to FLAT for visualisation) #42

(Released by Maarten van Gompel on 2018-11-21) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.1

PICCL v0.7.0

- added a document CONTRIBUTE.md with contributor guidelines and technical details
- added and expanded comments to aid @martinreynaert in understanding the Nextflow pipelines
- Restructured the webservice profiles (CLAM):
  - Publish relevant output of intermediate stages for the end-user, not just a single final end-result.
  - Less duplication
  - Some small fixes
  - Removed obsolete/implicit tokeniser option for Frog
- Fixes in the wrapper script
  - Fixes for text input
  - Fix: Output did not show up for download when only OCR is enabled #40
- Updated the startserver* scripts for the piccl webservice, made them more LaMachine-aware
- Prevent accidentally feeding Nextflow's trace.txt log as input
- Report input files to stdout for some pipelines (ticcl,frog, tokenize)
- Fix in nederlab pipeline, allow untokenised folia input and add --tok option to force tokenisation
- README fix. #41 See also https://github.com/proycon/clam/issues/69

(Released by Maarten van Gompel on 2018-11-19) https://github.com/LanguageMachines/PICCL/releases/tag/v0.7.0

Software Quality Guidelines

Project/task/deliverable references: CLARIAH-CORE WP2 Task 54.100

(no releases this period)

Timbl & Mbt

Project/task/deliverable references: Pre-CLARIN

mbt v3.4

- added an option '--tabbed' to set the word/tag separator to TABS
- better option parsing
- added some timers for debugging

(Released by Ko van der Sloot on 2018-11-28) https://github.com/LanguageMachines/mbt/releases/tag/v3.4

timbl v6.4.13

- added a '--limit' option to use only the most significant features.

(Released by Ko van der Sloot on 2018-11-28) https://github.com/LanguageMachines/timbl/releases/tag/v6.4.13

Ucto

- official documentation released, in online format

Maarten van Gompel, Ko van der Sloot, Iris Hendrickx and Antal van den Bosch. UCTO: Unicode Tokeniser. Reference Guide, Language and Speech Technology Technical Report Series 18-01, Radboud University, Nijmegen, October 30, 2018, Available from https://UCTO.readthedocs.io/.

Project/task/deliverable references: CLARIAH-CORE WP3 T55 [D1.1 (software), D1.2 (docs)]

ucto v0.14.1

- fixed textcat installation problems om Debian and OpenBSD (https://github.com/LanguageMachines/ucto/issues/59)
- typo in the man page fixed

(Released by Ko van der Sloot on 2018-12-10) https://github.com/LanguageMachines/ucto/releases/tag/v0.14.1

ucto v0.14

[Ko van der Sloot]

- updated usage() and removed -S option (never used)
- make sure the right textclass is assigned to <w> nodes in FoLiA
- minor code fixes/refactorings
- added more tests
- updated man.1 page

[Maarten van Gompel]

- updated README.md

[Iris Hendrickx]

- Updated and extended the manual

(Released by Ko van der Sloot on 2018-11-29) https://github.com/LanguageMachines/ucto/releases/tag/v0.14

uctodata v0.8

[Ko van der Sloot]

- separated .abr files from there main files for all Languages
- updated italian data (thanks to @texttheater)

[Iris Hendrickx]

- updated abbrev files for Portuguese Turkish and French based on https://en.wiktionary.org/wiki/Category:Portuguese_abbr and https://en.wiktionary.org/wiki/Category:Turkish_initialisms.
- added full list of French abbreviations.

- added 'aub' to Dutch list

(Released by Ko van der Sloot on 2018-11-29) https://github.com/LanguageMachines/uctodata/releases/tag/v0.8