



PANACEA WP 4 / 7



Corpus **A**cquisition and **A**notation / **T**ext **P**rocessing **C**omponent

Lexical Analysis Workflow

Components and evaluation

Munich

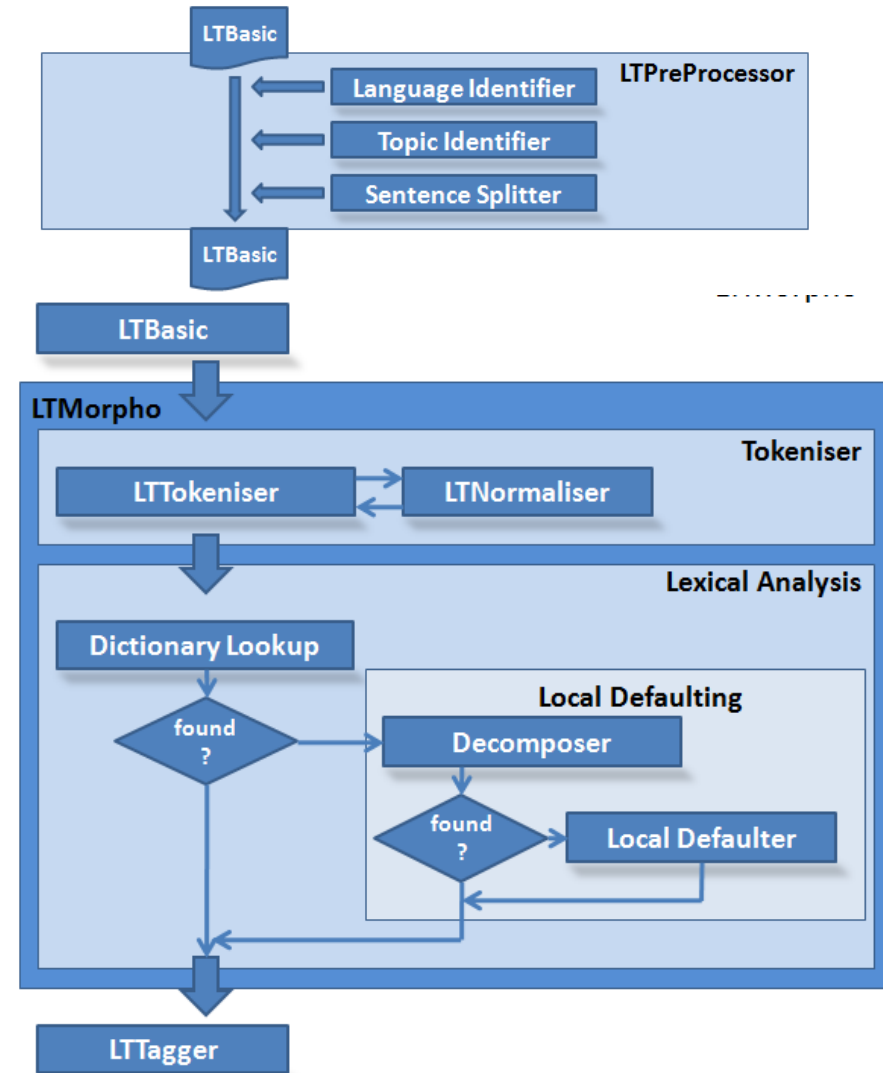
2011-Oct-11

*Gr. Thurmair, V. Aleksić,
Linguattec*

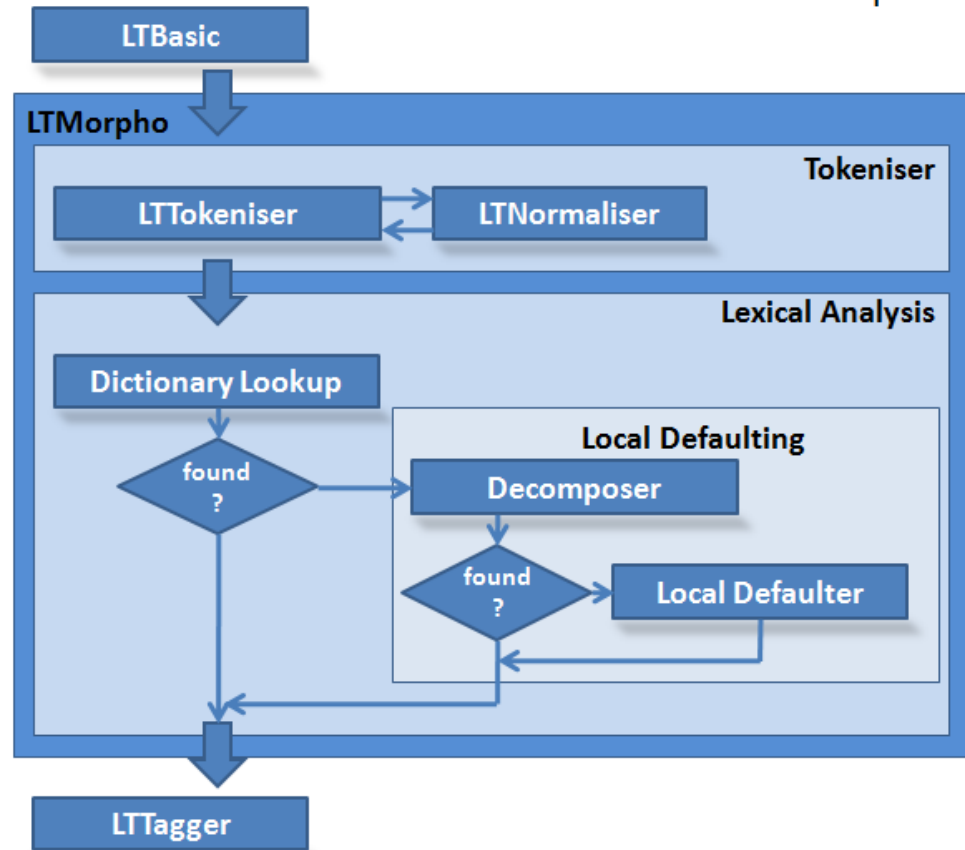
- Lexical Analysis Workflow
 - Definition and Embedding
 - Components
 - Resources
 - Processing
 - Evaluation

- Process crawled documents
 - As produced by the Crawler workflow
- Assign linguistic information to the tokens
 - ‚UNK‘ is not a linguistic category ...

- In: crawled XCES doc
- Assign doc-related info
 - Language identifier
 - Topic identifier
- Produce sentences
 - Sentence Splitter
- Produce tokens
 - Tokeniser
- Annotate tokens
 - Lexical analysis
- Out: lattice of readings



- LTMorpho
 - Tokeniser
 - Including normaliser
 - Lemmatiser
 - Including lexicon lookup
 - Decomposer
 - Unknowns could be compounds
 - Defaulter
 - Assign info to remaining unknowns
 - ,local‘, not contextual



- Purpose
 - To load the right resources
 - To avoid foreign language tokens
- Scope
 - On document, paragraph and sentence level
- Resources
 - Text form files, from speller lexicons (ca. 150K)

- Purpose
 - Assign topics to texts, used
 - For focused crawling
 - For linguistic analysis, e.g. in transfer selection
- Scope
 - Documents, paragraphs
- Resources
 - Taxonomy of about 40 topics (standard MT)
 - Features per topic (lemmata, multiwords)
 - De ca. 190 K (4.5K on avg.), en ca. 40K (1 K on avg.)
 - Supervised, manually corrected

Sentence Splitter

- Purpose:
 - separate texts into sentences by inserting `<s>` markup
- Scope:
 - On `<doc>` and `<p>` level
- Resources:
 - Typical startwords (ca. 10K / lang)
 - Typical end words (ca. 8K / lang)
 - Classification of ABBs (non-final, final, ...)

Evaluation

- Comparative evaluation
 - Ca. 75% of documents affected by different segmentation strategies
- Absolute evaluation
 - random sentences from 4-5 corpora de, en
 - Manual evaluation
 - De: 3844 sentences, error rate 0.44%
 - En: 3085 sentences, error rate 0.26%
 - (dirty input (HTML))

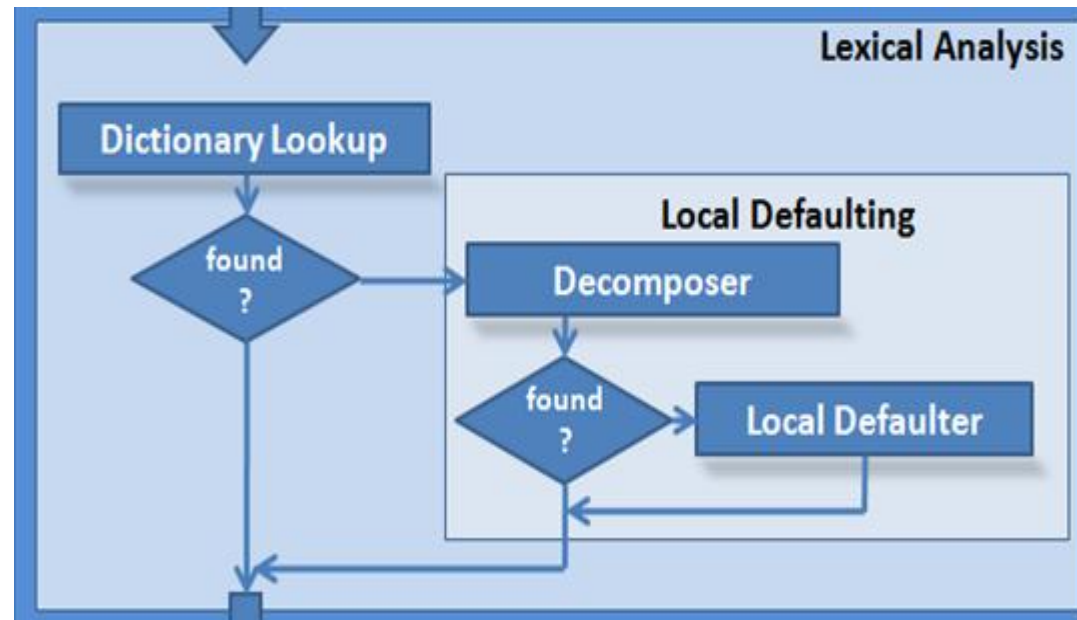
- Definition of Token:
 - a unit which can be found in the lexicon
- Purpose:
 - Prepare strings for lookup
 - Segmentation / Token definition (John's, arm-rest, 320i)
 - Normalisation on the canonical form of the lexicon
 - Casing
 - Spelling
- Resources
 - Normaliser files (ca. 8K per language)
 - De: Old spelling > new spelling. Ae/oe/ue > ä/ö/ü
 - En: US > UK spelling, hyphens

Evaluation

- Indirect criterion: lexicon coverage
 - In output files: Unknowns due to
 - Non-covered characters (nonbreakable spaces, ...)
 - (7K words)
 - Mis-segmented units
 - ...

Lexical Analysis

- Assign annotations to token strings
 - Simplest: tags; more complex: SCF, transfers
- Multi-step process
 - 1 Lemmatiser
 - 2 Decomposer
 - 3 local Defaulter



Lemmatiser

- Lexicon Lookup
- Lexicon = BLF format (textform \t lemma \t lexinfo)
 - Created from basic lexicon by flexers

• Tagsets and lexicon size	de	en
– Lemmata	3.7 m	41.3 K
– Basic (12 tags)	5.4 m	150.2 K
– Standard (80 tags)	5.8 m	173.2 K
– Extended (incl. Morph.)	17.7 m	272.6 K

Evaluation

- Testcorpus:
 - Europarl + emea + JRC-Acquis: 65.5 mio tokens
 - Evaluation
 - Lemmatised (31K words/sec)
 - Coverage: annotated with TAG info: 62.3 mio
 - 95.1% coverage
 - Remaining: 3.187 mio tokens
 - Lexicon errors
 - (evaluated random 1000 entries)
 - Standard tagset: 1,29%
 - Extended tagset: 2.5%
- > wrong: 804 K words

Decomposer

- Purpose:
 - Assign Lemma + Ling.Info to unknowns
- Resources:
 - Decomposer lexicon (legal strings + special tagset)
(470K entries, incl 12K irregulars)
 - Transition table (legal combinations of tags) (60x60)
 - Disambiguation rules (ca. 20)
- Decomposer en?
 - (*armrest, battleship, counterattack, ...*)

Evaluation

- Testcorpus
 - Unknowns of the previous component (3.187 m)
- Evaluation
 - Decompose (11.2 K words/sec)
 - Coverage:
 - Decomposed: 1.205 m tokens (37.8%)
 - Remaining 1.982 m tokens
 - Decomposer errors
 - (manual inspection of 14.3 K decompositions)
 - Error rate: 2.18%

-> wrong 26 K words

- Assign POS by string (suffix) similarity
 - ,*xxxxisation*' = noun, feminine, abstract
- Defaulting steps
 - Mixed case acronyms etc.: *AZ45/2004/1233*
 - Special noun category
 - Foreign language words
 - Special noun category
 - Learning of endings (890 K strings) (incl. spell errors)
- Resources:
 - Foreign language dictionary (70K)
 - Training result of word endings (890 K strings)

Evaluation

- Testcorpus
 - Tokens not decomposed (,hard‘ data ...) 1.98mio
 - Tokens derived from a lexicon (,good‘ data) 400K
- Evaluation
 - Run POS defaulter (30K words/sec)
 - Inspect results (ca. 16 K forms, different POS):
 - *All* input forms get a tag annotation (no surprise)
 - The question is: is it correct?
 - (Is the correct tag among the ones proposed, so that the tagger could have a chance)

Evaluation

- Correctness depends on
 - cleanness of data
 - Error rates for dictionary data: 5.9%
 - Error rates with (left-over) ,German‘ words of corpus: 10.6%
 - Error rates with all (left-over) words of test corpus: 20%
 - tags to be defaulted
 - NoC, NoP error rate < 10%
 - Vb, Ad > 30%

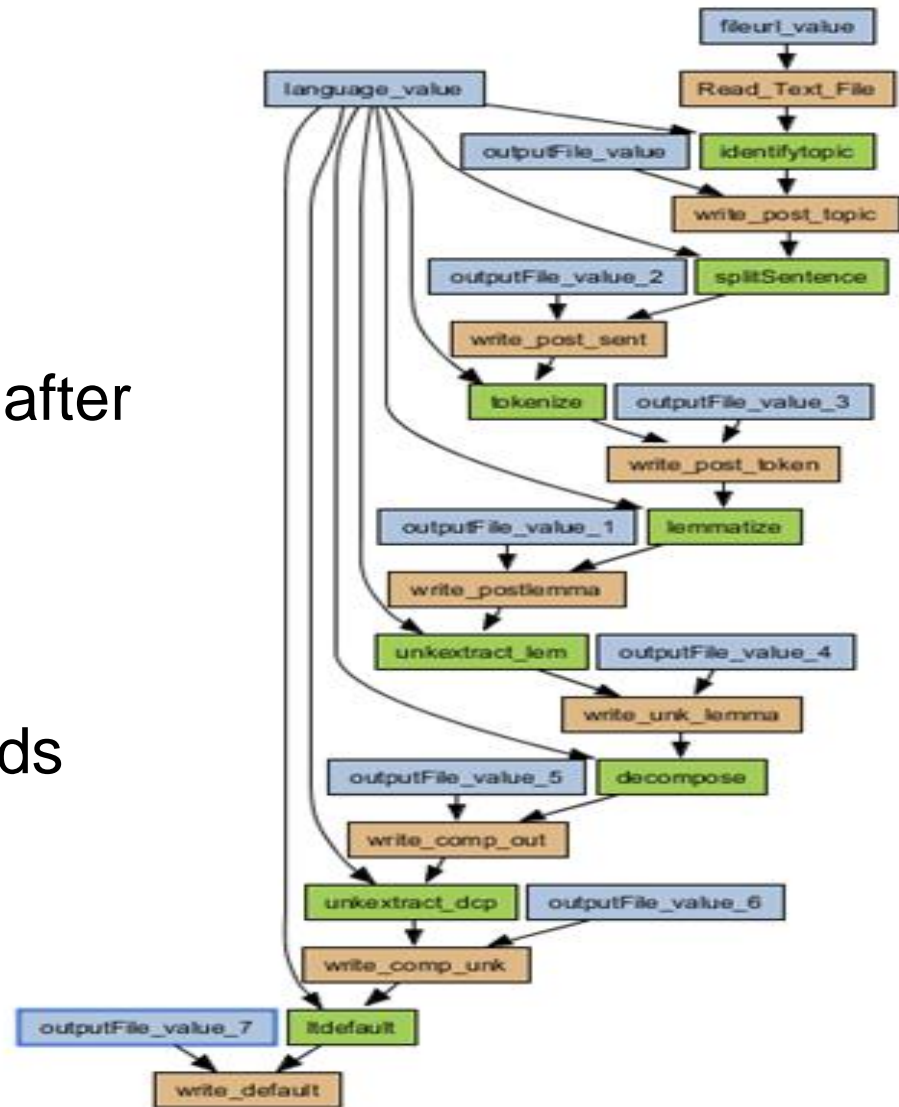
Assuming an error rate of 20%:

-> wrong: 408 K words

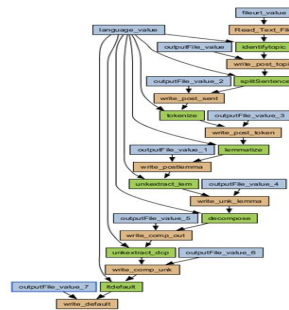
Complete Workflow

	tokens	coverage	error rate	wrong tokens
Europarl+JRC+EMEA	65.568.000			
lexicon lookup for	62.380.040	95,14%	1,29%	804.700
remaining	3.187.960			
decomposition for	1.205.200	37,8%	2,18%	26.270
remaining	1.982.760			
POS defaulting for	1.982.760	100%	20,58%	408.050
remaining	0			1.239.020
total correct assignments		98,11%		

- Input: ILSP-crawled file
- Workflow
 - All tools
 - (Extraction of unknowns after
 - Lemmatisation
 - Decomposition)
- Result
 - unknown / defaulted words
 - with POS proposals



DEMO ...



Panacea texts

Doc-ID (11 doc's)	2.xml	4.xml	22.xml	43.xml	61.xml	65.xml	98.xml	101.xml	103.xml	106.xml	118.xml	total
number tokens	548	317	395	640	226	2672	540	1244	1241	1438	416	9677
unknown lemmata	183	98	88	164	66	851	79	148	132	188	51	2048
coverage od lexicon in %	66,61	69,09	77,72	74,38	70,80	68,15	85,37	88,10	89,36	86,93	87,74	78,84
unknown after decomp	43	28	48	14	17	204	30	16	15	63	15	493
coverage of decomposer in %	76,50	71,43	45,45	91,46	74,24	76,03	62,03	89,19	88,64	66,49	70,59	75,93
errors of tokeniser	20	9	1		6	81	3		2	4	4	130 *)
errors of defaulter	5	2	8	3	1	19	3	5	3	8	3	60
coverage of defaulter in %	78,26	89,47	82,98	78,57	90,91	84,55	88,89	68,75	76,92	86,44	72,73	83,47
correct analysis in %	95,44	96,53	97,72	99,53	96,90	96,26	98,89	99,60	99,60	99,17	98,32	98,04

*) Tokeniser-Errors: 1.3%

Next

- Integrate into PANACEA platform
- (Cleanup some of the components)
- Extend the workflow to other **languages** (en, ...)
- Extend the Local Defaulter to other **features**
 - lemma, gender, number, ...
- **Target: From text to dictionary ...**
- Create proper Tagger / Parser input
 - originalstring casing no_readings (tag lemma)+
 - Multipart words?



Thank you!

