

Linguistic Processing Chains as Web Services: Initial Considerations

Maciej.Ogrodniczuk | @ipipan.waw.pl
Adam.Przepiorkowski



INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. J. K. Ordona 21, 01-237 Warszawa

Web Services and Processing Pipelines in HLT Workshop @ LREC
17-18 May 2010

WG 5.6 concentrates on:

- reviewing available web services implementing linguistic processing chains

WG 5.6 concentrates on:

- reviewing available web services implementing linguistic processing chains,
- specification of linguistic requirements on web services

WG 5.6 concentrates on:

- reviewing available web services implementing linguistic processing chains,
- specification of linguistic requirements on web services,
- selection of appropriate standards for the resources and tools to be integrated.

① D5R-3a, Dec 2009:

- identified 8 languages (English, German, Greek, Italian, Polish, Portuguese, Romanian, Spanish) for which mature web services providing linguistic tools exist, with a preference for web services implementing processing chains

① D5R-3a, Dec 2009:

- identified 8 languages (English, German, Greek, Italian, Polish, Portuguese, Romanian, Spanish) for which mature web services providing linguistic tools exist, with a preference for web services implementing processing chains,
- described each web service as a showcase of a processing chain

① D5R-3a, Dec 2009:

- identified 8 languages (English, German, Greek, Italian, Polish, Portuguese, Romanian, Spanish) for which mature web services providing linguistic tools exist, with a preference for web services implementing processing chains,
- described each web service as a showcase of a processing chain,
- intended to draw generalisations and preliminary recommendations.

1 D5R-3a, Dec 2009:

- identified 8 languages (English, German, Greek, Italian, Polish, Portuguese, Romanian, Spanish) for which mature web services providing linguistic tools exist, with a preference for web services implementing processing chains,
- described each web service as a showcase of a processing chain,
- intended to draw generalisations and preliminary recommendations.

2 D5R-3b, scheduled Dec 2010.

Scope of NLP functionalities offered by the reviewed tools

	Language identification	Sentence border detection	Tokenization	POS tagging / MSD	Named Entity recognition	Lemmatization	Parsing	TreeBank browsing	Co-occurrence annotation	Collocation extraction	Frequency analysis	Association measures	Semantic annotation	WordNet-related functionality	Thesaurus-related functionality	Lexicon access	Machine translation
WebLicht		•	•	•	•	•	•	•	•	•	•	•	•	•			
GATE		•	•	•	•	•	•		•	•	•	•	•	•			
IULA		•	•	•	•	•	•		•	•	•	•	•	•			
ILSP		•	•	•		•	•										
RACAI	•	•	•	•		•	•							•			•
WS-LexPI																•	
LXService		•	•	•													
WROCUT/ICS PAS		•	•	•		•	•		•	•	•	•		•			

Linguistic annotation output formats and tagsets

	Acknowledged standards						Proprietary formats	
	XML-based formats							
	LMF-XML	LMF-WordNet	MAF	SynAF	TIGER-XML	XCES	XCES proprietary extension	XML proprietary format
WebLicht			•		•			•
GATE			•	•				•
IULA						•		
ILSP						•		•
RACAI						•	•	•
WS-LexPI	•							
LXService								•
WROCUT/ICS PAS		•					•	

Linguistic annotation output formats and tagsets

	Acknowledged standards			Proprietary formats					
	XML-based formats								
	LMF-XML	LMF-WordNet	MAF	SynAF	TIGER-XML	XCES	XCES proprietary extension	XML proprietary format	Plain text proprietary format
WebLicht			•		•			•	
GATE			•	•				•	
IULA						•			
ILSP						•	•		
RACAI						•	•	•	•
WS-LexPI	•								
LXService							•	•	•
WROCUT/ICS PAS		•					•		

Standard tagsets					Proprietary tagsets				
CLAWS5									
EAGLES/ PAROLE									
MULTEXT-EAST									
Prague Dependency Treebank									
UPenn					•				
ICS PAS (PL)									
LX tagset (PT)									
RACAI tagset (EN, RO)									
SIMPLE-based tagset (IT)								•	
STTS (DE)									•

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence,
- XML is the lowest common denominator

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence,
- XML is the lowest common denominator,
- even proprietary formats are XML-related or at least interfacing XML world

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence,
- XML is the lowest common denominator,
- even proprietary formats are XML-related or at least interfacing XML world,
- standard vs. proprietary: proprietary formats are becoming „local standards” (→ ICS PAS tagset)

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence,
- XML is the lowest common denominator,
- even proprietary formats are XML-related or at least interfacing XML world,
- standard vs. proprietary: proprietary formats are becoming „local standards” (→ ICS PAS tagset),
- proprietary formats tend to maintain some extent of compatibility with standards (→ WebLicht TCF)

Generalizations:

- no common input/output format can be distinguished, no format is clearly taking precedence,
- XML is the lowest common denominator,
- even proprietary formats are XML-related or at least interfacing XML world,
- standard vs. proprietary: proprietary formats are becoming „local standards” (→ ICS PAS tagset),
- proprietary formats tend to maintain some extent of compatibility with standards (→ WebLicht TCF),
- tendency to standardize is obvious.

Three levels of interoperability:

Three levels of interoperability:

- technical interoperability — not relevant here (addressed in CLARIN deliverable D2R-6b)

Three levels of interoperability:

- technical interoperability — not relevant here (addressed in CLARIN deliverable D2R-6b),
- syntactic interoperability is maintained with XML-based interchange formats following official representation standards

Three levels of interoperability:

- technical interoperability — not relevant here (addressed in CLARIN deliverable D2R-6b),
- syntactic interoperability is maintained with XML-based interchange formats following official representation standards,
- semantic interoperability issues are still open, but appear to be solvable by providing formal mapping of proprietary categories to standard classes (→ ISOCat).

Three levels of interoperability:

- technical interoperability — not relevant here (addressed in CLARIN deliverable D2R-6b),
- syntactic interoperability is maintained with XML-based interchange formats following official representation standards,
- semantic interoperability issues are still open, but appear to be solvable by providing formal mapping of proprietary categories to standard classes (→ ISOCat).

Real interoperability between standards is ensured by providing conversion procedures.

Ideas for D5R-3b deliverable:

- a few (around three) case studies presenting chained Web Services with respect to standardization and interoperability requirements

Ideas for D5R-3b deliverable:

- a few (around three) case studies presenting chained Web Services with respect to standardization and interoperability requirements,
- synthesis referring to individual showcases

Ideas for D5R-3b deliverable:

- a few (around three) case studies presenting chained Web Services with respect to standardization and interoperability requirements,
- synthesis referring to individual showcases,
- summaries and recommendations

Ideas for D5R-3b deliverable:

- a few (around three) case studies presenting chained Web Services with respect to standardization and interoperability requirements,
- synthesis referring to individual showcases,
- summaries and recommendations

while several related issues are being covered by other deliverables, e.g.:

- review of the linguistic standards (→ D5C-4, see WG 5.7),
- Web Services requirements specification (→ D2R-6),
- Web Services creation (→ D2R-7).

Ideas for D5R-3b deliverable:

- a few (around three) case studies presenting chained Web Services with respect to standardization and interoperability requirements,
- syntesis referring to individual showcases,
- summaries and recommendations

while several related issues are being covered by other deliverables, e.g.:

- review of the linguistic standards (→ D5C-4, see WG 5.7),
- Web Services requirements specification (→ D2R-6),
- Web Services creation (→ D2R-7).

You can help: <http://www.clarin.eu/wp5/wg-56>.