Annotation Graphs and distant reading with Poio API and graf-python

Peter Bouda, pbouda@cidles.eu
Centro Interdisciplinar de Documentação Linguística e Social
Minde/Portugal

Poio API and GrAF Overview

- The goal: unified access to data from language documentation
- Support for common file formats: EAF (Elan), Toolbox, TCF (Weblicht), Typcraft XML, …
- Poio API adds description of "tier hierarchies" to GrAF
- With full power of annotation graphs
- Open source Python libraries:
  - http://media.cidles.eu/poio/poio-api/
  - http://media.cidles.eu/poio/graf-python/

Annotation Graphs

- the underlying data model for linguistic annotations
- pivot structure for linguistic data
- time vs. character offsets
- not hierarchical (but trees are also graphs)
- standoff annotation
- "It is important to recognize that translation into AGs does not magically create compatibility among systems whose semantics are different." [Bird & Liberman 2001]

GrAF

- GrAF: Graph Annotation Framework
- ISO 24612: Language resource management - Linguistic annotation framework (LAF)
- API and representation as data structures, not a file format
- GrAF/XML as XML representation
- Used for the MASC of the ANC
- Nodes, edges, regions, annotations, feature structures

Poio API

- Data/tier hierarchies
- Plugin mechanism for file formats
- Meta-data for additional information (Elan tier types etc.)
- Filters and filter chains for search

Search/Filters

PoioAnalyzer

File Edit About

Search 1 | New Search...

Utterances:

Words:

Morphemes:

Glosses: DEF

Translations:

Search Options
- AND
- OR
- NOT
- contained matches

Clear This Search | Close This Search | Save Searches... | **Search**

Files:

05IH_drying_sqash_IH_JL.eaf
05IH_ED_01_IH.eaf
05IH_ED_02_IH.eaf
05IH_ED_03_IH.eaf
05IH_ED_04_IH.eaf
05IH_maple_syrup_IH.eaf
05IH_New_Birth_IH.eaf
05IH_OAAT_IH_JL.eaf
05IH_REV_grizzly_bear.eaf
05IH_REV_haanp.eaf
05IH_REV_horses_JH.eaf
05IH_REV_napak.eaf
05IH_REV_twins.eaf
05IH_Richard_Mann_IH.eaf
05IH_Silent_Night_IH.eaf
A03_Saved_By_Grace_H.eaf
A04_Rich_Man_and_Lazarus1_H.eaf
A05_Rich_Man_and_Lazarus2_H.eaf
A06_Rich_Man_and_Lazarus_Song_H.eaf
B01_Matthew_24_1_H.eaf
B02_Matthew_24_2_H.eaf
B04_When_the_Roll_is_Called_Up_Yonder_H.

Result:

05IH_drying_sqash_IH_JL.eaf

WIC002 Wichąwąra kiisak wamącanąkga, rookra, suura hanąąc waigopnąkga, k'orok'oros jiinąkiregi 'eegi haruce waamąxaranąąną.

WORDS	Wichąwąra	kiisak	wamącanąkga,	rookra,	suura	hanąąc	waigopnąkga,	k'orok'oros	jiinąkiregi
MORPH	*wicąwą-ra*	*kiisak*	*wa-mąące-nąk-ga*	*rook-ra*	*suu-ra*	*hanąąc*	*wa-gigop-nąk-ga*	*k'orok'oros*	*jiinąk-ire-gi*
GLOSS	squash-DEF	half	OBJ.3PL-cut-POS.NTL-CONT	inside-DEF	seed-DEF	all	OBJ.3PL-hollow.out-POS.NTL-CONT	be.hollowed.out(OBJ.3SG)	become-SBJ.3PL-TOP

TRANS *Cut the sqash in half, scoop out the inside, and when it is hollowed out it may be sliced crosswise.*

WIC003 S~ooga waamąxaraanąkga, mąąhį huuną hošogara hikisge waamąxaranąkgiži nąąksikšįąra hižą wookiaxurucnąkga cąąkeja, hotakaceja hoicga taawus ważunąąna

WORDS	S~ooga	waamąxaraanąkga,	mąąhį	huuną	hošogara	hikisge	waamąxaranąkgiži	nąąksikšįąra
MORPH	*šooga*	*wa-hamąxara-nąk-ga*	*mąąhį*	*huuną*	*hošoga-ra*	*hikisge*	*wa-hamąxara-nąk-giži*	*nąąksik-šįąą-ra*
GLOSS	be.thick(OBJ.3SG)	OBJ.3PL-slice-POS.NTL-CONT	knife	handle	thickness-DEF	resemble(SBJ.3SG&OBJ.3SG)	OBJ.3PL-slice(SBJ.3SG)-POS.NTL-TOP	stick-be.hard(OBJ.3SG)

TRANS *The slices should be thick, or equal to the thickness of a knife handle. Use a sturdy stick and pick up circles of squash, or lace them through with a heavy stick. This stick may then be hung outside*

WIC005 Z~eegųgi, 'ee wii hotakacra wirokų hi'ųnąąną.

WORDS	Z~eegųgi,	'ee	wii	hotakacra	wirokų	hi'ųnąąną.
MORPH	*žeegų-gi*	*'ee*	*wii*	*hotakac-ra*	*wa-hirokų*	*hi-'ųų-nąą-ną*
GLOSS	thus-TOP	3EMPH	sun	warm.place-DEF	OBJ.3PL-utilize(SBJ.3SG)	APPL.INST-do/make-POT-DECL

TRANS *That way, it will utilize the hot rays of the sun to advantage.*

05IH_ED_01_IH.eaf

ED1002 wąąkšik hit'e raažra Mąąšųųrukąnąbiga hiigaire

WORDS	wąąkšik	hit'e	raažra	Mąąšųųrukąnąbiga	hiigaire
MORPH	*wąąkšik*	*hit'e*	*raaš-ra*	*mąąšųų-rukąnąp-ik-ga*	*hi-hige-ire*
GLOSS	Indian/person	speak	name-DEF	feather-be.shiny(OBJ.3SG)-DIM-PROP	1E.U-say.to-SBJ.3PL

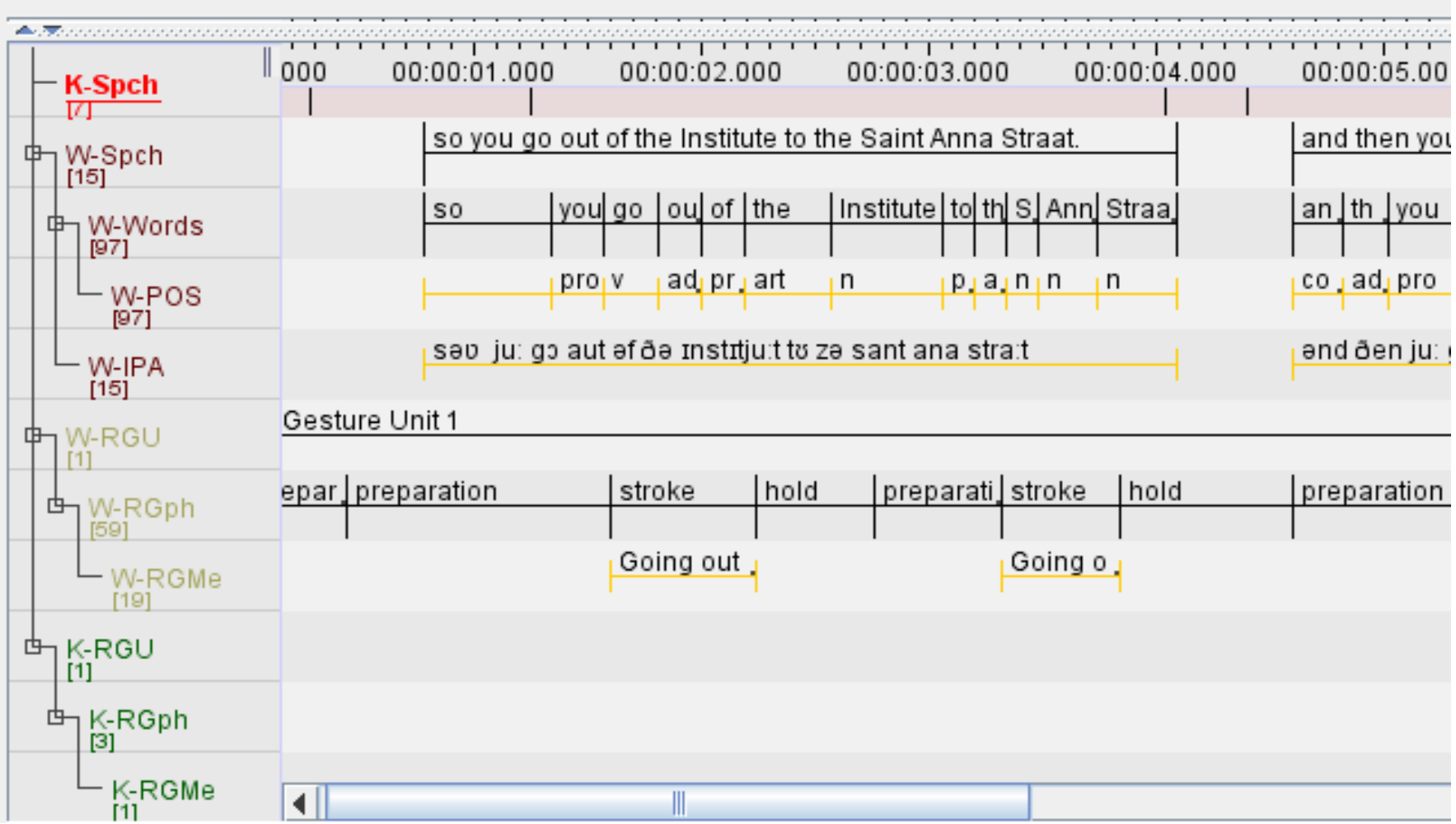
TRANS *my Hocank name is Shining Feather*

ED1003 'eegi haapte'e 'eegi Hoocak vaat'ekiane 'airera

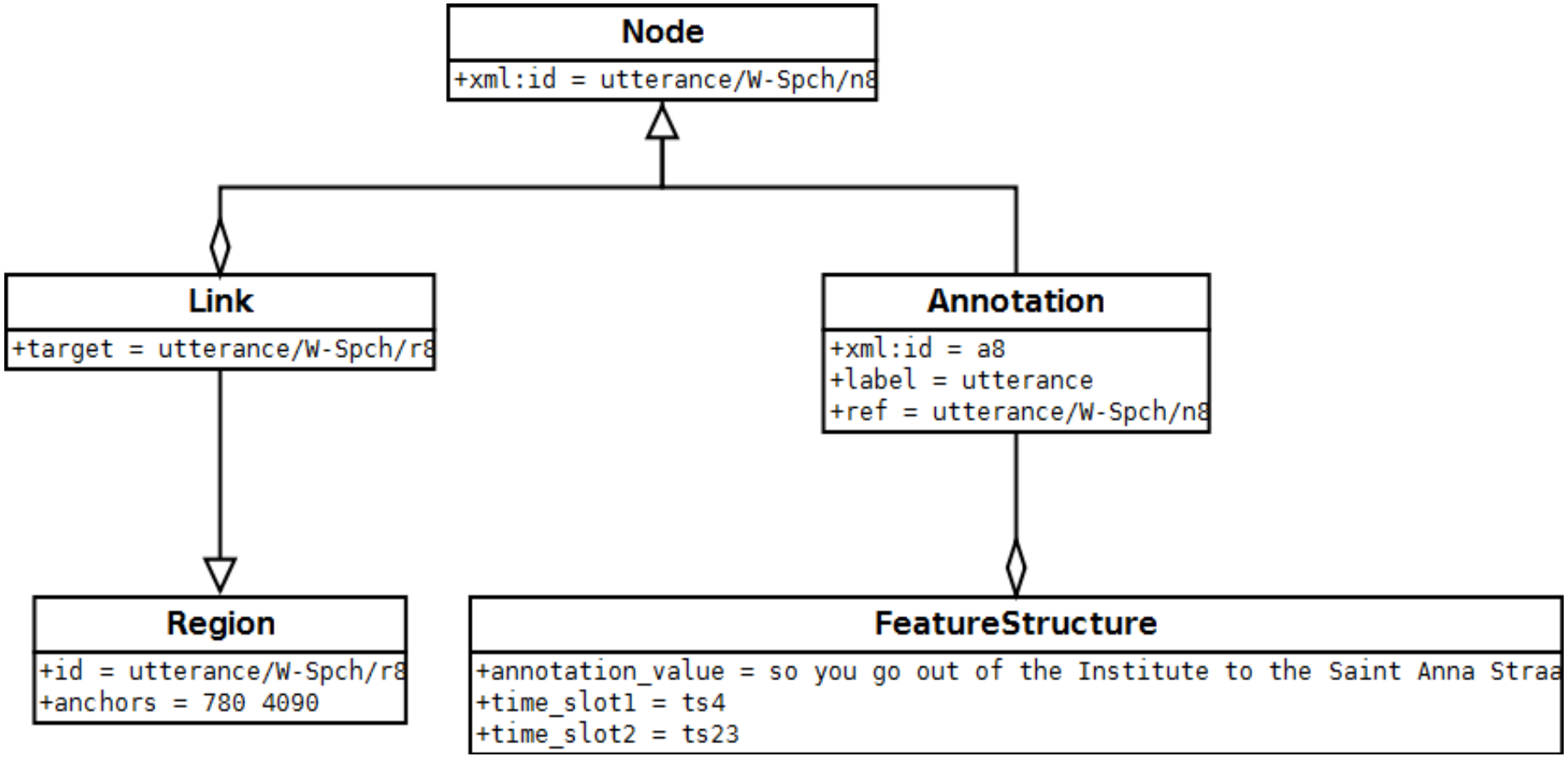
Add files... | Remove files

Quick Search:

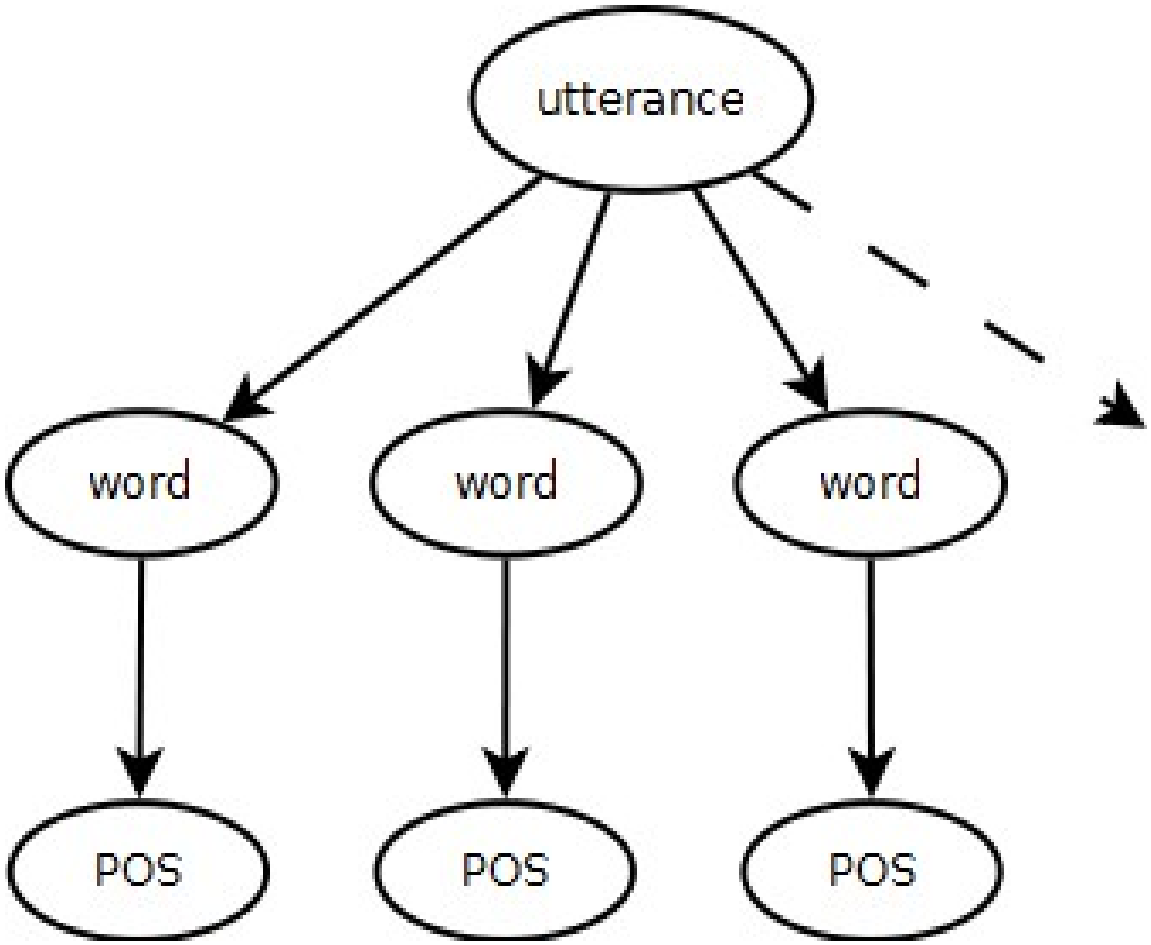
Mapping of file formats



GrAF entities



GrAF structure



Data hierarchies

```
[['utterance/K-Spch'],
 ['gestures/W-RGU',
   ['gesture_phases/W-RGph',
     ['gesture_meaning/W-RGMe']]],
 ['utterance/W-Spch',
   ['words/W-Words',
     ['part_of_speech/W-POS']],
   ['phonetic_transcription/W-IPA']],
 ['gestures/K-RGU',
   ['gesture_phases/K-RGph',
     ['gesture_meaning/K-RGMe']]]]
```

Editor from tier hierarchy

File Edit About

1	UTTERANCE	guš-īt:
	CLAUSE UNIT	guš-īt:
	WORD	guš-īt:
	WFW	say.PRS-3SG
	GRAID1	
	GRAID2	nc
	TRANSLATION	They say:
	COMMENT	

2	UTTERANCE	ki yag bādišā=(y)ē=at	
	CLAUSE UNIT	ki yag bādišā=(y)ē=at	
	WORD	ki	yag	bādišā=(y)ē=at
	WFW	SUB	one	king=IND=COP.PST.3SG
	GRAID1	comp	deti	np.h:s=cop:predp
	GRAID2			
	TRANSLATION	that there was a king.		
	COMMENT			

3	UTTERANCE	ē bādišā bi=m-ē wat-ī šār-ay wasat-(t)ā yakk tīr=i barγ-ē dāšt									
	CLAUSE UNIT	ē bādišā bi=m-ē wat-ī šār-ay wasat-(t)ā yakk tīr=i barγ-ē dāšt									
	WORD	ē	bādišā	bi=m-ē	wat-ī	šār-ay	wasat-(t)ā	yakk	tīr=i	barγ-ē	dāšt
	WFW	DEM	king	in=EMPH-DEM	RFL-GEN	town-GEN	middle-OBL	one	pole=IZ	electricity-IND	have.P
	GRAID1	det	np.h:a	prep	refl:poss	np:poss	np:l	deti	np:p	other	possv
	GRAID2										
	TRANSLATION	This king had a light post in the centre of his town									
	COMMENT										

Pipelines in Linguistics

- Idea of a workflow as in bioinformatics, for example
- graf-python and Poio API can be used in first step, when the data enters the pipeline
- GrAF is supported in GATE (General Architecture for Text Engineering) and Apache UIMA (Unstructured Information Management)
- Will become more important when data gets bigger and/or computational needs increase
- Scientific workflow systems: Kepler, Taverna, VisTrails, pypipegraph, ...

Using the Graphs

- Conversion: HTML, LaTex, etc.
- Interaction with other Python libraries
  - Transformation into networks (Python: networkx library), execution of graph-based algorithms
  - Corpus linguistics and computational linguistics (NLTK)
  - Scientific computation and statistics with numpy, scipy, pandas, etc., for example to feed cooccurence/collocation matrices, in Interactive Python (IPython)

Example 1: HTML output

- Conversion from:
  - Elan
  - Typecraft
  - Poio
  - Excel
- File format conversion as simple workflow
- Or: count word orders

Example 1: HTML output

- Conversion from:
  - Elan
  - Typecraft
  - Poio
  - Excel
- File format conversion as simple workflow
- Another: count word orders

Example 2: Semantics in dictionaries

- Data from QuantHistLing project of Michael Cysouw
- Data published as GrAF
- Digitized and annotated dictionaries from native South-American Languages
- Build a „translation graph“
- Visualization of polysemy of bodyparts
- All heads and translations into one big graph: how many paths of length 2 exist between two Spanish bodyparts?

Example 3: POS in translation tier

- Data from Johannes Helmbrecht of DoBeS project „Hoocąk“
- How does comparison work in Hoocąk?
- Add sub-tier for „ft“ in hierarchy and add POS tags from NLTK tagger as sub-nodes for each „ft“ node
- Search for POS tags „comparative“ or „superlative“ in English translation

Thanks!

pbouda@cidles.eu