CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

Annotation in AnnCor

Jan Odijk

CLARIAH Annotation Workshop
Amsterdam, 2017-05-11

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

OVERVIEW

- AnnCor
- Annotation in AnnCor
- Annotation of Syntactic Structures
  - Annotation Procedure & Guidelines
  - Preprocessing
  - Annotation Application
  - Annotation checks
- Search and Analysis Application
- Future Work

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

ANNCOR

- UU infrastructure project
- **Ann**otation of **Cor**pora (Dutch language only)
  - https://anncor.sites.uu.nl/ (under construction)
- Different types of annotations:
  - Syntactic structures (treebanks)
  - Discourse annotations
  - Error annotations
  - Narrative annotations
- Different Types of Corpora
  - Dutch CHILDES corpora (CHAT format)
  - Various UU CHAT format language acquisiton corpora
  - Learner Corpora
  - Newspapers
  - UU Narrative Corpora

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

SYNTACTIC ANNOTATION

- Assign a syntactic structure to each utterance in the corpus
- Corpus = CHILDES http://childes.talkbank.org/
- Annotated transcriptions of speech in child/adult interactions
- Dutch: 8 subcorpora, 534K utterances, 2.5 million tokens
- CHILDES sample CHILDES Sample2

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

PREPROCESSING

- Convert available metadata and annotations to suitable format (PaQu Plain Text Metadata Format)
  - Metadata conversion tool (chamd)
  - Convert CHAT Metadata format into PaQu metadata format
- Clean each utterance so that Alpino can deal with it
  - Cleaning tool (cleanCHILDESMD)
  - CHILDES sample after cleaning
- Bootstrap using the Alpino Parser
  - Example Parse

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

MANUAL ANNOTATION

- Manually verify and correct a small percentage (10-20%)
  - Creating a fully manually verified representative subset
  - And corrections for utterances that are likely to contain errors (based on certain heuristics)
- Manual Annotation
  - Based on expert's linguistic knowledge
  - in accordance with existing annotation guidelines
    - Part of speech guideline (Van Eynde 2005, 88 pages)
    - LASSY Annotation manual (Van Noord et al., 2011, 208 pages)
    - CGN Annotation Manual (Hoekstra et al., 2003; 77 pages)
- Update/adapt the guidelines to the specific corpus at hand

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

ANNOTATION APP

- TrED (**Tr**ee **Ed**itor) 2.0
  - https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-48F7-8
  - Part of the CLARIN infrastructure (Czech CLARIN / LINDAT project)
  - **Desktop** application for Windows, Linux, OS X
  - fully customizable and programmable graphical editor and viewer for tree-like structures
  - Very stable software!
  - screenshots

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

CHECKING ANNOTATIONS

- Fact of life:
  - Humans: intelligent but sloppy
  - Computers: stupid but precise
- What we need:
  - Intelligent and precise
- How? Combine work by humans & computers
- Selection of annotations checked by a different annotator
  - Discussed and guidelines updated
- All annotations checked by
  - Built in normalisation tool
  - Specially developed AnnCor Check Engine (ACE)

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

CHECKING ANNOTATIONS

- ACE
  - Checks for
    - Formal errors (serious errors)
    - Annotation errors (errors)
    - Likely errors (warnings)
  - 33 different types of errors, thousands of local configurations
  - False error message occurrences can be overruled by annotator
    - and need never be considered again
  - Error / Warning messages based on
    - Formal definition of well-formed syntactic structures
    - (some) guidelines in the annotation manuals
    - Statistics of configurations occurring in existing treebanks (LASSY-SMALL, CGN), manually adapted where needed

ACE ERROR REPORT SNAPSHOT

Status	File	Mother	Node	ErrNr	Arg1	Message	Arg2
Message						generated by anncor-checks.py version 0.14 on 2016-12-05T21:13:38	
Warning	checktest file.xml	top	--/conj	5	tsw	unlikely child of	conj
Warning	checktest file.xml	top	--/conj	13	cnj/tsw	unusual child rel/cat of	conj
Warning	checktest file.xml	top	--/conj	5	let	unlikely child of	conj
Error	checktest file.xml	top	--/conj	10	--	illegal child relation of	conj
Error	checktest file.xml	top	--/conj	12	--/let	illegal child rel/cat of	conj

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

PROBLEMS

- Wrongly transcribed / untranscribed words
  - *Foor* (*voor*), *dese* (*deze*), *saap* (*slaap*)
  - *Das* (*da(t i)s*)
  - *(i)khoe(f)nie(t) s(t)oel* ➔ *ikhoefniet stoel*
  - *maar nu gaat hij xxx.*
- False starts, stuttering and retracing
  - *Ik wil knu knuffel*
  - *Mama heb een… mama heeft een koek*
- Child utterances that are ill-formed in adult language
    - Pop gaat niet huil (finite verb in infinitival construction?)
    - boekje lesen [: lezen]
    - en ik **deze boekje** , he.

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

ANNOTATION CONSISTENCY

- Annotation Consistentency is extremely important
  - Even more important than correctness of the annotations
  - Determines the usability of the resource for research
- Small inter-annotator agreement experiment
  - Compare the syntactic structures for one sentence created independently by two annotators
  - 100 sentences, 3 annotators, 6 comparisons
  - Probably too small
  - Syntactic structure ➔a set of dependency tuples
  - Comparison by means of recall, precision, F-score
  - Average F-score of 86%
  - Cohen's kappa or related metrics?

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

SEARCH & ANALYSIS APPLICATION

- GrETEL original version
- GrETEL extended version being developed at UU
- New functionality
  - Upload one's own corpus incl metadata
    - Plain text, LASSY XML and CHAT format (more to follow)
  - Faceted search for filtering search results by metadata
  - Component for full analysis : select, group, sort, filter search results on arbitrary combinations of data and metadata elements (under development)
  - Additional user interface (to be developed)

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

FUTURE WORK

- Integrate re-parse option with Alpino with Bracketed Input
  - give hints to Alpino about the correct constituent structure by putting straight brackets around constituents
  - Phantom words, skip words, …
  - (followed by manual adaptation)
- Translate certain CHAT annotations into bracketed input
- Larger inter-annotator agreement test
- Tools for distributing the work over multiple annotators (?)
- Publish new versions of the data with updated transcriptions
- Create exports to CHAT
  - %MOR tier: part of speech tags
  - %GRA tier: syntactic structure

CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

Thanks for your Attention!

REFERENCES

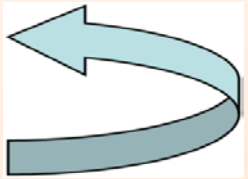
H. Hoekstra (2003). CGN Syntactische Annotatie. CGN Rapport. Utrecht, december 2003. http://www.mpi.nl/tools/synpathy/syn_prot.pdf s

G. Van Noord et al. (2011). Lassy Syntactische Annotatie Revision : 19455. LASSY Rapport, Groningen, 7 september 2011. http://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf

Frank Van Eynde (2005). Part of speech tagging en lemmatisering van het D-COI CORPUS. D-COI rapport, KU Leuven, juli 2005. http://www.ccl.kuleuven.be/Papers/DCOIpos.pdf

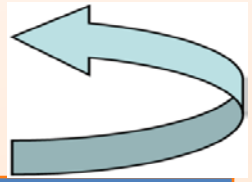
CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

CHAT METADATA



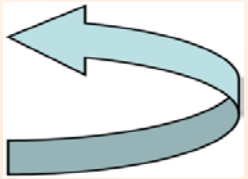
- @UTF8
- @Begin
- @Languages: nld
- @Participants: SAR Sarah Child , JAC Jacqueline Mother
- @ID: nld|vankampen|SAR|4;8.03||||Child|||
- @ID: nld|vankampen|JAC||||Mother|||@Date: 05-JAN-1994
- @Date: 30-JAN-1994@Transcriber: Simone Boezewinkel
- *SAR: en xxx die xxx?
- *JAC: ja.

PAQU METADATA



Meta All	UTT1	UTT2
##META text charencoding = UTF8 ##META date date = 1994-01-30 ##META text transcriber = Simone Boezewinkel	##META int uttid = 0 ##META text speaker = SAR ##META text origutt = en xxx die xxx? ##META text role = Child ##META text name = Sarah ##META text language = nld ##META text corpus = vankampen ##META text age = 4;8.03 ##META int months = 56 ##META text sex = ##META text group = ##META text SES = ##META text role = Child ##META text education = ##META text custom = en xxx die xxx?	##META int uttid = 1 ##META text speaker = JAC ##META text origutt = ja. ##META text role = Mother ##META text name = Jacqueline ##META text language = nld ##META text corpus = vankampen ##META text age = ##META text months = ##META text sex = ##META text group = ##META text SES = ##META text role = Mother ##META text education = ##META text custom = ja.

CHILDES SAMPLE 1



*JAC: weet je hoev [//] hoeveelste het vandaag is?

*JAC: ehm.

*SAR: nee , ik [/] ik moet in de microfoon praten , nee.

…

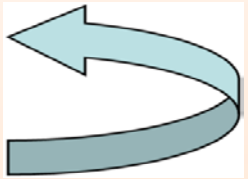
*SAR: mama?

*JAC: ja.

*SAR: heb jij kaasjes [= kaarsen] ?

- Retracing [//], repetition [/], explanation ([= …]
- And many others ….
- (Van Kampen Sarah46 utts 13-15 (14-16))

CLEAN SAMPLE 1



*JAC: weet je hoev hoeveelste het vandaag is?

*JAC: ehm.

*SAR: nee , ik moet in de microfoon praten , nee.

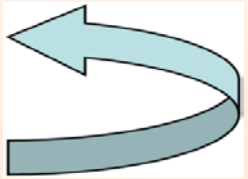
...

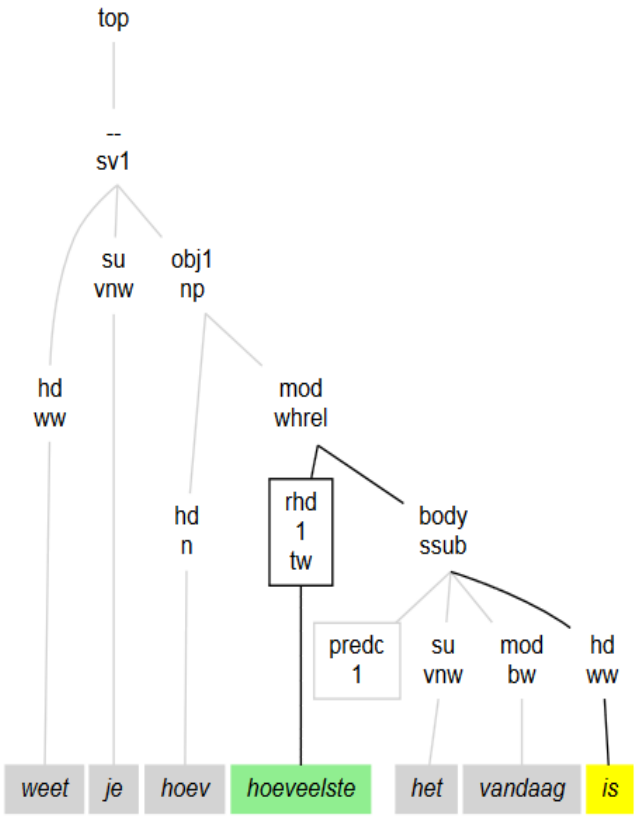
*SAR: mama?

*JAC: ja.

*SAR: heb jij kaasjes ?

ALPINO



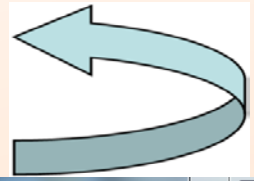


opslaan als: dot

bestand: childesdutchmeta::VanKampen/sarah46.14

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities

TRED SHORTCUTS



TrEd ver. 2.5049 Default(4/4): D:\jodijk\Dropbox\jodijk\presentations\2017\201705-11 Annotation Workshop\data\FL1-3(1EM2RvdV3JvE)\FL1(1EM-2RvdV-3JvE)\VanKampen_sarah46_u00000000013.xml

File Node Tree View Macros Setup Help

Mode: Alpino

Style: Alpino

weet je hoev hoeveelste

1/1

List Of Macros ...

Current Mode

All Modes

Reload Macros

Add Terminal Left	(ALT+l)
Add Terminal Right	(ALT+r)
Add Terminal Right Punt	(ALT+p)
AssignCatOfParent	(U)
CjCj	(ALT+c)
Cut Subtree	(CTRL+x)
Cut Terminal	(ALT+x)
DpDp	(ALT+d)
Edit Cat	(c)
Edit Comments (Annotate)	(CTRL+ALT+a)
Edit Index	(i)
Edit Pos	(p)
Edit Rel	(r)
Edit Root	(l)
Edit Word	(w)
Flatten	(CTRL+ALT+f)
Fold/unfold current subtree	(space)
Move to parent directory and Next File	(F12)
Move to skipped directory and Next File	(ALT+s)
MwuFlatten	(CTRL+ALT+m)
More...	

top

num TW(rang,prenom,stan)
hoev (hoev : hoev)

hd
verb WW(pv,tg
weet (weet : w

obj1
whsub

whd :1
ang,prenom,stan)
eveelste : hoeveelste)

body
ssub

predc :1

mod
adv BW()
vandaag (vandaag : vandaag)

su
det VNW(pers,pron,stan,red,3,ev,onz)
het (het : het)

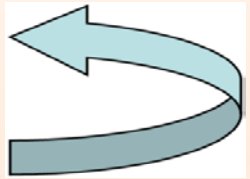
hd
verb WW(pv,tgw,ev)
is (ben : zijn)

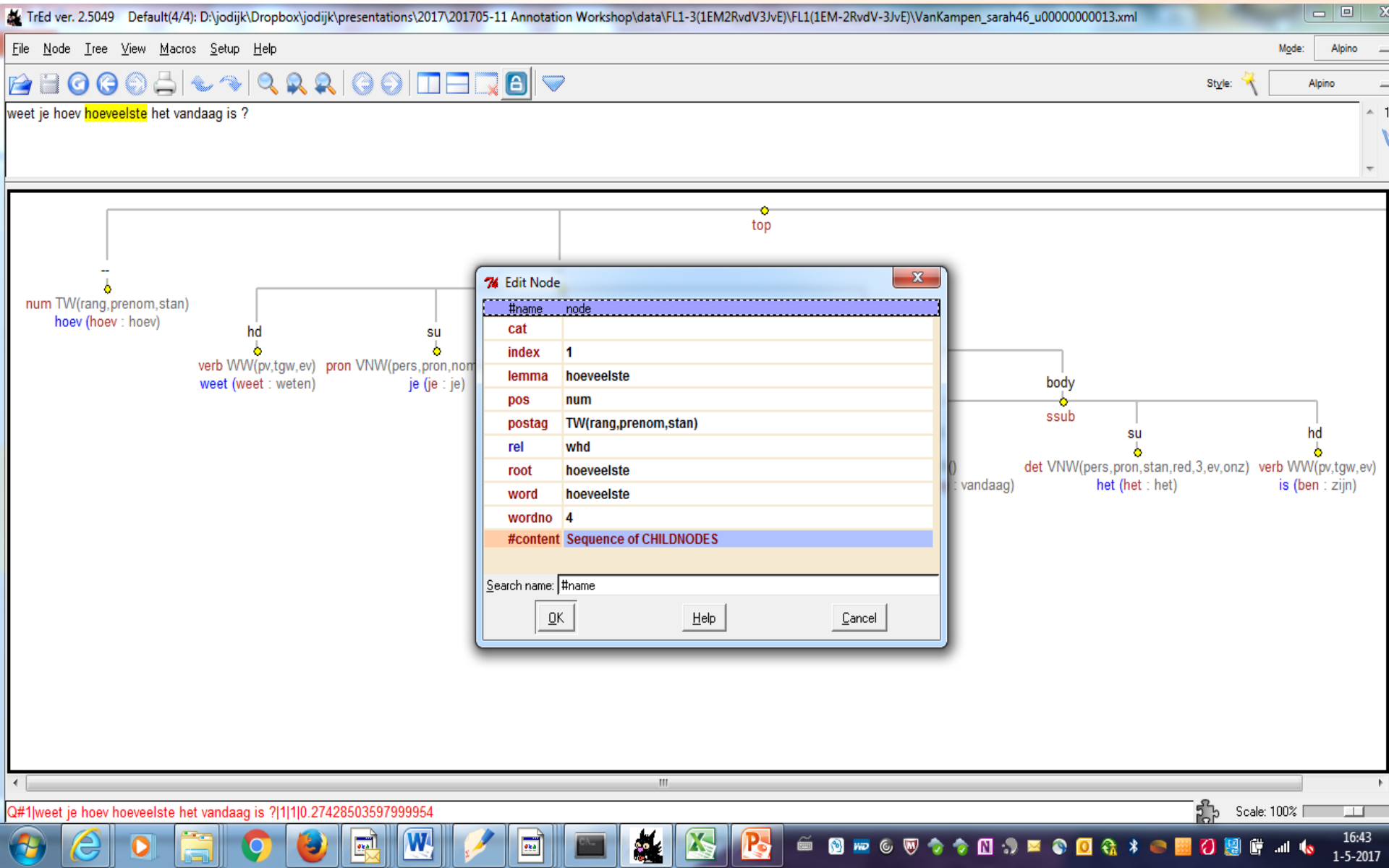
Q#1|weet je hoev hoeveelste het vandaag is ?|1|1|0.27428503597999954

Scale: 100%

16:44
1-5-2017

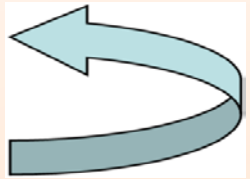
EDIT NODE

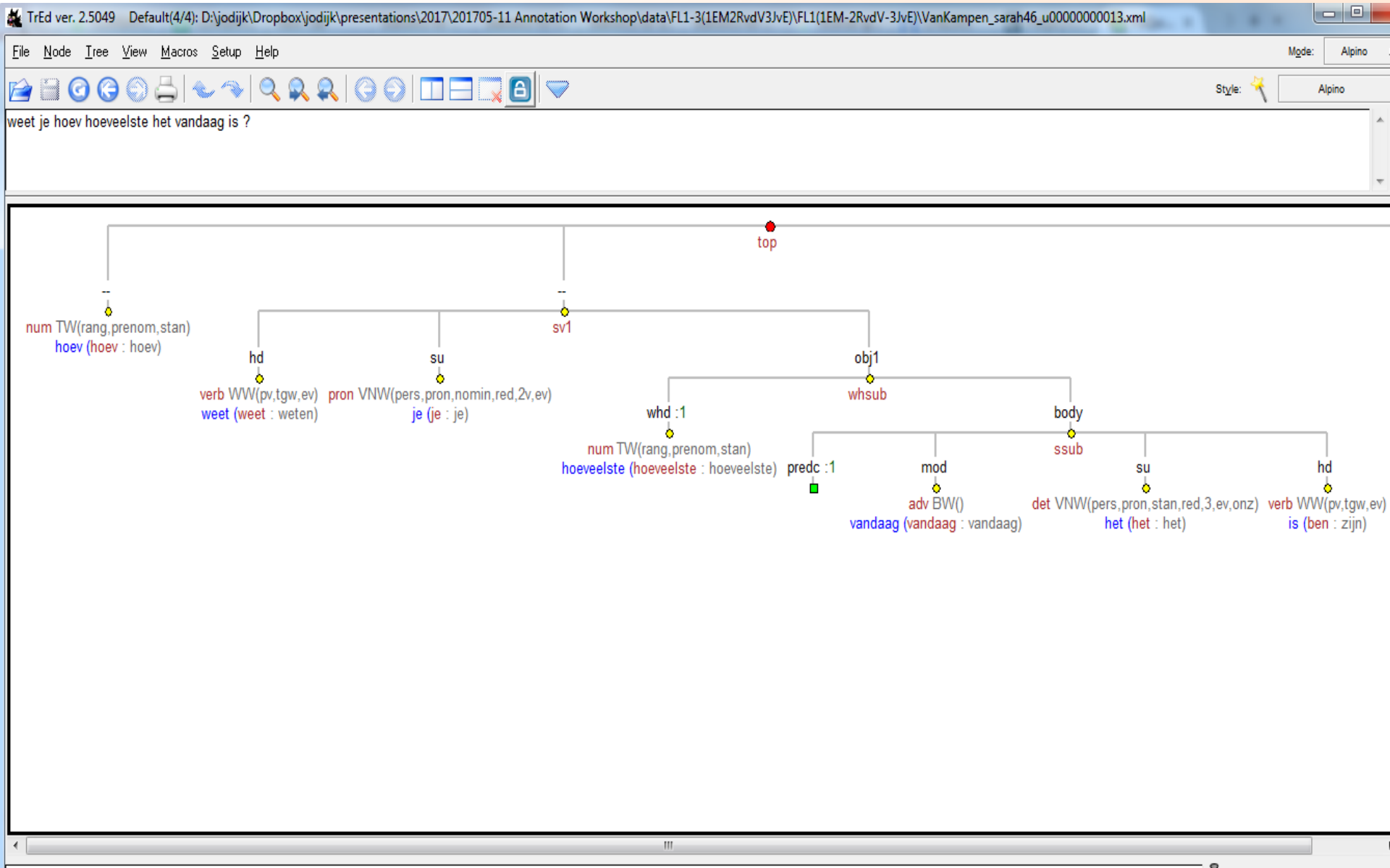



TrEd ver. 2.5049 Default(4/4): D:\jodijk\Dropbox\jodijk\presentations\2017\201705-11 Annotation Workshop\data\FL1-3(1EM2RvdV3JvE)\FL1(1EM-2RvdV-3JvE)\VanKampen_sarah46_u00000000013.xml
File Node Tree View Macros Setup Help
Mode: Alpino
Style: Alpino
weet je hoev hoeveelste het vandaag is ?
top
num TW(rang,prenom,stan)
hoev (hoev : hoev)
hd
verb WW(pv,tgw,ev)
weet (weet : weten)
su
pron VNW(pers,pron,nom
je (je : je)
body
ssub
su
det VNW(pers,pron,stan,red,3,ev,onz)
het (het : het)
hd
verb WW(pv,tgw,ev)
is (ben : zijn)
: vandaag)
Edit Node
#name node
cat
index 1
lemma hoeveelste
pos num
postag TW(rang,prenom,stan)
rel whd
root hoeveelste
word hoeveelste
wordno 4
#content Sequence of CHILDNODES
Search name: #name
OK
Help
Cancel
Q#1|weet je hoev hoeveelste het vandaag is ?|1|1|0.27428503597999954
Scale: 100%
16:43
1-5-2017

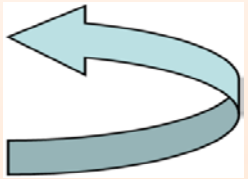
CLARIAH
Common Lab Research Infrastructure
for the Arts and Humanities

CORRECTED TREE





CHILDES SAMPLE 2



*JAC: zo moet ie.

*LAU: doet de ande nou?

*JAC: xxx afgekoeld , de andere nou.

*JAC: is [//] weet je , die andere hoeft er niet bij.

*JAC: ehm.

*LAU: alleen foor [/] foor dese daar.

*LAU: foor kindertjes.

*JAC: hij moet xxx branden.

- Speech event annotations: retracing ([//]), repetition ([/])
- Incomprehensible words (xxx)
- Contains wrong spellings (imitation of phonetics)