Parsing the BNC with RASP4UIMA

Øistein E. Andersen
Julien Nioche
Ted Briscoe
John Carroll

LREC 2008

The British National Corpus

- British English
- Late 20th cent.
- 100M words
- Balanced
  - written (90%) / spoken (10%)
  - informative (75%) / imaginative (25%)
  - books (60%) / periodicals (25%) / other publications / letters and diaries / speeches and stageplays

The British National Corpus

- Sentences, tokens
- Part of speech
- Written: italics, sections, …
  Spoken: turn taking, laughter, false starts, …
- Metadata: title, source, genre, date, author's name, sex and origin, …

The British National Corpus

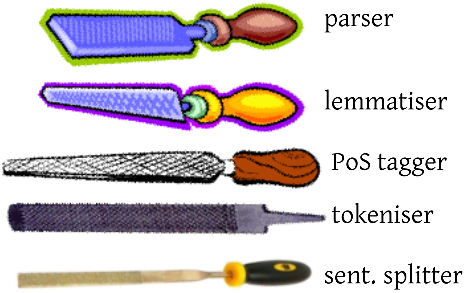
<xml>

- XML edition 2007
- Revisions and corrections
- Still no syntactic annotation

</xml>

The Robust Accurate Statistical Parsing system

- Domain-independent, robust parsing system for English
- Free for research purposes
- PoS tagger:
  - CLAWS-style tags
  - *c.* 150 parts of speech
- Parser output:
  - Trees
  - Grammatical relations



parser

lemmatiser

PoS tagger

tokeniser

sent. splitter

The RASP system

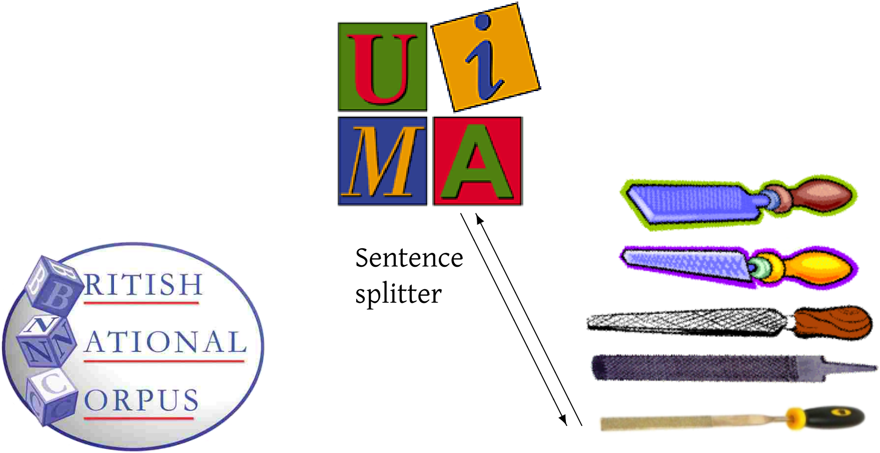
The Unstructured Information Management Architecture

- Analysis framework for adding structured data to natural-language documents
- IBM, Apache

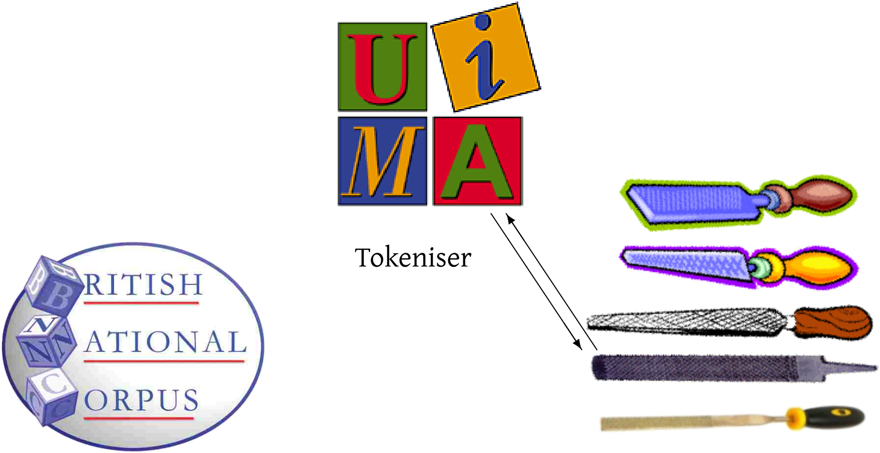
- Read/import a document
- Interface with *processing engines*
- Write/export annotated document

U
i
M
A

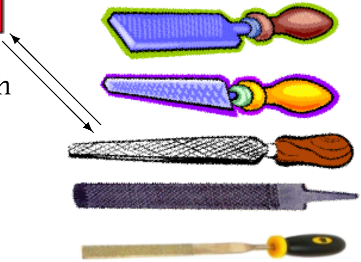
B
N
C
RITISH
ATIONAL
ORPUS



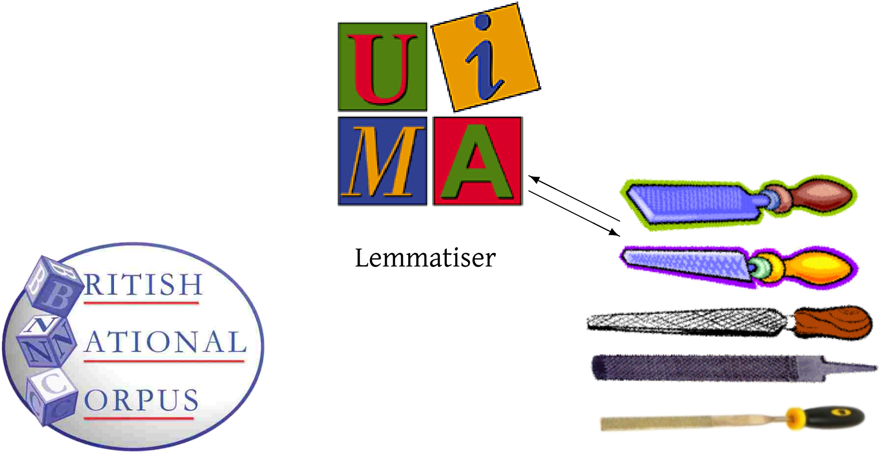
Sentence splitter



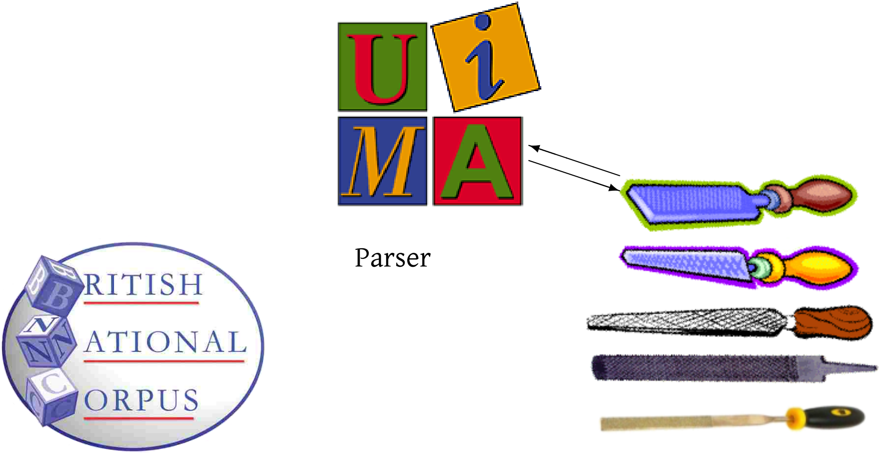
Tokeniser



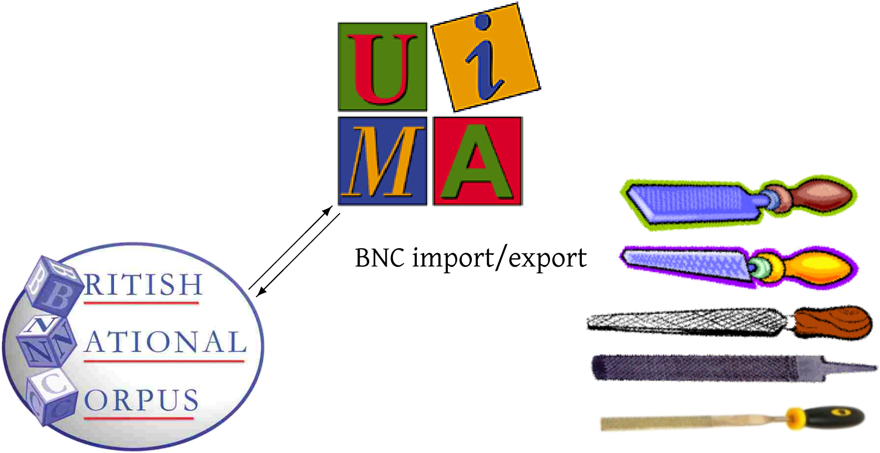
Part-of-speech tagger



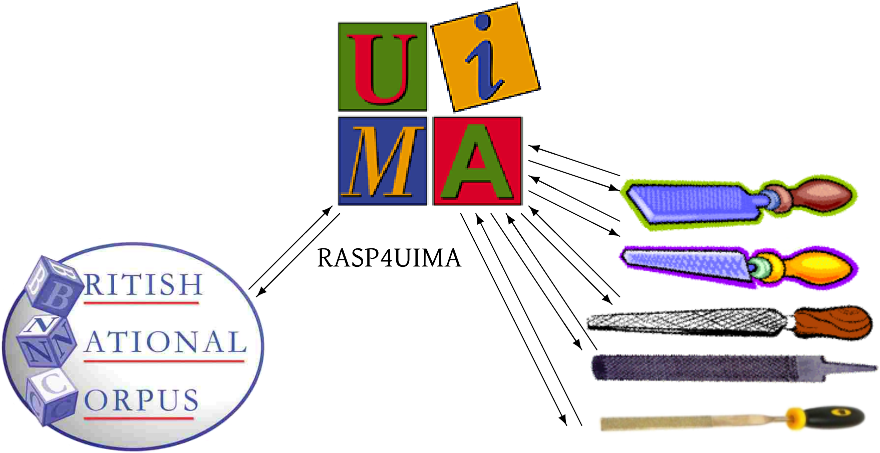
Lemmatiser



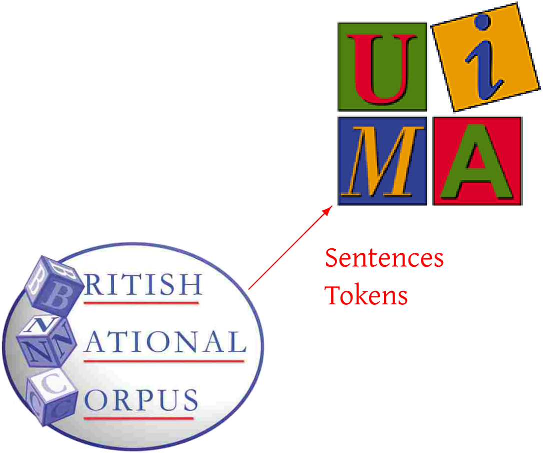
Parser



BNC import/export

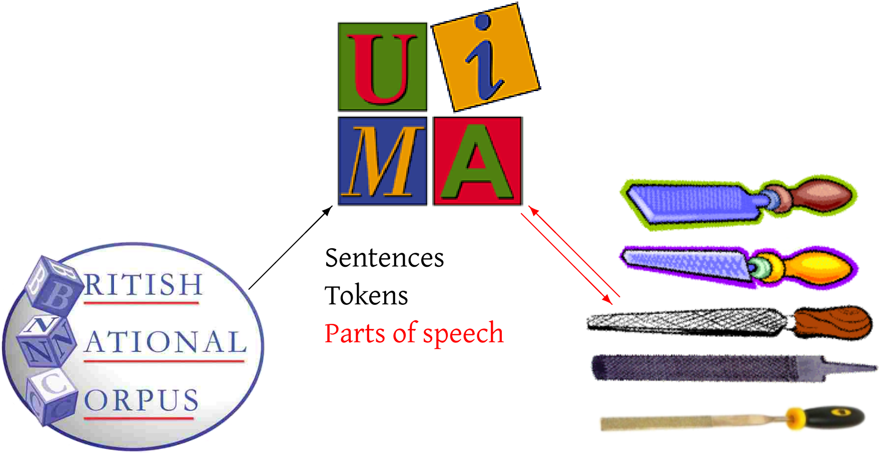


RASP4UIMA



Sentences

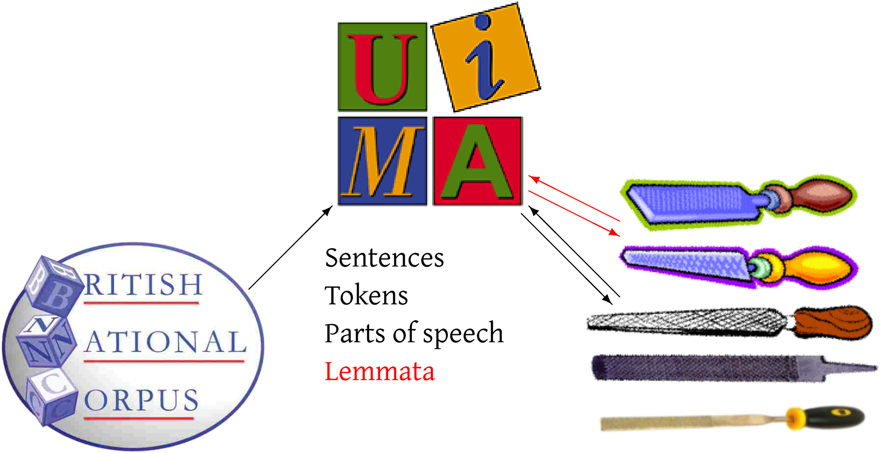
Tokens



Sentences

Tokens

Parts of speech

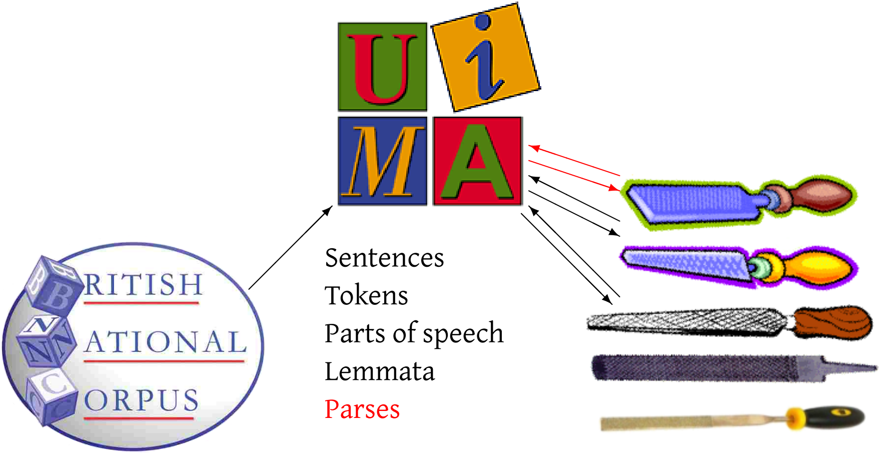


Sentences

Tokens

Parts of speech

Lemmata



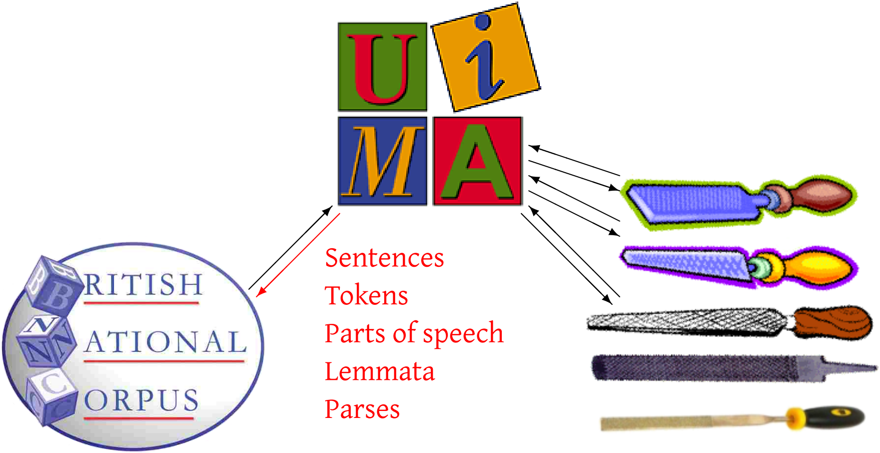
Sentences

Tokens

Parts of speech

Lemmata

Parses



Sentences

Tokens

Parts of speech

Lemmata

Parses

Example sentence

```
<s n="1">
  <trunc>
    <w c5="UNC" hw="any" pos="UNC">Any </w>
  </trunc>
  <w c5="PNI" hw="anyone" pos="PRON">anyone </w>
  <w c5="PNQ" hw="who" pos="PRON">who </w>
  <w c5="AJ0-VVN" hw="dissolved" pos="ADJ">dissolved </w>
  <mw c5="AV0">
    <w c5="AV0" hw="more" pos="ADV">more </w>
    <w c5="CJS" hw="than" pos="CONJ">than </w>
  </mw>
  <w c5="UNC" hw="½" pos="UNC">½</w>
  <gap desc="formula"/>
  <w c5="PRP" hw="in" pos="PREP"> in </w>
  <w c5="NN1" hw="rivers/lakes" pos="SUBST"
```

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> ½ <gap desc="formula"/> <w> </w>*in rivers/lakes is n't gon na forget his pilgrimage , y'know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

`<trunc>`*Any*`</trunc>` *anyone who dissolved* `<mw>`*more than*`</mw>` ½ `<gap desc="formula"/>` `<w> </w>`*in rivers/lakes is n't gon na forget his pilgrimage , y'know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> ½ <gap desc="formula"/> <w> </w>*in rivers/lakes is n't gon na forget his pilgrimage , y'know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> ½ <gap desc="formula"/> <w> </w>*in rivers/lakes is n't gon na forget his pilgrimage , y'know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> *½* <gap desc="formula"/> <w> </w>*in rivers/lakes is n't gon na forget his pilgrimage , y'know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> *½* <gap desc="formula"/> <w> </w>*in rivers / lakes is n't gon na forget his pilgrimage , y' know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> ½ <gap desc="formula"/> <w> </w>*in rivers / lakes is n’t gon na forget his pilgrimage , y’ know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

<trunc>*Any*</trunc> *anyone who dissolved* <mw>*more than*</mw> ½ <gap desc="formula"/> <w> </w>*in rivers / lakes is n't gon na forget his pilgrimage , y' know .*

Corrections/adaptations during reading

- Empty tokens
- Multi-word expressions
- Incomplete tokenisation
- Gaps
- Spoken data :-(

anyone who dissolved more than ½ [gap] in rivers / lakes

is n't gon na forget his pilgrimage , y' know .

Processing

- Part-of-speech tagger
- Lemmatiser
- Parser

anyone	*who*	*dissolved*	*more*	*than*	*½*	*[gap]*	*in*	*rivers*	*/*	*lakes*
PN1	PNQS	VVD	DAR	CSN	MC	&FO	II	NNL2	CC	NN2

is	*n't*	*gon*	*na*	*forget*	*his*	*pilgrimage*	*,*	*y'*	*know*	*.*
VBZ	XX	VVG	TO	VV0	APP$	NN1	,	PPY	VV0	.

Processing

- Part-of-speech tagger
- Lemmatiser
- Parser

anyone	*who*	*dissolve+ed*	*more*	*than*	*½*	*[gap]*	*in*	*river+s*	*/*	*lake+s*
PN1	PNQS	VVD	DAR	CSN	MC	&FO	II	NNL2	CC	NN2

be+s	*not+*	*gon*	*na*	*forget*	*his*	*pilgrimage*	*,*	*y'*	*know*	*.*
VBZ	XX	VVG	TO	VV0	APP$	NN1	,	PPY	VV0	.

Processing

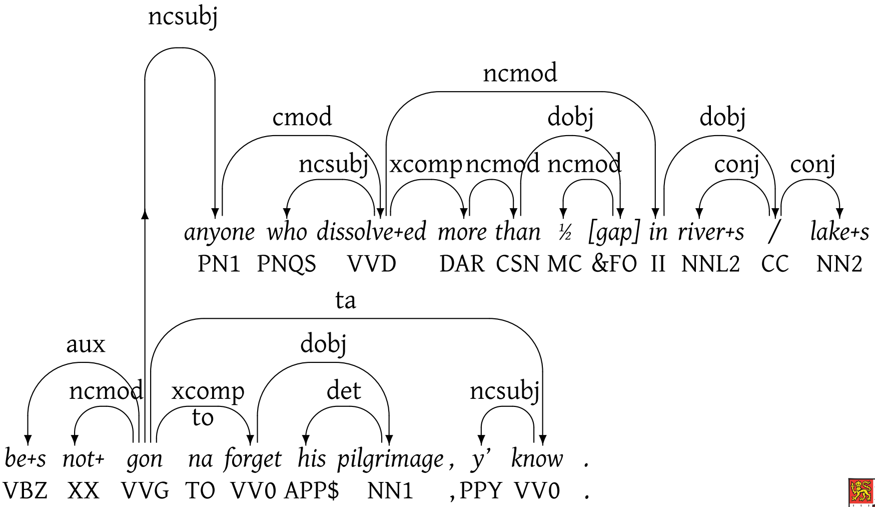
- Part-of-speech tagger
- Lemmatiser
- Parser

anyone	*who*	*dissolve+ed*	*more*	*than*	*½*	*[gap]*	*in*	*river+s*	*/*	*lake+s*
PN1	PNQS	VVD	DAR	CSN	MC	&FO	II	NNL2	CC	NN2

be+s	*not+*	*gon*	*na*	*forget*	*his*	*pilgrimage*	*,*	*y'*	*know*	*.*
VBZ	XX	VVG	TO	VV0	APP$	NN1	,	PPY	VV0	.

Processing

- Part-of-speech tagger
- Lemmatiser
- Parser


ncsubj
ncmod
cmod
dobj
dobj
ncsubj
xcomp
ncmod
ncmod
conj
conj
anyone who dissolve+ed more than ½ [gap] in river+s / lake+s
PN1 PNQS VVD DAR CSN MC &FO II NNL2 CC NN2
ta
aux
dobj
ncmod
xcomp
to
det
ncsubj
be+s not+ gon na forget his pilgrimage , y' know .
VBZ XX VVG TO VV0 APP$ NN1 , PPY VV0 .

```
<s n="1">
  <trunc>
    <w c5="UNC" hw="any" pos="UNC">Any </w>
  </trunc>
  <w n="1" c5="PNI" hw="anyone" pos="PRON"
     rpos="PN1" lem="anyone">anyone </w>
  <w n="2" c5="PNQ" hw="who" pos="PRON"
     rpos="PNQS" lem="who">who </w>
  <w n="3" c5="AJ0-VVN" hw="dissolved" pos="ADJ"
     rpos="VVD" lem="dissolve" affix="+ed">dissolved </w>
  <mw c5="AV0">
    <w n="4" c5="AV0" hw="more" pos="ADV"
       rpos="DAR" lem="more">more </w>
    <w n="5" c5="CJS" hw="than" pos="CONJ"
       rpos="CSN" lem="than">than </w>
  </mw>
  <w n="6" c5="UNC" hw="½" pos="UNC" rpos="MC"
```

```
<w n="20 21" c5="VVB-NN1" hw="y'know" pos="VERB"
    rpos="PPY VV0" lem="y' know">y'know</w>
<c n="22" c5="PUN" rpos="." lem=".">.</c>
<grlist parse="1" score="-40.848">
  <gr type="ncsubj" head="14" dep="1"/>
  <gr type="cmod" subtype="_" head="1" dep="3"/>
  <gr type="ncsubj" head="3" dep="2"/>
  <gr type="ncmod" subtype="_" head="3" dep="8"/>
  <gr type="xcomp" subtype="_" head="3" dep="4"/>
  <gr type="ncmod" subtype="_" head="4" dep="5"/>
  <gr type="dobj" head="5" dep="7"/>
  <gr type="ncmod" subtype="_" head="7" dep="6"/>
  <gr type="dobj" head="8" dep="10"/>
  <gr type="conj" head="10" dep="9"/>
  <gr type="conj" head="10" dep="11"/>
  <gr type="aux" head="14" dep="12"/>
  <gr type="ncmod" subtype="_" head="14" dep="13"/>
```

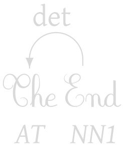
Available resources

RASP4UIMA

- DigitalPebble: *http://www.digitalpebble.com*

Parsed BNC XML

- Oxford Text Archive (OTA)
- *En attendant Godot*



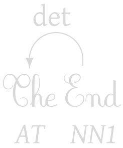
Available resources

RASP4UIMA

- DigitalPebble: *http://www.digitalpebble.com*

Parsed BNC XML

- Oxford Text Archive (OTA)
- *En attendant Godot*



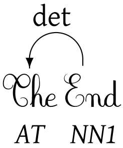
Available resources

RASP4UIMA

- DigitalPebble: *http://www.digitalpebble.com*

Parsed BNC XML

- Oxford Text Archive (OTA)
- *En attendant Godot*



det

The End

AT *NN1*