Institut für
Technische Informatik und
Kommunikationsnetze

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich
Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo

Character Stream Parsing of Mixed-lingual Text

Harald Romsdorfer and Beat Pfister

Text Analysis for TTS: Traditional

- "... Studio. They record the next record. Then ..."
  ⇒ [ðe͡ɪ ? ðə nekst ?]

they	ðe͡ɪ
record	?
the	ðə
next	nekst
record	?

Text Analysis for TTS: SVOX, . . .

- ". . . Studio. They record the next record. Then . . . "
  ⇒ [ðe͡ɪ rɪˈkɔːd ðə nekst ˈrekɔːd]

they	ðe͡ɪ	personal pronoun
record	rɪˈkɔːd	verb
the	ðə	determiner
next	nekst	adjective
record	ˈrekɔːd	noun

⇒ Syntactic analysis is a must for TTS synthesis!

Text Analysis for TTS: Problems I & II

Contracted Word Forms & Ambiguous Punctuation Symbols

- "It's in St. Mary's St. Mary's at home."
  ⇒ [ɪts ɪn ? mɛ͡əriz ? mɛ͡əriz ət hə͡ʊm]

it's	ɪts	?
in	ɪn	preposition
st.	?	?
mary's	m'ɛ͡əriz	?
st.	?	?
mary's	m'ɛ͡əriz	?
at	ət	preposition
home	hə͡ʊm	?

Text Analysis for TTS: Problems I & II

Contracted Word Forms & Ambiguous Punctuation Symbols

- "It's in St. Mary's St. Mary's at home."
  ⇒ [ɪts ɪn sənt meə̯riz striːt] [meə̯riz ət həʊ̯m]

it	ɪt	personal pronoun	mary	m'eə̯ri	proper noun
's	s	auxiliary "be"	's	z	auxiliary "be"
in	ɪn	preposition	at	ət	preposition
st.	sənt	noun title "Saint"	home	həʊ̯m	noun
mary's	m'eə̯riz	noun (possessive form)	.		full stop
st	striːt	noun "street"			
.		full stop			

⇒ Use syntactic words instead of graphemic words as tokens!

Text Analysis for TTS: Problems III

Multi-word lexemes

- "He's in fine conditions in fine."
  ⇒ [hiz ɪn faɪn kəndɪʃ(ə)nz ɪn faɪni]

he	hi	personal pronoun
's	z	auxiliary "be"
in	ɪn	preposition
fine	faɪn	adjective
conditions	kəndɪʃ(ə)nz	noun
in fine	ɪn faɪni	adverb

Text Analysis for TTS: Problems IV

Languages without word delimiter: Chinese, Japanese, ...

研究生 亦般年闹大

他宅研究 生 命起源

Text Analysis for TTS: Problems IV

Languages without word delimiter: Chinese, Japanese, ...

研究生	亦般	年闹	大
yan2-jiu1-sheng1	yi4-ban1	nian2-ling2	da4
'Master student'	'generally'	'age'	'old'

他	宅	研究	生 命起源
ta1	zhai	yan2-jiu1	sheng1-ming4-qi3-yuan2
'He'	'doing'	'research'	'the origin of life'

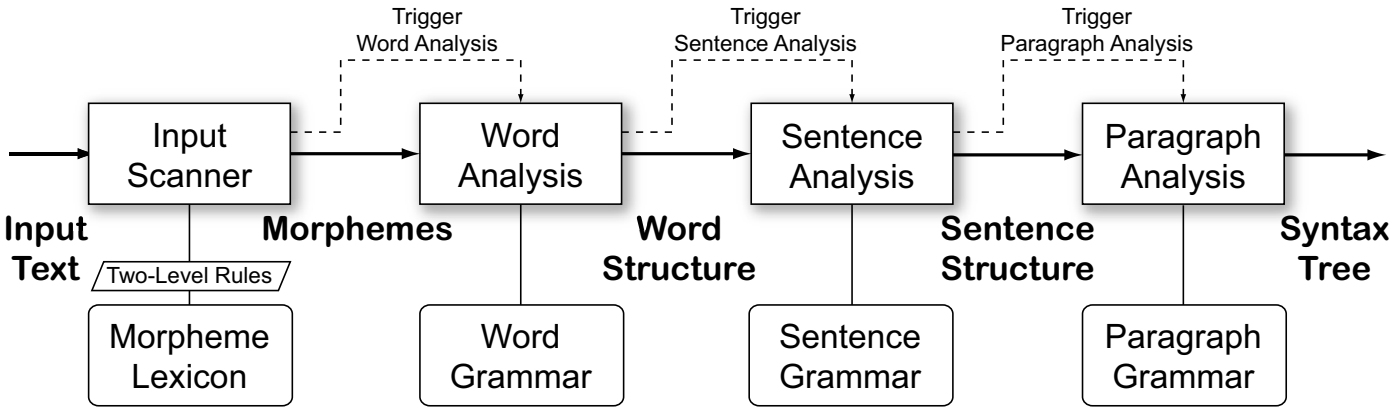
⇒ Boundary of syntactic words depends on morphological and syntactic context!

Basic Problem

$\Rightarrow$ Syntactic words are the input to syntactic analysis.

$\Rightarrow$ But boundaries of syntactic words depend on result of syntactic analysis!

Text Analysis of polySVOX



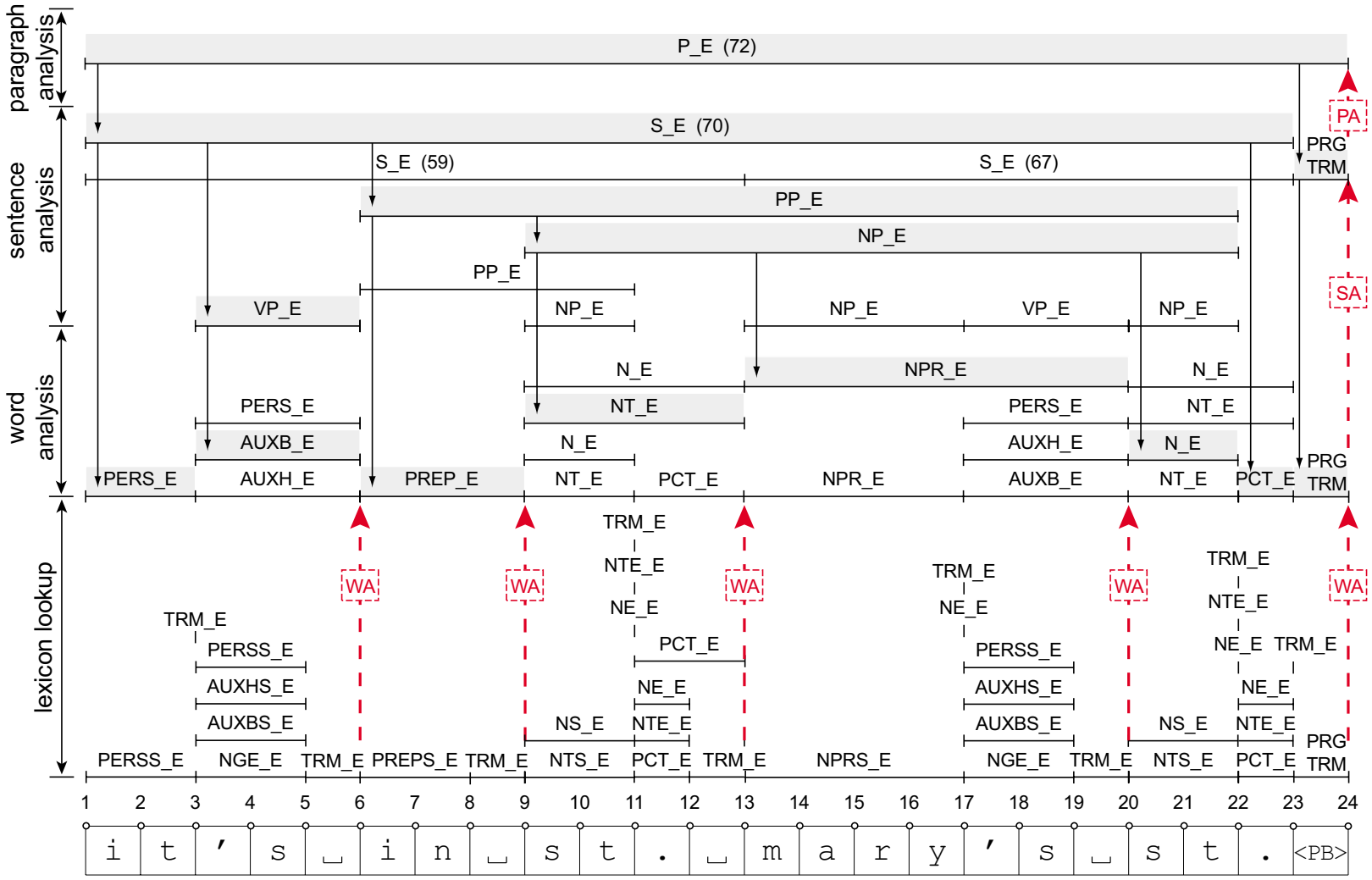
⇒ Scan input text character by character

⇒ Treat blanks, punctuation symbols, etc. like any other character

⇒ Define word boundaries to be blanks, “empty” symbols and certain punctuation symbols

⇒ Each module provides all alternatives for subsequent module

⇒ Each module triggers next module at unambiguous analysis positions


paragraph analysis
sentence analysis
word analysis
lexicon lookup
P_E (72)
S_E (70)
S_E (59)
S_E (67)
PP_E
NP_E
PP_E
VP_E
NP_E
NP_E
VP_E
NP_E
N_E
NPR_E
N_E
PERS_E
NT_E
PERS_E
NT_E
AUXB_E
N_E
AUXH_E
N_E
PERS_E
AUXH_E
PREP_E
NT_E
PCT_E
NPR_E
AUXB_E
NT_E
PCT_E
PRG
TRM
PA
SA
WA
TRM_E
NTE_E
NE_E
PCT_E
PERSS_E
AUXHS_E
AUXBS_E
NS_E
PERSS_E
NGE_E
TRM_E
PREPS_E
NTS_E
NPRS_E
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
i t ' s ␣ i n ␣ s t . ␣ m a r y ' s ␣ s t . <PB>

Text Analysis for TTS: Chinese/Japanese texts

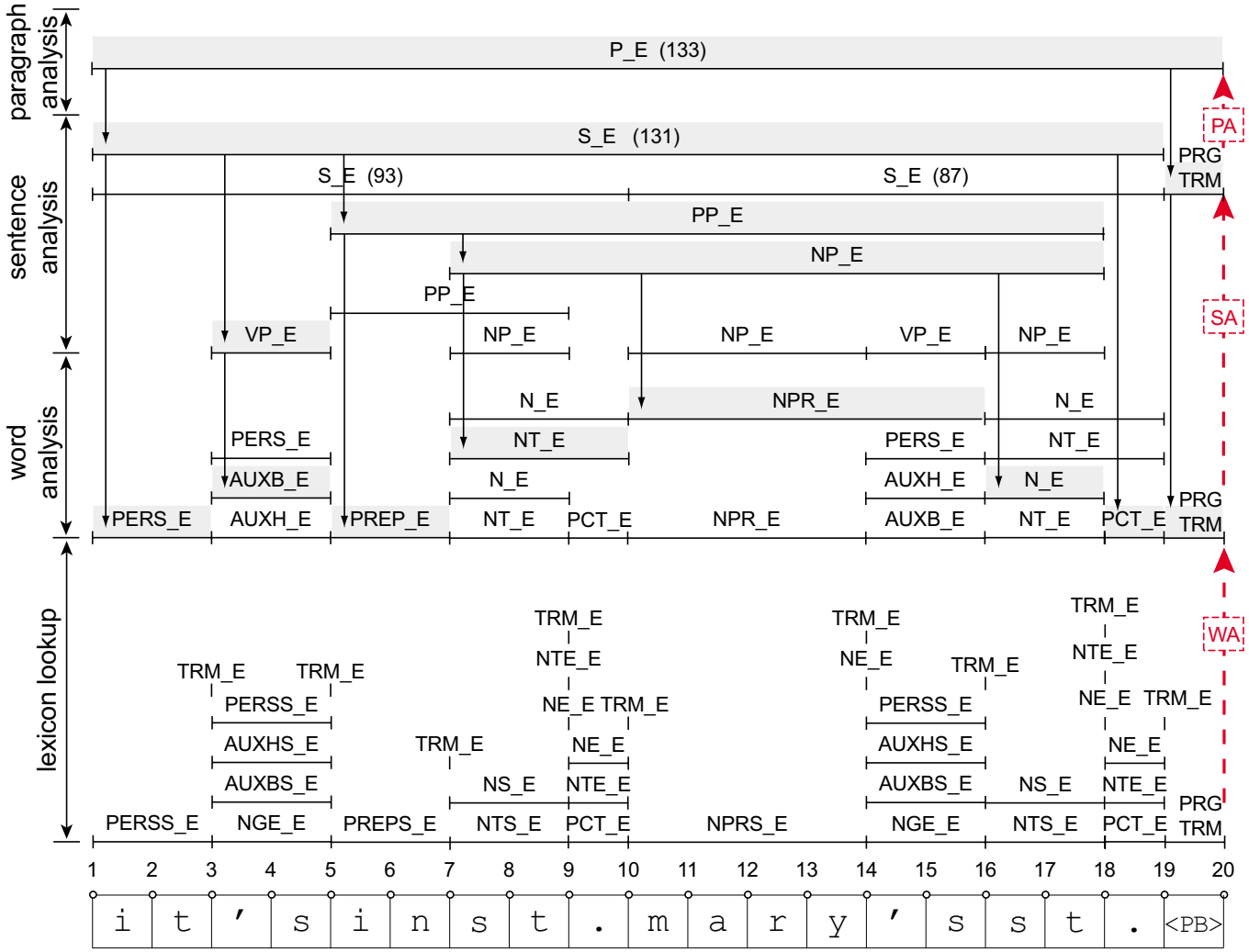
研究生	亦般	年闹	大
yan2-jiu1-sheng1	yi4-ban1	nian2-ling2	da4
'Master student'	'generally'	'age'	'old'

他	宅	研究	生 命起源
ta1	zhai	yan2-jiu1	sheng1-ming4-qi3-yuan2
'He'	'doing'	'research'	'the origin of life'

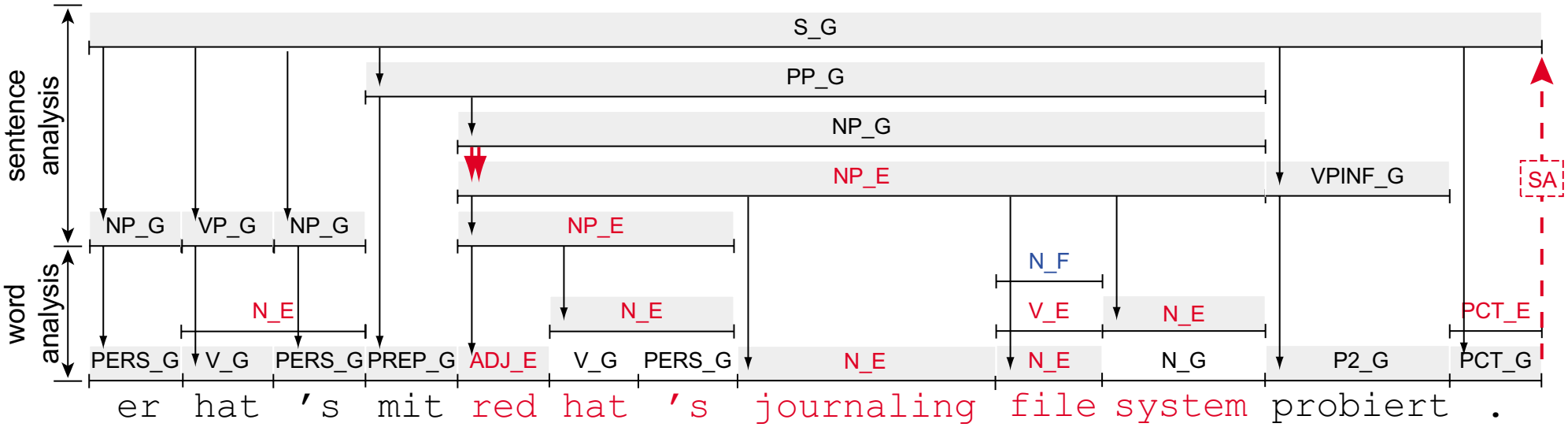
"Equivalent" example in English:

"it'sinst.mary'sst." ⇒ "It's in St. Mary's St."

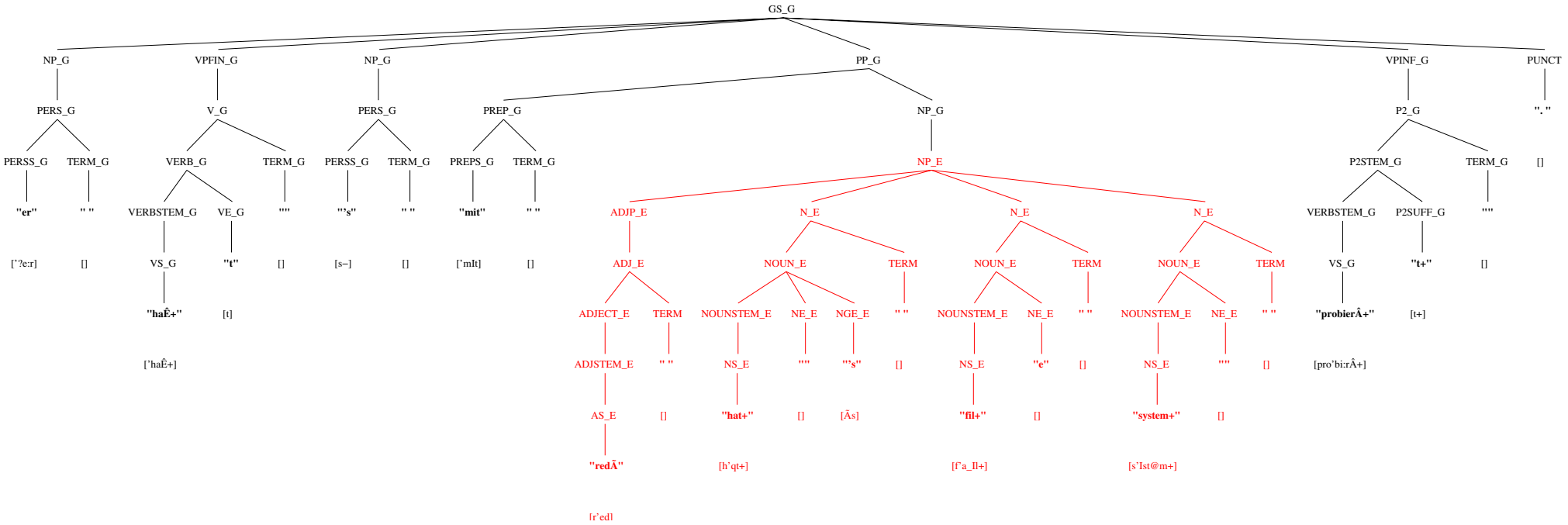
"itsinstwice." ⇒ "It sins twice."


paragraph analysis
sentence analysis
word analysis
lexicon lookup
P_E (133)
S_E (131)
S_E (93)
S_E (87)
PP_E
NP_E
PP_E
VP_E
NP_E
NP_E
VP_E
NP_E
N_E
NPR_E
N_E
PERS_E
NT_E
PERS_E
NT_E
AUXB_E
N_E
AUXH_E
N_E
PERS_E
AUXH_E
PREP_E
NT_E
PCT_E
NPR_E
AUXB_E
NT_E
PCT_E
PA
SA
WA
PRG
TRM
TRM_E
PERSS_E
AUXHS_E
AUXBS_E
NGE_E
NTE_E
NE_E
NS_E
NTS_E
PREPS_E
NPRS_E
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
i t ' s i n s t . m a r y ' s s t . <PB>

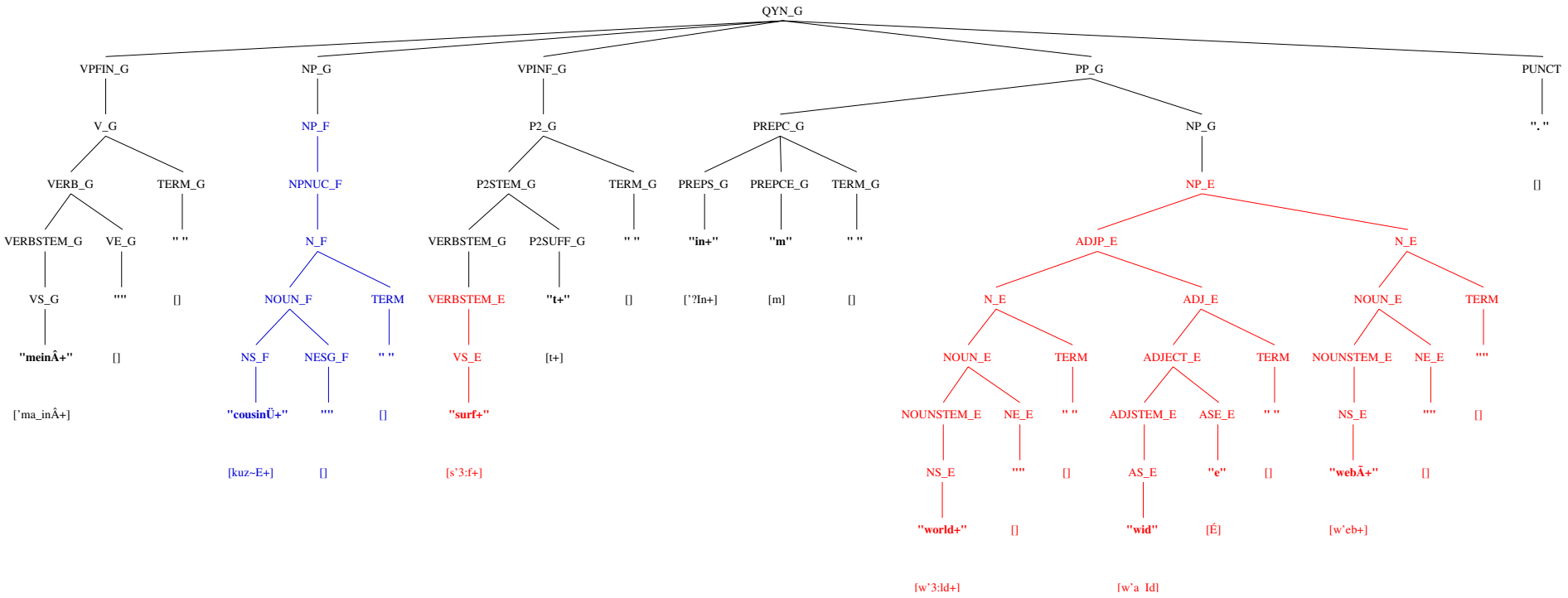
Mixed-lingual Sentences



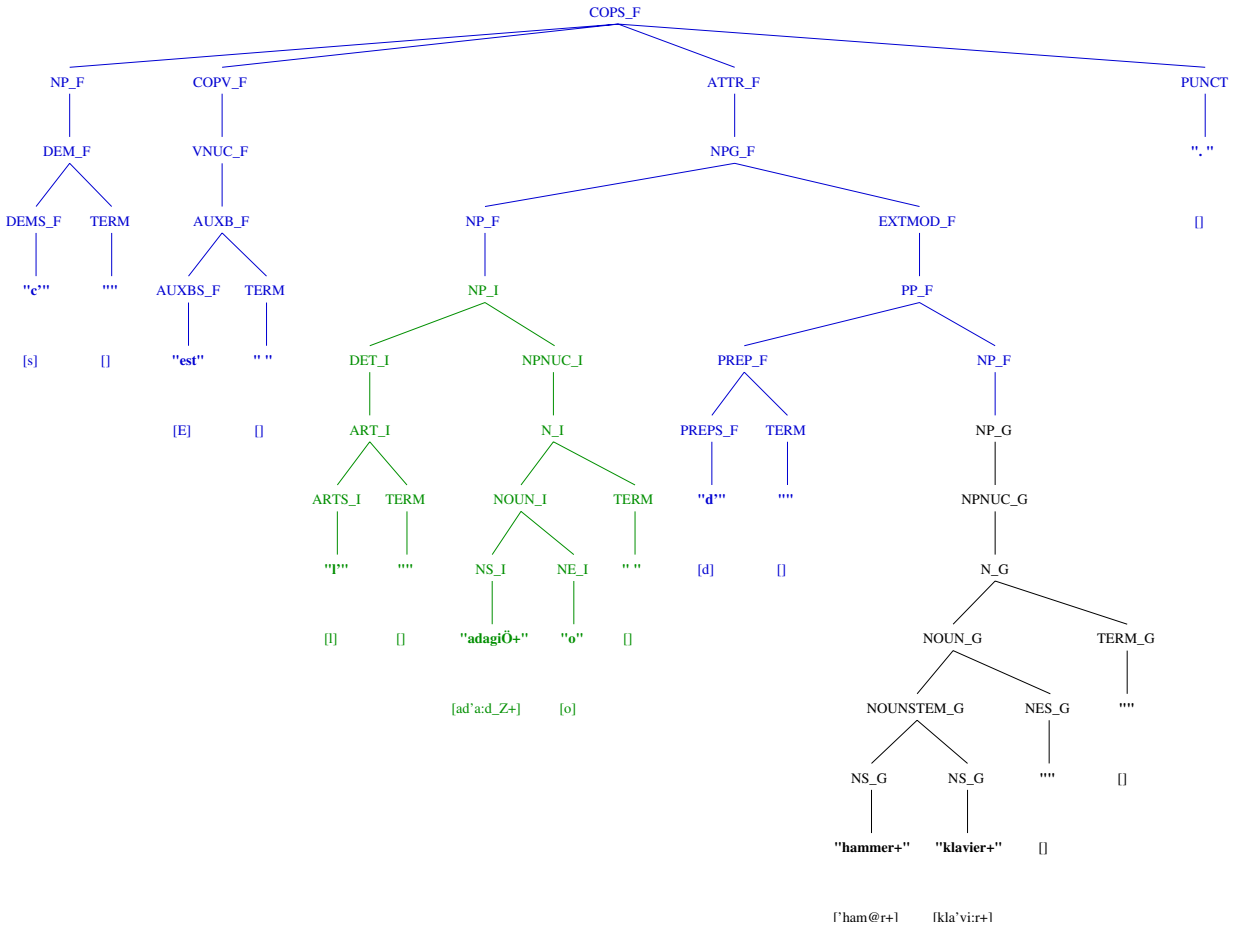
(He tried it with Red Hat's journaling file system.)



"Er hat's mit Red Hat's File System probiert."



"Mein Cousin surft im World Wide Web."



"C'est l'Adagio d'Hammerklavier."

<http://www.tik.ee.ethz.ch/~spr/SVOX/polysvoxdemo/>

Thank you!