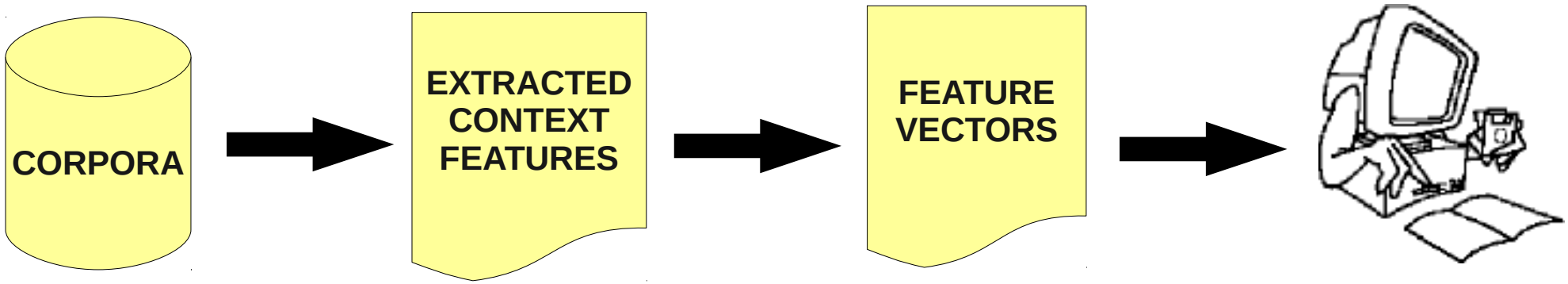
Systém pro automatickou extrakci lingvistických kontextů z textových korpusů

Implementace s využitím prostředí Treex

Lenka Smejkalová

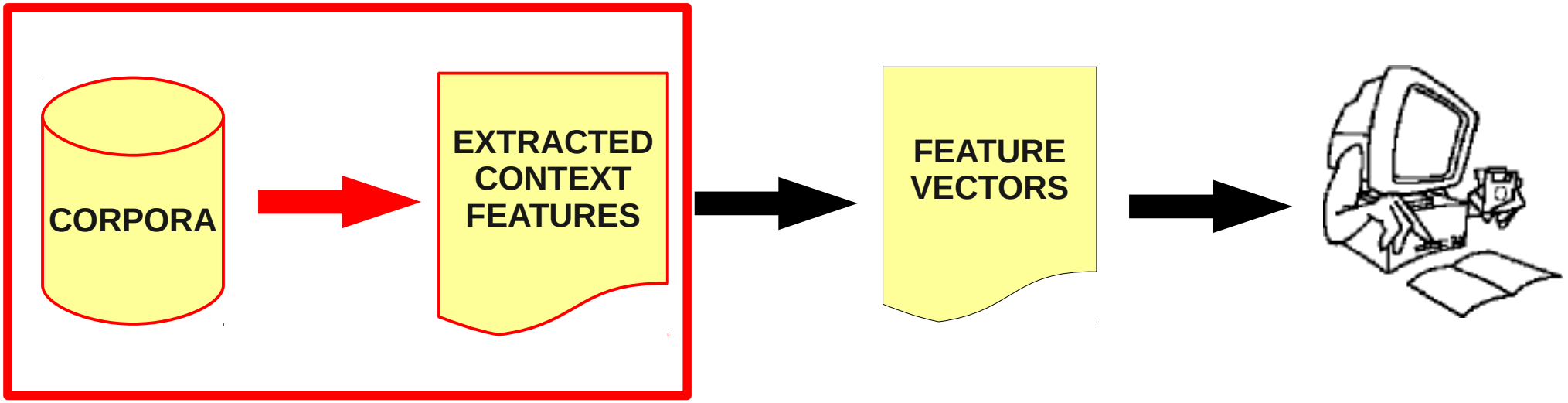
25. 1. 2012
Seminář o sémantice anglických sloves

Motivace



- potřebujeme prostředek, jak z textu získat informace (kontexty slov)
- příprava dat pro další zpracování

Motivace

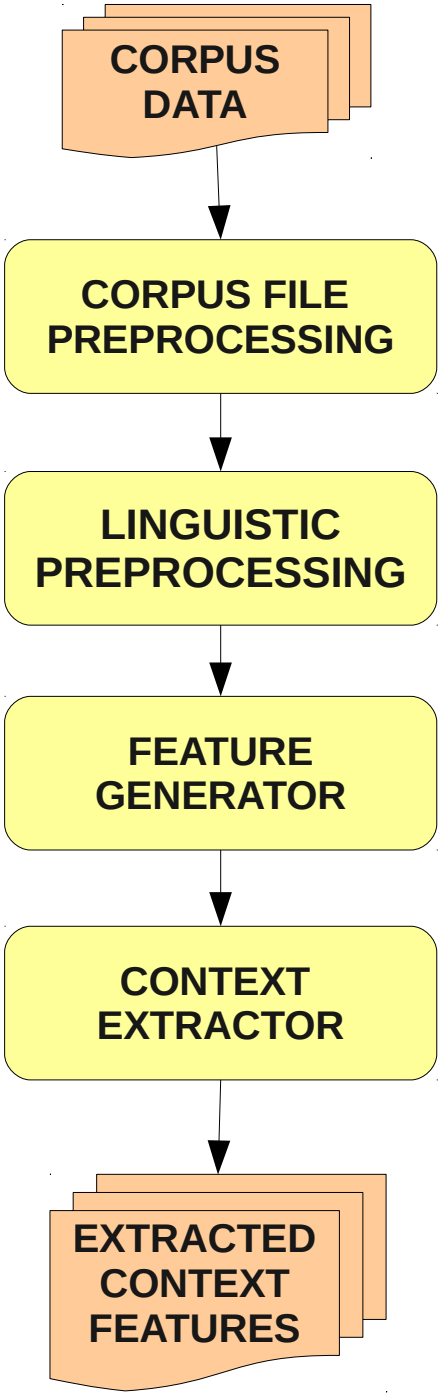


- potřebujeme prostředek, jak z textu získat informace (kontexty slov)
- příprava dat pro další zpracování

Základní pojmy

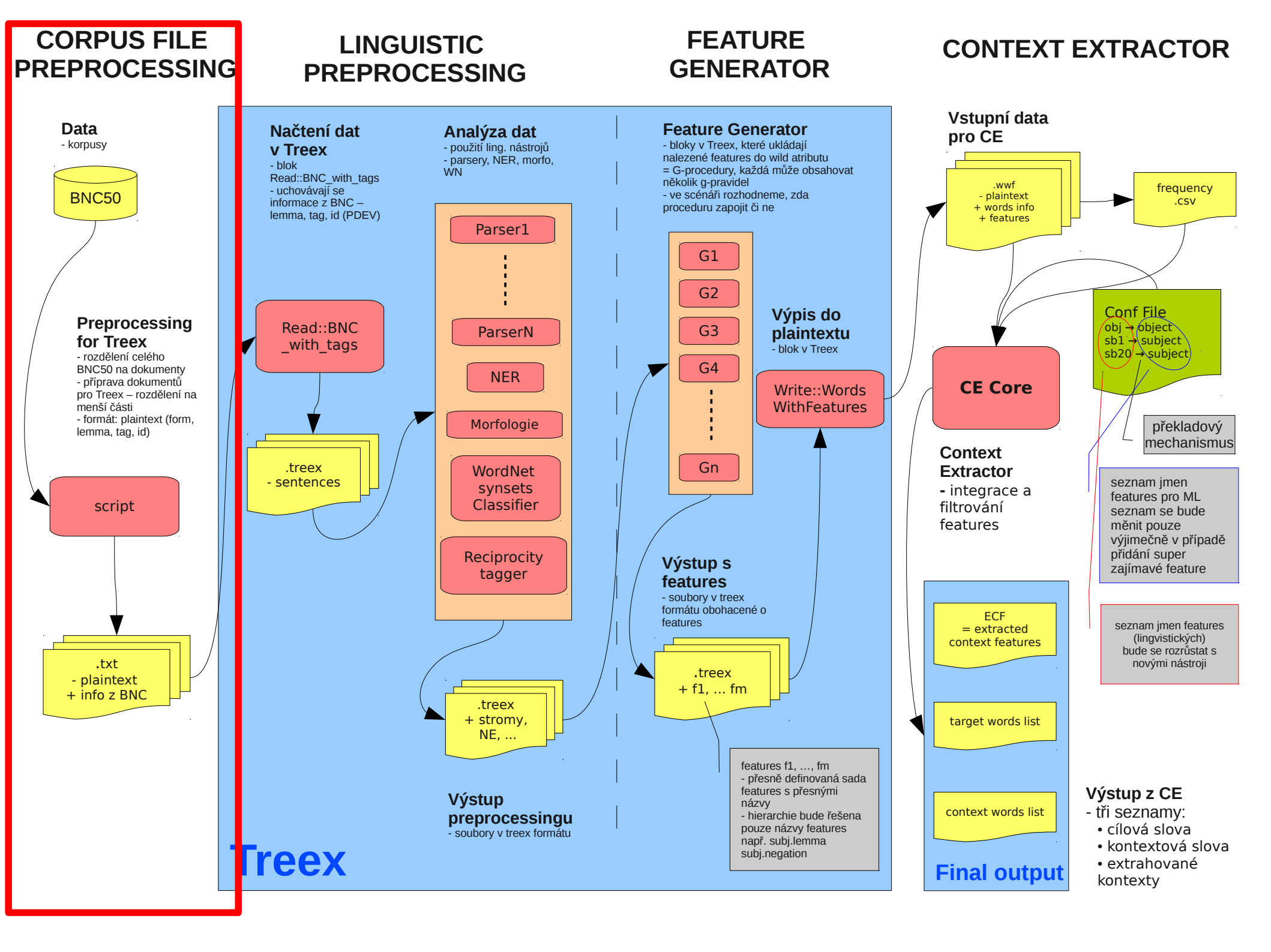
Popis systému

- Technická příprava dat
- Lingvistické předzpracování
- Generátory kontextových atributů
- Extraktor kontextů



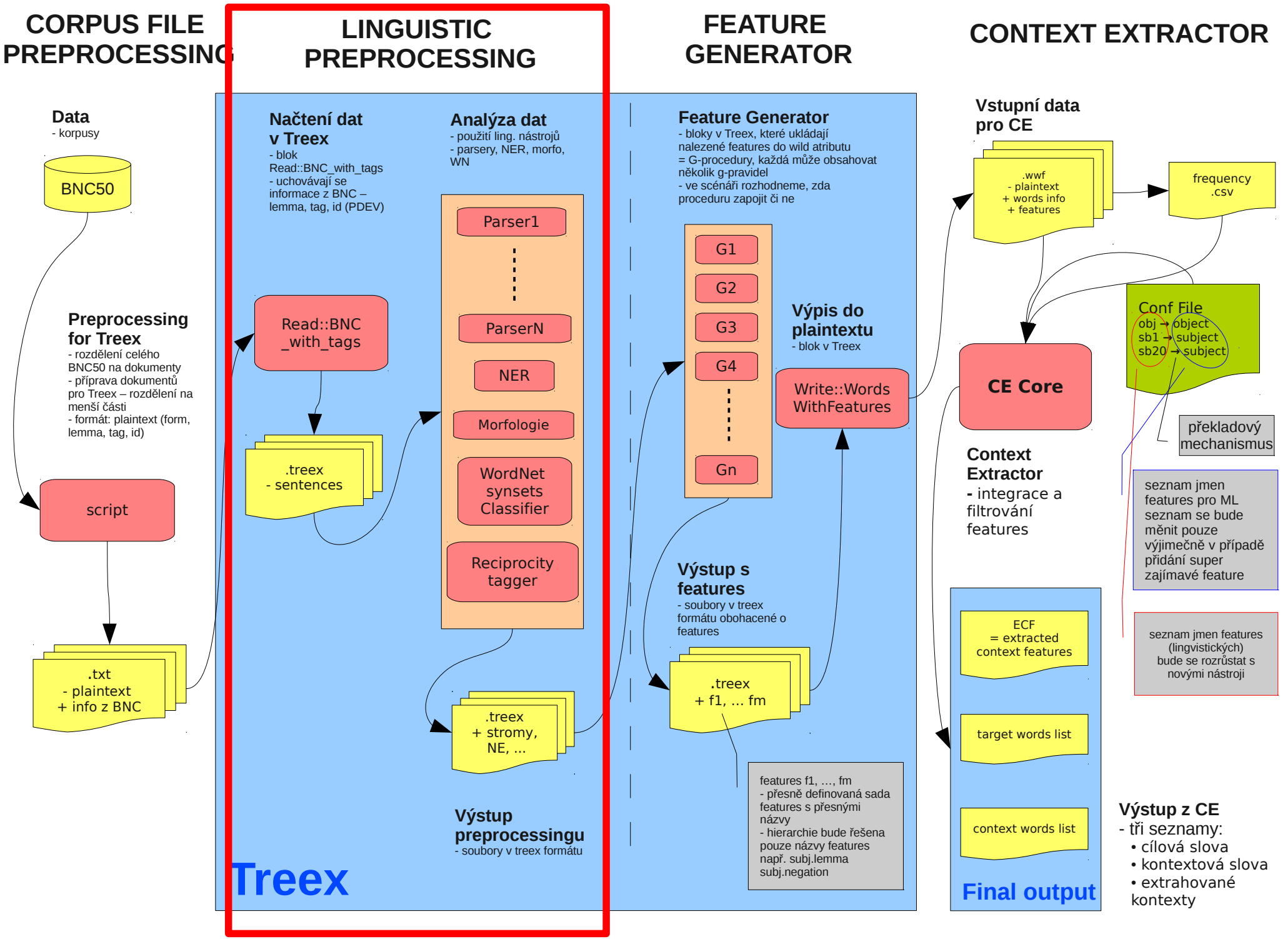
1. Technická příprava dat

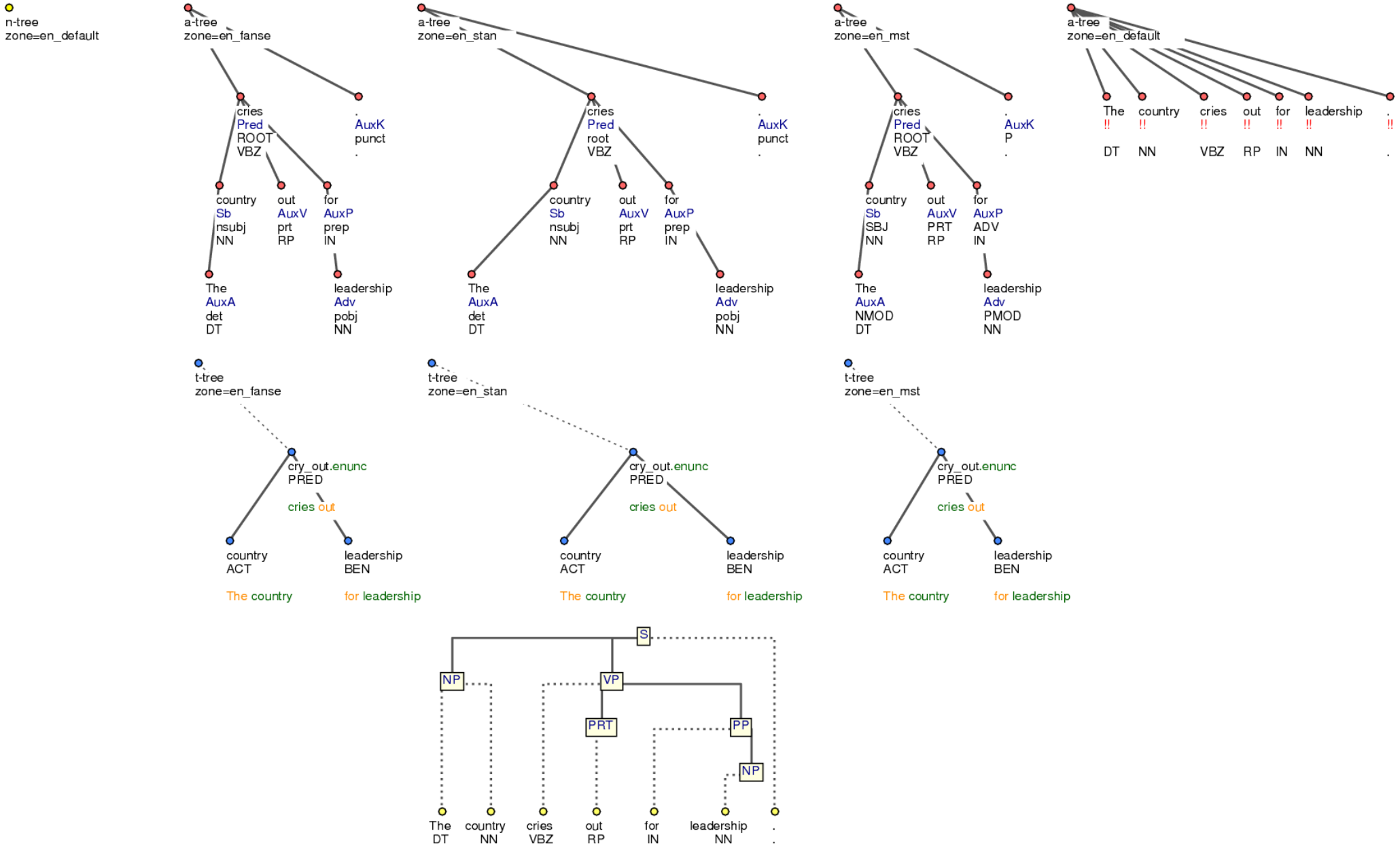
- čištění dat
- dělení na dokumenty
  - celý BNC50 byl jako jeden file
- příprava pro Treex
  - rozdělení dokumentů na soubory (po 50 větách)
  - Treex block pro načtení

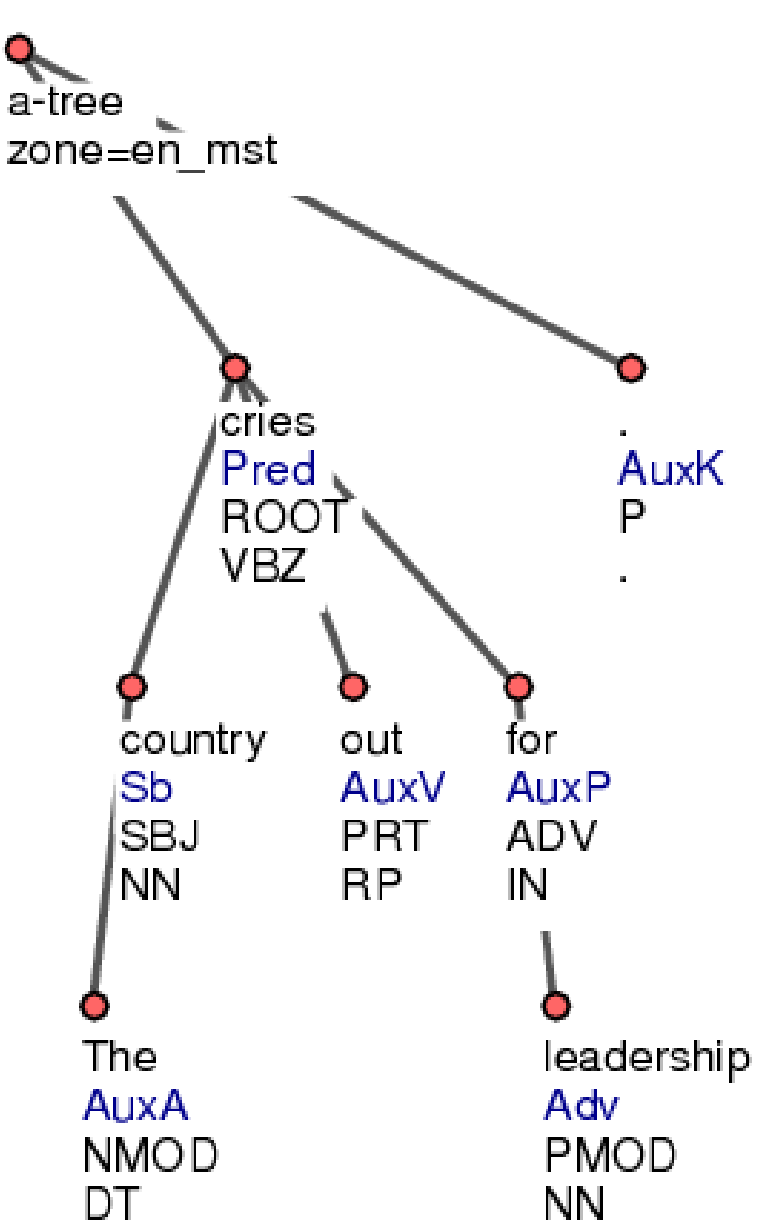

CORPUS FILE PREPROCESSING
LINGUISTIC PREPROCESSING
FEATURE GENERATOR
CONTEXT EXTRACTOR
Data
- korpusy
BNC50
Preprocessing for Treex
- rozdělení celého BNC50 na dokumenty
- příprava dokumentů pro Treex – rozdělení na menší části
- formát: plaintext (form, lemma, tag, id)
script
.txt
- plaintext
+ info z BNC
Načtení dat v Treex
- blok Read::BNC_with_tags
- uchovávají se informace z BNC – lemma, tag, id (PDEV)
Read::BNC _with_tags
.treex
- sentences
Analýza dat
- použití ling. nástrojů
- parsery, NER, morfo, WN
Parser1
ParserN
NER
Morfologie
WordNet synsets Classifier
Reciprocity tagger
.treex
+ stromy, NE, ...
Výstup preprocessingu
- soubory v treex formátu
Feature Generator
- bloky v Treex, které ukládají nalezené features do wild atributu
= G-procedury, každá může obsahovat několik g-pravidel
- ve scénáři rozhodneme, zda proceduru zapojit či ne
G1
G2
G3
G4
Gn
Výpis do plaintextu
- blok v Treex
Write::Words WithFeatures
Výstup s features
- soubory v treex formátu obohacené o features
.treex
+ f1, ... fm
features f1, ..., fm
- přesně definovaná sada features s přesnými názvy
- hierarchie bude řešena pouze názvy features např. subj.lemma subj.negation
Treex
Vstupní data pro CE
.wwf
- plaintext
+ words info
+ features
frequency .csv
Conf File
obj → object
sb1 → subject
sb20 → subject
CE Core
překladový mechanismus
Context Extractor
- integrace a filtrování features
seznam jmen features pro ML seznam se bude měnit pouze výjimečně v případě přidání super zajímavé feature
seznam jmen features (lingvistických) bude se rozrůstat s novými nástroji
ECF
= extracted context features
target words list
context words list
Final output
Výstup z CE
- tři seznamy:
• cílová slova
• kontextová slova
• extrahované kontexty

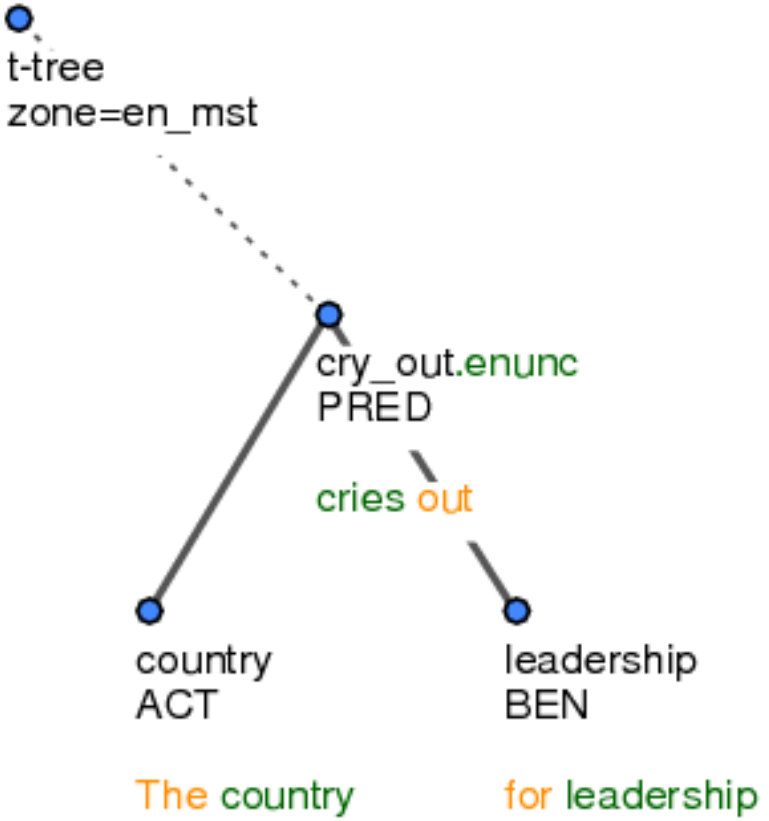
2. Lingvistické předzpracování

- lingvistická anotace
  - [segmentace]
  - [tokenizace]
  - morfologická analýza (Morče)
  - lemmatizace
  - rozpoznávání jmenných entit (Stanford NER)
  - syntaktická analýza
    - závislostní parsery: MST parser, Malt parser, Zpar, Fanse parser
    - složkové: Stanford parser, Charniak parser
    - převod: Penn Converter, Stanford Converter
  - tektogramatická analýza
  - přiřazování wordnet hyperchain


CORPUS FILE PREPROCESSING
LINGUISTIC PREPROCESSING
FEATURE GENERATOR
CONTEXT EXTRACTOR
Data
- korpusy
BNC50
Preprocessing for Treex
- rozdělení celého BNC50 na dokumenty
- příprava dokumentů pro Treex – rozdělení na menší části
- formát: plaintext (form, lemma, tag, id)
script
.txt
- plaintext
+ info z BNC
Načtení dat v Treex
- blok Read::BNC_with_tags
- uchovávají se informace z BNC – lemma, tag, id (PDEV)
Read::BNC _with_tags
.treex
- sentences
Analýza dat
- použití ling. nástrojů
- parsery, NER, morfo, WN
Parser1
ParserN
NER
Morfologie
WordNet synsets Classifier
Reciprocity tagger
.treex
+ stromy, NE, ...
Výstup preprocessingu
- soubory v treex formátu
Treex
Feature Generator
- bloky v Treex, které ukládají nalezené features do wild atributu
= G-procedury, každá může obsahovat několik g-pravidel
- ve scénáři rozhodneme, zda proceduru zapojit či ne
G1
G2
G3
G4
Gn
Výpis do plaintextu
- blok v Treex
Write::Words WithFeatures
Výstup s features
- soubory v treex formátu obohacené o features
.treex
+ f1, ... fm
features f1, ..., fm
- přesně definovaná sada features s přesnými názvy
- hierarchie bude řešena pouze názvy features např. subj.lemma subj.negation
Vstupní data pro CE
.wwf
- plaintext
+ words info
+ features
frequency .csv
Conf File
obj → object
sb1 → subject
sb20 → subject
CE Core
překladový mechanismus
Context Extractor
- integrace a filtrování features
seznam jmen features pro ML seznam se bude měnit pouze výjimečně v případě přidání super zajímavé feature
seznam jmen features (lingvistických) bude se rozrůstat s novými nástroji
ECF
= extracted context features
target words list
context words list
Final output
Výstup z CE
- tři seznamy:
• cílová slova
• kontextová slova
• extrahované kontexty

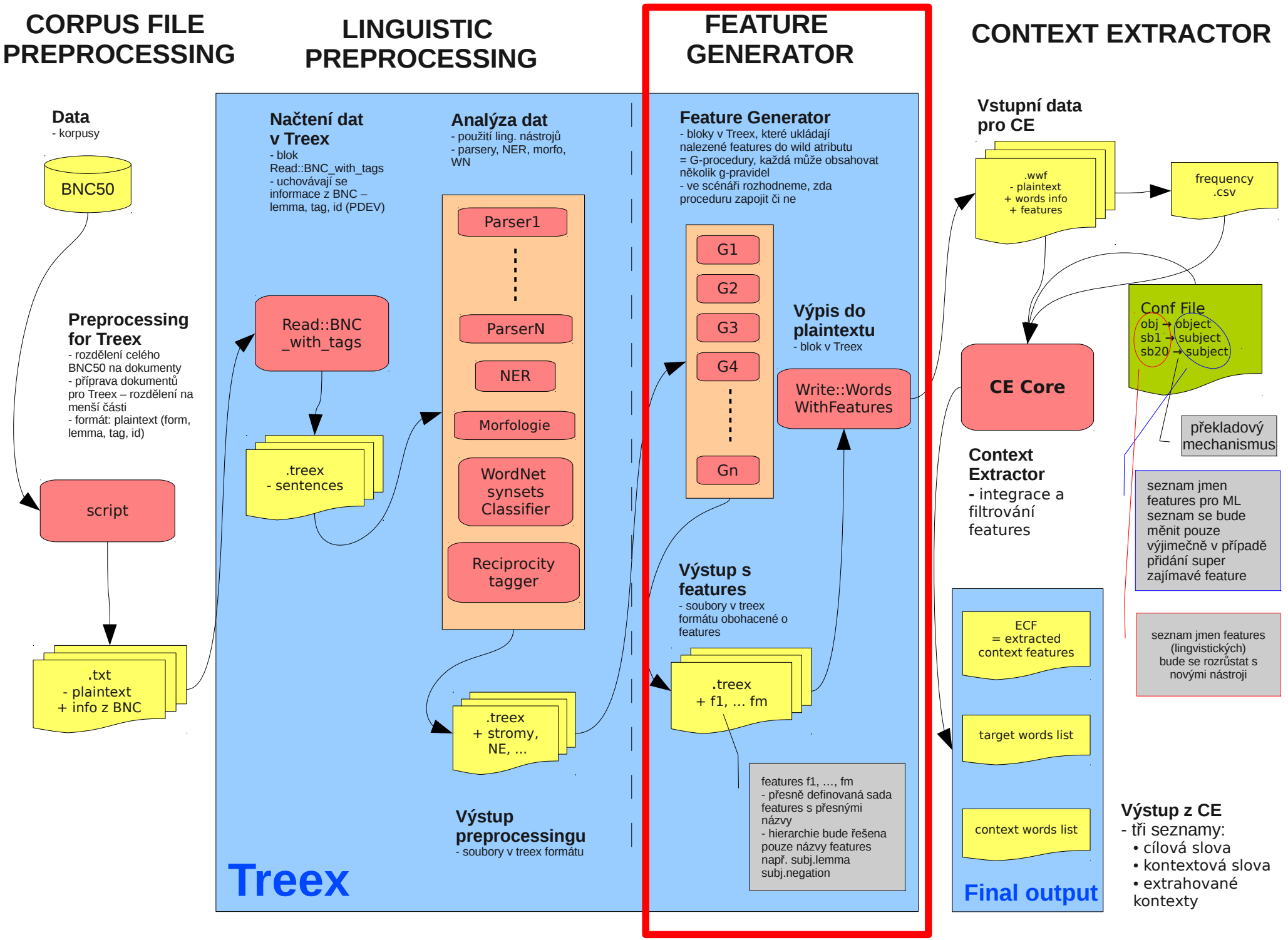

n-tree
zone=en_default
a-tree
zone=en_fanse
cries
Pred
ROOT
VBZ
.
AuxK
punct
.
country
Sb
nsubj
NN
out
AuxV
prt
RP
for
AuxP
prep
IN
The
AuxA
det
DT
leadership
Adv
pobj
NN
a-tree
zone=en_stan
cries
Pred
root
VBZ
.
AuxK
punct
.
country
Sb
nsubj
NN
out
AuxV
prt
RP
for
AuxP
prep
IN
The
AuxA
det
DT
leadership
Adv
pobj
NN
a-tree
zone=en_mst
cries
Pred
ROOT
VBZ
.
AuxK
P
.
country
Sb
SBJ
NN
out
AuxV
PRT
RP
for
AuxP
ADV
IN
The
AuxA
NMOD
DT
leadership
Adv
PMOD
NN
a-tree
zone=en_default
The
!!
DT
country
!!
NN
cries
!!
VBZ
out
!!
RP
for
!!
IN
leadership
!!
NN
.
!!
.
t-tree
zone=en_fanse
cry_out.enunc
PRED
cries out
country
ACT
The country
leadership
BEN
for leadership
t-tree
zone=en_stan
cry_out.enunc
PRED
cries out
country
ACT
The country
leadership
BEN
for leadership
t-tree
zone=en_mst
cry_out.enunc
PRED
cries out
country
ACT
The country
leadership
BEN
for leadership
S
NP
VP
PRT
PP
NP
The
DT
country
NN
cries
VBZ
out
RP
for
IN
leadership
NN
.
.


a-tree
zone=en_mst
cries
Pred
ROOT
VBZ
.
AuxK
P
.
country
Sb
SBJ
NN
out
AuxV
PRT
RP
for
AuxP
ADV
IN
The
AuxA
NMOD
DT
leadership
Adv
PMOD
NN


t-tree
zone=en_mst
cry_out.enunc
PRED
cries out
country
ACT
leadership
BEN
The country
for leadership

3. Generátor kontextových atributů

- přidání dalších informací do lingvisticky předzpracovaných dat
- g-pravidlo
  - blok v kódu, většinou jednoduchý podmíněný příkaz
  - generuje vždy právě jeden kontextový atribut
- G-procedura
  - block v Treex, logicky sdružuje více g-pravidel

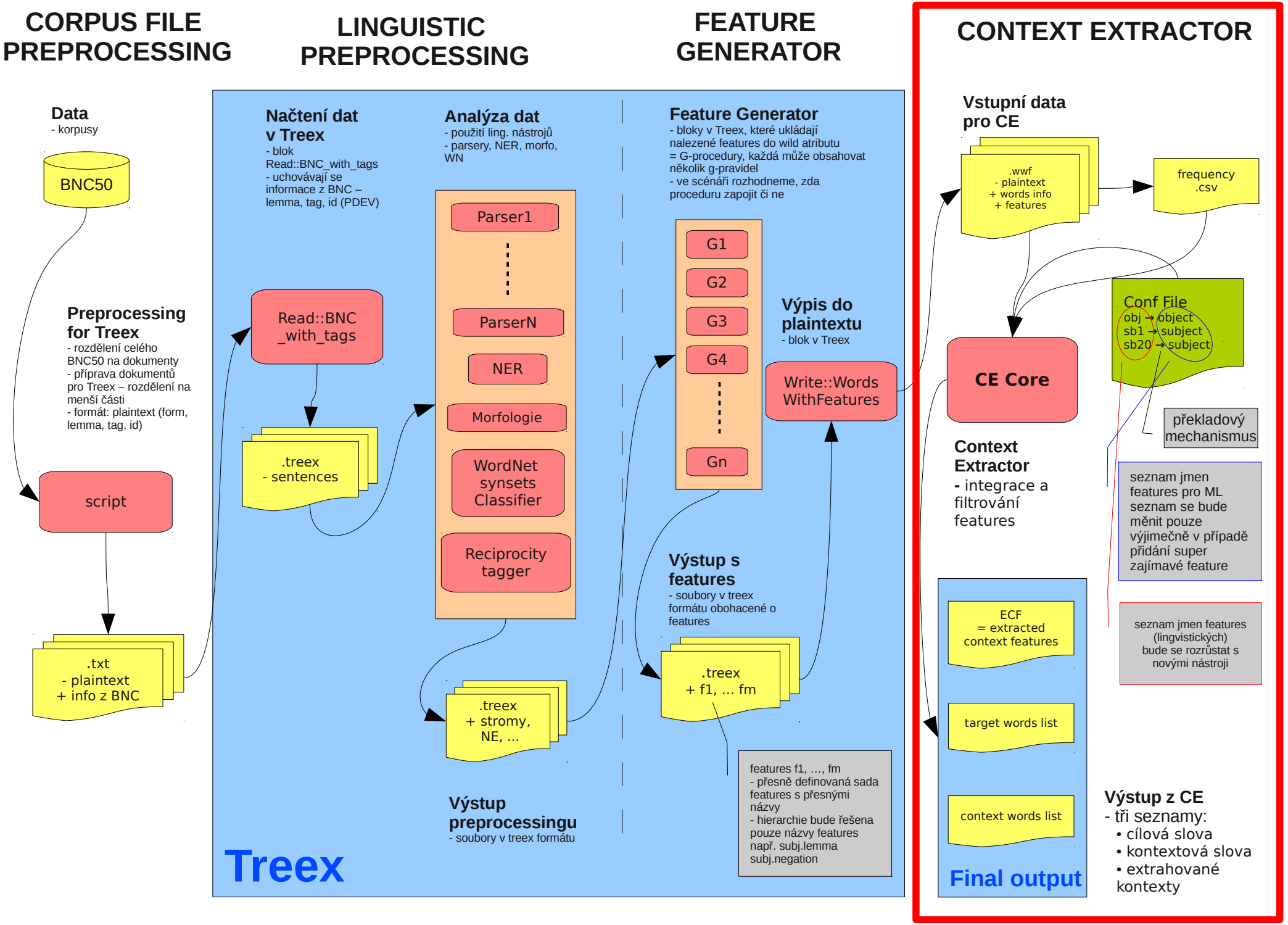

CORPUS FILE PREPROCESSING
LINGUISTIC PREPROCESSING
FEATURE GENERATOR
CONTEXT EXTRACTOR
Data
- korpusy
BNC50
Preprocessing for Treex
- rozdělení celého BNC50 na dokumenty
- příprava dokumentů pro Treex – rozdělení na menší části
- formát: plaintext (form, lemma, tag, id)
script
.txt
- plaintext
+ info z BNC
Načtení dat v Treex
- blok Read::BNC_with_tags
- uchovávají se informace z BNC – lemma, tag, id (PDEV)
Read::BNC _with_tags
.treex
- sentences
Analýza dat
- použití ling. nástrojů
- parsery, NER, morfo, WN
Parser1
ParserN
NER
Morfologie
WordNet synsets Classifier
Reciprocity tagger
.treex
+ stromy, NE, ...
Výstup preprocessingu
- soubory v treex formátu
Treex
Feature Generator
- bloky v Treex, které ukládají nalezené features do wild atributu
= G-procedury, každá může obsahovat několik g-pravidel
- ve scénáři rozhodneme, zda proceduru zapojit či ne
G1
G2
G3
G4
Gn
Výpis do plaintextu
- blok v Treex
Write::Words WithFeatures
Výstup s features
- soubory v treex formátu obohacené o features
.treex
+ f1, ... fm
features f1, ..., fm
- přesně definovaná sada features s přesnými názvy
- hierarchie bude řešena pouze názvy features např. subj.lemma subj.negation
Vstupní data pro CE
.wwf
- plaintext
+ words info
+ features
frequency .csv
Conf File
obj → object
sb1 → subject
sb20 → subject
CE Core
překladový mechanismus
Context Extractor
- integrace a filtrování features
seznam jmen features pro ML seznam se bude měnit pouze výjimečně v případě přidání super zajímavé feature
seznam jmen features (lingvistických) bude se rozrůstat s novými nástroji
ECF
= extracted context features
target words list
context words list
Final output
Výstup z CE
- tři seznamy:
• cílová slova
• kontextová slova
• extrahované kontexty

wwf - ukázka

```
XXX  0001  1   The          the        DT   a_tree-s16-n1115  AuxA features=0
XXX  0001  2   country      country    NN   a_tree-s16-n1116  Sb   features=0
XXX  0001  3   cries        cry        VBZ  a_tree-s16-n1117  Pred features=2
XXX  0001  tense   VBZ
XXX  0001  sb      a_tree-s16-n1116
XXX  0001  4   out          out        RP   a_tree-s16-n1118  AuxV features=0
XXX  0001  5   for          for        IN   a_tree-s16-n1119  AuxP features=0
XXX  0001  6   leadership   leadership NN   a_tree-s16-n1120  Adv  features=0
XXX  0001  7   .            .          .    a_tree-s16-n1121  AuxK features=0
XXX  0001
XXX  0001  1   So           so         IN   a_tree-s27-n1231  AuxC features=0
XXX  0001  2   the          the        DT   a_tree-s27-n1232  AuxA features=0
XXX  0001  3   old          old        JJ   a_tree-s27-n1233  Atr  features=0
XXX  0001  4   wine         wine       NN   a_tree-s27-n1234  Sb   features=0
XXX  0001  5   is           be         VBZ  a_tree-s27-n1235  AuxV features=1
XXX  0001  tense   VBZ
XXX  0001  6   being        be         VBG  a_tree-s27-n1236  AuxV features=1
XXX  0001  tense   VBG
XXX  0001  7   poured       pour       VBN  a_tree-s27-n1237  Pred features=2
XXX  0001  tense   VBN
XXX  0001  sb      a_tree-s27-n1234
XXX  0001  8   into         into       IN   a_tree-s27-n1238  AuxP features=0
XXX  0001  9   new          new        JJ   a_tree-s27-n1239  Atr  features=0
XXX  0001  10  bottles      bottle     NNS  a_tree-s27-n1240  Adv  features=0
XXX  0001  11  .            .          .    a_tree-s27-n1241  AuxK features=0
XXX  0001
```

4. Extraktor kontextů

- z výstupu generátorů extrahuje kontexty na základě podmínek z konfiguračního souboru
- konfigurační soubor
  - cílová a kontextová slova
  - kontextové atributy
  - další nastavení (jména souborů, kam ukládat, apod.)
- výstupem je:
  - seznam cílových slov
  - seznam kontextových slov
  - extrahované kontexty


CORPUS FILE PREPROCESSING
LINGUISTIC PREPROCESSING
FEATURE GENERATOR
CONTEXT EXTRACTOR
Data
- korpusy
BNC50
Preprocessing for Treex
- rozdělení celého BNC50 na dokumenty
- příprava dokumentů pro Treex – rozdělení na menší části
- formát: plaintext (form, lemma, tag, id)
script
.txt
- plaintext
+ info z BNC
Načtení dat v Treex
- blok Read::BNC_with_tags
- uchovávají se informace z BNC – lemma, tag, id (PDEV)
Read::BNC _with_tags
.treex
- sentences
Analýza dat
- použití ling. nástrojů
- parsery, NER, morfo, WN
Parser1
ParserN
NER
Morfologie
WordNet synsets Classifier
Reciprocity tagger
.treex
+ stromy, NE, ...
Výstup preprocessingu
- soubory v treex formátu
Feature Generator
- bloky v Treex, které ukládají nalezené features do wild atributu
= G-procedury, každá může obsahovat několik g-pravidel
- ve scénáři rozhodneme, zda proceduru zapojit či ne
G1
G2
G3
G4
Gn
Výpis do plaintextu
- blok v Treex
Write::Words WithFeatures
Výstup s features
- soubory v treex formátu obohacené o features
.treex
+ f1, ... fm
features f1, ..., fm
- přesně definovaná sada features s přesnými názvy
- hierarchie bude řešena pouze názvy features např. subj.lemma subj.negation
Treex
Vstupní data pro CE
.wwf
- plaintext
+ words info
+ features
frequency .csv
Conf File
obj → object
sb1 → subject
sb20 → subject
CE Core
překladový mechanismus
Context Extractor
- integrace a filtrování features
seznam jmen features pro ML seznam se bude měnit pouze výjimečně v případě přidání super zajímavé feature
seznam jmen features (lingvistických) bude se rozrůstat s novými nástroji
ECF
= extracted context features
target words list
context words list
Final output
Výstup z CE
- tři seznamy:
• cílová slova
• kontextová slova
• extrahované kontexty

Konfigurační soubor

```
# ========== tword ==========
tword     YES    pos="verb"    fmin=1000
tword     NO     lemma="be have"

# ========== cword ==========
cword     YES    pos="noun"    fmin=100

# ========== context ==========
context   distance="10"
context   features="sb->subject \
                    obj->object\
                    prt->particle"
context   features="adv->adverbial"
```

target/context words list

lemma	t_id	f_s	f_all
enlist-V	0	1	348
claim-V	1	2	12031
plug-V	4	4	318
wake-V	8	3	926
cry-V	11	6	1198
tell-V	19	2	21594
arrive-V	25	5	6056
enlarge-V	28	2	499
pour-V	31	5	1076

lemma	c_id	f_s	f_all
government-N	22	1	43580
option-N	23	1	3758
tomorrow-N	24	1	1575
dream-N	25	1	2097
anyone-N	26	1	5325
decibel-N	27	1	47
head-N	28	2	12359
country-N	35	1	31874
leadership-N	36	1	3990
wine-N	60	1	1283
bottle-N	61	1	1464
soul-N	62	1	1535

Extracted context features

cry-V	**11**	**dXXX-s16-w3**	**4**	
	subject	country-N	35	dXXX-s16-w2
	l1-s0	country-N	35	dXXX-s16-w2
	l9-s-1	wound-N	34	dXXX-s15-w16
	r3-s0	leadership-N	36	dXXX-s16-w6
tell-V	**19**	**dXXX-s17-w2**	**3**	
	l3-s-1	leadership-N	36	dXXX-s16-w6
	l7-s-1	country-N	35	dXXX-s16-w2
	r6-s0	fuss-N	37	dXXX-s17-w8
...				
pour-V	**31**	**dXXX-s26-w4**	**5**	
	subject	cash-N	58	dXXX-s26-w2
	l2-s0	cash-N	58	dXXX-s26-w2
	l5-s-1	story-N	57	dXXX-s25-w9
	r2-s0	marketing-N	59	dXXX-s26-w6
	r7-s1	wine-N	60	dXXX-s27-w4
pour-V	**31**	**dXXX-s27-w7**	**4**	
	subject	wine-N	60	dXXX-s27-w4
	l3-s0	wine-N	60	dXXX-s27-w4
	l8-s-1	marketing-N	59	dXXX-s26-w6
	r3-s0	bottle-N	61	dXXX-s27-w10

Shrnutí

- systém pro extrakci kontextů
  - relativně univerzální
    - nezávislý na jazyku
    - nezávislý na korpusu
  - konfigurovatelný
  - může posloužit i v jiných projektech