```
а
ода
а
одзе
д
а

ыў
е

buntu@ubuntu-VirtualBox:/media/sf_hse/compling/apertium/apertium-bel$ cut -f2 -d' ' be_freq.txt | apertium
-d. bel-morph | head -15
і/і<cnjcoo>$
у/у<pr>$
ў/у<pr>$
з/з<pr>$

на/на<pr>$ ^да/да<pr>$
года/год<n><m><nn><sg><gen>$
па/па<pr>$
годзе/годзе<adv>/год<n><m><nn><sg><loc>$
ад/ад<pr>$
за/за<pr>$
а/а<cnjcoo>$
быў/быць<vblex><impf><past><m><sg>$
не/не<ij>$
г/*г$
buntu@ubuntu-VirtualBox:/media/sf_hse/compling/apertium/apertium-bel$ cut -f2 -d' ' be_freq.txt | apertium
-d . bel-morph | grep -v -P "^\s*$" | head -15
і/і<cnjcoo>$
у/у<pr>$
ў/у<pr>$
з/з<pr>$
на/на<pr>$ ^да/да<pr>$
года/год<n><m><nn><sg><gen>$
па/па<pr>$
```

Как выбрать нужные метки?

- Слова, которые apertium не анализирует (*)
- Много анализов

Как данные выглядят:

3326181 ^v/v<pr>$

3233979 ^a/a<cnjcoo>$

1919556 ^se/se<prn><ref><acc>$

1744502 ^na/na<pr>$

1296633 ^je/být<vbser><pres><p3><sg>/prpers<prn><p3><mf><pl><acc>$

33460 ^b/*b$

48730 ^často/často<adv><sint>/častý<adj><sint><cpd>/častý<adj><sint><cpd>$

Скрипт разделяет слова и анализы благодаря знакам препинания:

566709 ^roce/rok<n><mi><sg><loc>$
483610
^roku/rok<n><mi><sg><loc>/rok<n><mi><sg><voc>/rok<n><mi><sg><dat>/rok<n><mi><sg><gen>$
169189 ^mezi/mez<n><f><sg><loc>/mez<n><f><sg><voc>/mez<n><f><sg><dat>$
165921 ^letech/léto<n><nt><pl><loc>$
48730 ^často/častý<adj><sint><cpd>/častý<adj><sint><cpd>$

Переводим метки в формат UD.

```
roce    rok NOUN  Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
roku    rok NOUN  Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing
roku    rok NOUN  Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing
roku    rok NOUN  Animacy=Inan|Case=Dat|Gender=Masc|Number=Sing
roku    rok NOUN  Animacy=Inan|Case=Voc|Gender=Masc|Number=Sing
často   častý   ADJ
Gender=Neut|Number=Sing|Polarity=Pos|Variant=Short|VerbForm=Part|Voice=Pas
s
```

```
mkdir crh
wget
https://dumps.wikimedia.org/crhwiki/20190120/crhwiki-20190120-pages-articles.x
ml.bz2
$ python3 ~/scripts/WikiExtractor.py --infn
crhwiki-20190120-pages-articles.xml.bz2 >/dev/null
$ cat wiki.txt | sed 's/[^a-zA-Z]\+/\n/g' | sort -u > crh-tokens.txt

cd ..
mkdir tur
wget
https://dumps.wikimedia.org/trwiki/latest/trwiki-latest-pages-articles.xml.bz2
$ python3 ~/scripts//WikiExtractor.py --infn
trwiki-latest-pages-articles.xml.bz2 >/dev/null
$ cat wiki.txt | sed 's/[^a-zA-Z]\+/\n/g' | sort -u > tur-tokens.txt

cd ..

# Discover cognates in the simplest way possible!
for i in `cat crh/crh-tokens.txt`; do cat tur/tur-tokens.txt | grep ^$i$; done
> crh-tur-intersection.txt

$ wget
"https://github.com/ftyers/calma/blob/master/sharedtaskdata/train/tur-uncovere
d?raw=true" -O tur-uncovered

$ for i in `cat crh-tur-intersection.txt`; do cat tur-uncovered | grep -P
"^tur\t$i\t"; done

# Create synthetic training data!
$ for i in `cat ../../crimean-tatar/wikipedia/crh-tur-intersection.txt`; do
cat tur-uncovered | grep -P "^tur\t$i\t"; done | sed 's/^tur/crh/g'
```

Romance languages (`roa`)

Code	Language
ast	Asturian
cat	Catalan
fra	French
ita	Italian
por	Portuguese
spa	Spanish
???	Surprise Language

Turkic languages (`trk`)

Code	Language
bak	Bashkir
crh	Crimean Tatar
kaz	Kazakh
kir	Kyrgyz
tat	Tatar
tur	Turkish
???	Surprise Language

О попытках

- бейзлайн:
  - seq2seq
  - энкодер-декодер на lstm
  - + attention
- наивные улучшения:
  - отдельно предсказывать POS, отдельно остальное
  - обучить эмбеддинги для всех языков на вики
  - перевести эмбеддинги для всех языков в одно пространство
  - найти для слов из тестовой выборки когнаты в обучающей выборке
    - с этим не очень очевидно, что делать, но очень хочется
  - **???**

```
siam.head()
```

	language	iso_code	gloss	global_id	local_id	transcription	cognate_class	notes
0	ANCIENT_GREEK	grc	sharp	1396	sharp	oksis	sharp:C	NaN
1	GREEK	ell	sharp	1396	sharp	kofteros	sharp:J	NaN
2	CLASSICAL_ARMENIAN	xcl	sharp	1396	sharp	sowr	sharp:D	NaN
3	ARMENIAN_EASTERN	hye	sharp	1396	sharp	sur	sharp:D	NaN
4	OSSETIC	oss	sharp	1396	sharp	ts3rG	sharp:A	NaN

```
roa.head()
```

	iso_code	wordform	lemma	POS	analysis
0	cat	revelat	revelar	VERB	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
1	por	optou	optar	VERB	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tens...
2	cat	concedit	concedir	VERB	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
3	cat	erigida	erigir	VERB	Gender=Fem|Number=Sing|Tense=Past|VerbForm=Part
4	fra	affirmé	affirmer	VERB	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part