Feldman & Hana 2010

NPFL096 Computational Morphology 2011

Jirka Hana

Based on slides for an ESSLLI 2010 course
by Anna Feldman & Jirka Hana

Cohen's kappa (Cohen 1960)

- The most popular measure of agreement between two annotators.
- Takes into account (somewhat) the possibility of chance agreement.
- $\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$

  $Pr(a)$ - the relative observed agreement
  $Pr(e)$ - the hypothetical probability of chance agreement
  $Pr(e) = \sum_t \frac{t_a * t_b}{N}$
  t_a – number of tags t assigned by annotator a
  N – number of all tags
- Weighted kappa – gives different weights to different errors .

- (Variant of) Kendall's tau - the minimal number of operations necessary to turn one annotation into the other.
- There are other measures.
- High agreement is important but it is not everything:
  - One can use use a tagset with a single tag.
  - The annotation manual can be purely formal (Tag all sentence initial words as topics).
  - On the other hand, if iaa is below the accuracy of a tagger ...

A Resource Light MA (of Czech)

- Motivation
- Guesser
- Lexicon Acquisition
- Results

Reminder: What is MA?

MA: form $\rightarrow$ set(lemma $\times$ set(tag))

English:	*her*	$\rightarrow$	{ (*she*,	{PP }),
			(*her* ,	{PP$}) }

Czech:	*ženou*	$\rightarrow$	{ (*žena* 'woman',	{noun fem sing inst }),
			(*hnát* 'hurry',	{verb pres pl 3rd }) }
	ženy	$\rightarrow$	{ (*žena* 'woman',	{noun fem sing gen,
				noun fem pl nom,
				noun fem pl acc,
				noun fem pl voc }) }

Focus on nouns

We focus exclusively on nouns.

- Hard & interesting
  - High homonymy
  - The most open class (Names!)
- We cannot do everything at once

Two extreme approaches to MA

- Provide all information manually – e.g. (Hajic 2004)
  - \+ High accuracy (Recall 98.5%)
  - − Very costly (300K lexicon)
- Learn all information automatically – e.g. (Goldsmith 2001)
  - \+ Cheap to use, good for understudied languages
  - − Low accuracy

Corpus coverage by lemma frequency

tr1 corpus				tr2 corpus
Lemma freq decile	Number of tokens	Corpus noun coverage (%)	Cumulative coverage (%)	Lemmas not present (%)
10	164 643	74	74	0.2
9	22 515	10	84	6.7
8	11 041	5.0	89	22
7	6 741	3.0	92	36
6	4 728	2.1	94	48
5	3 179	1.4	96	61
4	2 365	1.0	97	65
3	2 364	1.0	98	70
2	2 364	1.0	99	75
1	2 364	1.0	100	77

tr1/tr2: each 700K tokens; newspapers, magazine; similar
Each decile contains 2364 or 2365 noun lemmas.

What does it mean? – The good news

tr1 corpus				tr2 corpus
Lemma freq decile	Number of tokens	Corpus noun coverage (%)	Cumulative coverage (%)	Lemmas not present (%)
10	164 643	74	74	0.2
9	22 515	10	84	6.7
8	11 041	5.0	89	22
7	6 741	3.0	92	36
6	4 728	2.1	94	48
5	3 179	1.4	96	61
4	2 365	1.0	97	65
3	2 364	1.0	98	70
2	2 364	1.0	99	75
1	2 364	1.0	100	77

What does it mean? – The good news

tr1 corpus				tr2 corpus
Lemma freq decile	Number of tokens	Corpus noun coverage (%)	Cumulative coverage (%)	Lemmas not present (%)
10	164 643	74	74	0.2
9	22 515	10	84	6.7
8	11 041	5.0	89	22
7	6 741	3.0	92	36
6	4 728	2.1	94	48
5	3 179	1.4	96	61
4	2 365	1.0	97	65
3	2 364	1.0	98	70
2	2 364	1.0	99	75
1	2 364	1.0	100	77

Complete Goldsmith is not necessary

- 2.5K most frequent lemmas cover 3/4 of tokens
- 7K most frequent lemmas cover nearly 90% of tokens

What does it mean? – The bad news

tr1 corpus				tr2 corpus
Lemma freq decile	Number of tokens	Corpus noun coverage (%)	Cumulative coverage (%)	Lemmas not present (%)
10	164 643	74	74	0.2
9	22 515	10	84	6.7
8	11 041	5.0	89	22
7	6 741	3.0	92	36
6	4 728	2.1	94	48
5	3 179	1.4	96	61
4	2 365	1.0	97	65
3	2 364	1.0	98	70
2	2 364	1.0	99	75
1	2 364	1.0	100	77

What does it mean? – The bad news

tr1 corpus				tr2 corpus
Lemma freq decile	Number of tokens	Corpus noun coverage (%)	Cumulative coverage (%)	Lemmas not present (%)
10	164 643	74	74	0.2
9	22 515	10	84	6.7
8	11 041	5.0	89	22
7	6 741	3.0	92	36
6	4 728	2.1	94	48
5	3 179	1.4	96	61
4	2 365	1.0	97	65
3	2 364	1.0	98	70
2	2 364	1.0	99	75
1	2 364	1.0	100	77

Complete Hajič is impossible, nearly complete is hard

- Coverage gains drop quickly – each of the 5 lower deciles adds ca 1%
- Infrequent lemmas are text specific – 70% (!!) of the less frequent half of the lemmas from `tr1` do not occur in `tr2`

Guesser

- Looks at endings (sometimes also at the ends of stems)
- Uses manually supplied info about Czech noun paradigms:
    - endings + tags
    - permissible stem-tails
    - some stem alternation (regular tail changes, epenthesis)
    - 13 linguistic paradigms are encoded as 64 paradigms.
    - a book for general public used as a reference (Karlík et al. 1996)
- Massively overgenerates – good recall, bad precision

Czech noun paradigms

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	kr-a	pár-a	mouch-a
S2	žen-y	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-**e**	koz-**e**	k**ř**-**e**	pá**ř**-e	mou**š**-**e**
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-**e**	koz-**e**	k**ř**-**e**	pá**ř**-e	mou**š**-**e**
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-y	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	k**e**r-0	p**a**r-0	m**u**ch-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-y	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P5	žen-y	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

Czech noun paradigms – Ending variation

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	kr-a	pár-a	mouch-a
S2	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	ker-0	par-0	much-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P5	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

- Ending variation: *žen-ě*, *sov-ě* vs. *burz-e*, *kř-e*, *pář-e*
  The dative and local sg. ending is -*ě* after alveolar stops (*d*, *t*, *n*) and labials (*b*, *p*, *m*, *v*, *f*). It is -*e* otherwise.

Czech noun paradigms – Ending variation

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	kr-a	pár-a	mouch-a
S2	žen-**y**	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-**e**
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-**y**	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	ker-0	par-0	much-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-**y**	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P5	žen-**y**	sov-y	skic-**i**	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

- Ending variation: *žen-y* vs. *skic-i*.
  Czech spelling rules require the ending *-y* to be spelled as *-i* after certain consonants, in this case: *c, č, ď, ň, š*. The pronunciation is the same ([ɪ]).

Czech noun paradigms – Stem change

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	kr-a	pár-a	mouch-a
S2	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	ker-0	par-0	much-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P5	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

- Palatalization of the stem final consonant:
  kr-a – kř-e, *mouch-a – mouš-e*.
  The -*ě*/*e* ending affects the preceding consonant: *ch* [x] → *š*, *g*/*h* → *z*, *k* → *c*, *r* → *ř*.

Czech noun paradigms – Stem change

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	**kr**-a	pár-a	mouch-a
S2	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	k**e**r-0	par-0	much-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P5	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

- Epenthesis: *kr-a* – *ker*.
  Sometimes, there is an epenthesis (insertion of -*e*-) in genitive plural.

Czech noun paradigms – Stem change

Table: Examples of the *žena* 'woman' paradigm nouns

	woman	owl	draft	goat	iceberg	vapor	fly
S1	žen-a	sov-a	skic-a	koz-a	kr-a	pár-a	mouch-a
S2	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
S3	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S4	žen-u	sov-u	skic-u	koz-u	kr-u	pár-u	mouch-u
S5	žen-o	sov-o	skic-o	koz-o	kr-o	pár-o	mouch-o
S6	žen-ě	sov-ě	skic-e	koz-e	kř-e	pář-e	mouš-e
S7	žen-ou	sov-ou	skic-ou	koz-ou	kr-ou	pár-ou	mouch-ou
P1	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P2	žen-0	sov-0	skic-0	koz-0	ker-0	par-0	much-0
P3	žen-ám	sov-ám	skic-ám	koz-ám	kr-ám	pár-ám	mouch-ám
P4	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P5	žen-y	sov-y	skic-i	koz-y	kr-y	pár-y	mouch-y
P6	žen-ách	sov-ách	skic-ách	koz-ách	kr-ách	pár-ách	mouch-ách
P7	žen-ami	sov-ami	skic-ami	koz-ami	kr-ami	pár-ami	mouch-ami

- Stem internal vowel shortening: *pár-a* – *par*.

Czech noun paradigms (cont.)

- Roughly 13 basic noun paradigms:
  - 4 neuter
  - 3 feminine
  - 6 masculine
  - 2 paradigms for nouns with adjectival declension
- Many subparadigms and subsubparadigms, great amount of irregularity, variation, and homonymy
- Some forms have official and colloquial variants

Encoding Czech noun paradigms

Ending Homony

Table: Homonymy of the *a* ending in Czech

form	lemma	gloss		category
měst-a	město	town	NS2	noun neut sg gen
			NP1 (5)	noun neut pl nom (voc)
			NP4	noun neut pl acc
tém-a	téma	theme	NS1 (5)	noun neut sg nom (voc)
			NS4	noun neut sg acc
žen-a	žena	woman	FS1	noun fem sg nom
pán-a	pán	man	MS2	noun masc anim sg gen
			MS4	noun masc anim sg acc
ostrov-a	ostrov	island	IS2	noun masc inanim sg gen
předsed-a	předseda	president	MS1	noun masc anim sg nom
vidě-l-a	vidět	see		verb past fem sg
				verb past neut pl
vidě-n-a				verb passive fem sg
				verb passive neut pl
vid-a				verb transgressive masc sg
dv-a	dv-a	two		numeral masc sg nom
				numeral masc sg acc

Ending Homony (cont.)

Table: Ending *-e* and noun cases in Czech

case	form	lemma	gender	gloss
nom	kuř-e	kuře	neuter	chicken
gen	muž-e	muž	masc.anim.	man
dat	mouš-e	moucha	feminine	fly
acc	muž-e	muž	masc.anim.	man
voc	pan-e	pán	masc.anim.	mister
loc	mouš-e	moucha	feminine	fly
inst	–	–		

Lexicon Acquisition

Guesser overgenerates. Use a raw corpus to prune the results.

Lexicon Acquisition

Guesser overgenerates. Use a raw corpus to prune the results.

Lemma of *talking* ?

Lexicon Acquisition

Guesser overgenerates. Use a raw corpus to prune the results.

Lemma of *talking* ?

- *talk*?
- *talking* (à la *sibling*)?

Lexicon Acquisition

Guesser overgenerates. Use a raw corpus to prune the results.

Lemma of *talking* ?

- *talk*?
- *talking* (à la *sibling*)?

Also found *talk*, *talks*, *talked* – clear

Did you see *sible*, *sibles*, *sibled*?

An Example & A Problem

forms	tokens
atom-0	48
atom-u	28
atom-em	1
atom-y	22
atom-ů	30
atom-ům	1
atom-ech	1

	inanim	found
S1	hrad-0	+
S2	hrad-ě/u	−/+
S3	hrad-u	+
S4	hrad-0	+
S5	hrad-e	
S6	hrad-ě/u	−/+
S7	hrad-em	+
P1	hrad-y	+
P2	hrad-ů	+
P3	hrad-ům	+
P4	hrad-y	+
P5	hrad-y	+
P6	hrad-ech	+
P7	hrad-y	+
Total		7

An Example & A Problem'

forms	tokens
atom-0	48
atom-u	28
atom-em	1
atom-y	22
atom-ů	30
atom-ům	1
atom-ech	1

	inanim	found	anim	found
S1	hrad-0	+	pán-0	+
S2	hrad-ě/u	−/+	pán-a	
S3	hrad-u	+	pán-u/ovi	+/−
S4	hrad-0	+	pán-a	
S5	hrad-e		pan-e	
S6	hrad-ě/u	−/+	pán-u	+
S7	hrad-em	+	pán-em	+
P1	hrad-y	+	pán-i/ové	−
P2	hrad-ů	+	pán-ů	+
P3	hrad-ům	+	pán-ům	+
P4	hrad-y	+	pán-y	+
P5	hrad-y	+	pán-i	
P6	hrad-ech	+	pán-ech	+
P7	hrad-y	+	pán-y	+
Total		7		7

An Example & A Problem”

forms	tokens
atom-0	48
atom-u	28
atom-em	1
atom-y	22
atom-ů	30
atom-ům	1
atom-ech	1
atom-ové	200

	inanim	found	anim	found
S1	hrad-0	+	pán-0	+
S2	hrad-ě/u	−/+	pán-a	
S3	hrad-u	+	pán-u/ovi	+/−
S4	hrad-0	+	pán-a	
S5	hrad-e		pan-e	
S6	hrad-ě/u	−/+	pán-u	+
S7	hrad-em	+	pán-em	+
P1	hrad-y	+	**pán-i/ové**	**−/+**
P2	hrad-ů	+	pán-ů	+
P3	hrad-ům	+	pán-ům	+
P4	hrad-y	+	pán-y	+
P5	hrad-y	+	pán-i	
P6	hrad-ech	+	pán-ech	+
P7	hrad-y	+	pán-y	+
Total		7		8

An Example & A Problem"

We can connect inflectional paradigms related by derivation into "super-paradigms".

Alleviates two important problems:

- The *ové* problem above *ové* = *ov-é*
- Data sparsity.

Very rough (overgenerating) information seems to be enough.

Algorithm

1. MA of a corpus & Create all possible hypothetical lexical entries
2. Cluster entries & Filter out the bad ones.
   Simply put: the entry that covers the highest number of forms wins.
   - Size of the wining crust can be specified. In relative or absolute terms.
   - Minimal number of tokens for an entry can be specified.
   - Exclude strange entries – contains infrequent forms (voc), but not frequent (nom)
   - Etc.

Limited memory: several passes, etc.

MA modules

Running a cascade of modules. High precision first, high recall last.

- Word list
- Abbreviation identification
- Numbers
- Lexicon based analyzer
- Paradigm-based guesser

Results (on nouns)

Lexicon	–	–	–	+	+	+	+	Hajič[1]
Top forms list	0K	5K	10K	0K	5K	10K	10K	
Derivation suff:	0	0	0	0	0	0	20	
Error rate	3.6	2.9	2.7	5.8	3.9	**3.6**	**3.4**	1.3
Ambiguity tag/w	19.6	13.1	11.5	11.7	8.5	**7.8**	**4.0**	3.8

Results for other POS than noun are better.

[1] (Hajic 2004, p.c.): 300K lexicon

Evaluation of the Russian morphological analyzer

Lexicon		no	yes	no	yes
LEO		no	no	yes	yes
All	Recall error:	2.9	4.3	12.7	6.6
	ambiguity (tag/w)	9.7	4.4	3.3	2.8
N	Recall error:	2.6	4.9	41.6	13.7
	ambiguity (tag/w)	18.6	6.8	6.5	4.3
A	Recall error:	6.2	7.0	8.1	7.5
	ambiguity (tag/w)	21.6	10.8	3.3	5.7
V	Recall error:	0.8	2.0	2.3	2.3
	ambiguity (tag/w)	14.7	4.8	1.5	1.5

No Top-frequency lists, no derivation used.

Resource light morphology – Why?

- Traditional taggers and analyzers are very accurate, but very costly (money, time, resources)
- Most languages and dialects have no realistic prospect for morphological tools created in this way

Main Assumption

- target-language model can be approximated by language models of related source language(s)
- inclusion of a limited amount of high-impact and/or low-cost manual resources is greatly beneficial and desirable

Using TnT (Brants 2000), a second order Markov Model tagger

- emissions: approximated by the source-language emissions + resource-light morphological analysis
- transitions: approximated by the source-language transitions

See (Feldman and Hana 2010)

Languages

- We have experimented with several language pairs
  - Russian via Czech
  - Catalan via Spanish
  - Portuguese via Spanish
- Currently working on
  - Lithuanian via Russian/Czech
  - Romanian via Bulgarian/Spanish
- Planning to do Old Czech.

Here, we present our approach on Czech and Russian.

Russian vs. Czech

Russian East Slavonic, Czech West Slavonic

Syntax/Morphosyntax

- Grammatical functions by inflection
- Constituent order determined mostly by Information Structure.
- Agreement: subj-verb (person, nr), subj-participle (gender, nr), within NP (gender, nr, case)
- No articles; (in)definiteness is expressed using other means, e.g., word order.
- Certain rigid word order combinations, such as noun modifiers, clitics (in Czech), and negation (in Russian).

Russian & Czech Morphology

- The order and value of morphemes nearly identical
- Similar shape of morphemes (modulo scripts)
- Nominal categories inflect for gender, number, case.
  - 3 genders (masculine, feminine, neuter)
  - 2 numbers (some remnants of dual in Czech).
  - 6 cases with roughly the same meaning (nominative, genitive, dative, accusative, local, instrumental).
    In addition, Czech has vocative.
- Nouns and verbs are grouped into paradigms.
- Numerals use declensional strategies which range from near indeclinability to adjective-like declension.

Czech and Russian paradigms

	Czech	Russian	Gloss
sg.			
nom	žen-a	ženščin-a	‘woman’
gen	žen-y	ženščin-y	
dat	žen-ě	ženščin-e	
acc	žen-u	ženščin-u	
voc	žen-o	–	
loc	žen-ě	ženščin-e	
ins	žen-ou	ženščin-oj/ou	
pl.			
nom	žen-y	ženščin-y	
gen	žen	ženščin	
dat	žen-ám	ženščin-am	
acc	žen-y	ženščin	
voc	žen-y	–	
loc	žen-ách	ženščin-ax	
ins	žen-ami	ženščin-ami	

Czech and Russian Morphology

Morphology in both languages exhibits

- a high number of fusion – several morphemic categories whose values are combined in clusters, each of which is expressed by a single ending (e.g., number, gender, and case with nouns or adjectives, or tense, number, and person with finite verbs),
  - the Russian *knig-oj*, 'book', *-oj* stands for feminine, singular, instrumental;
  - *pročital-a -a* stands for past tense and feminine.
- a high degree of ambiguity of the endings. See the two next slides.
- a relatively common synonymy of the endings.

Questions we try to address

- Are word order properties of Czech and Russian similar enough to approximate the target language word order by the source language word order?
- What kind of morpho-syntactic descriptions are relevant for these languages in general and for the annotation transfer in particular?
- How close is a particular pair of languages in the lexicon?
- Can lexical similarities be used to improve the morpho-syntactic transfer?
- How can the data sparsity problem be addressed in the cross-lingual induction of morpho-syntactic features of highly inflected languages?

Tagging Russian via Czech

- Direct
- Approximating Emissions
  - Even
  - Cognates
- Approximating Transitions

Using TnT (Brants 2000), a second order Markov Model tagger

- emissions: approximated by the source-language emissions + resource-light morphological analysis
- transitions: approximated by the source-language transitions

Resources

- Limited language dependent resources:
  - Manually created list of paradigms and closed class words
  - Annotated development corpus: 1,788 tokens from Orwell's *1984*
  - Raw Russian corpus: 1M tokens of Uppsala Corpus[2]
- Testing corpus: 4,011 tokens from Orwell's *1984*
- Russian Positional tagset
  Size: Russian 2000+; Czech 4000+, English 45 (Penn Treebank)

[2] http://www.slaviska.uu.se/ryska/corpus.html

Tagset

Table: Overview and comparison of the Czech and Russian tagsets

Pos	Description	Abbr.	No. of values	
			Czech	Russian
1	POS	p	12	12
2	SubPOS – detailed POS	s	69	45
3	Gender	g	11	5
4	Number	n	6	4
5	Case	c	9	8
6	Possessor's Gender	f	5	5
7	Possessor's Number	m	3	3
8	Person	e	5	5
9	Tense	t	5	5
10	Degree of comparison	d	4	4
11	Negation	a	3	3
12	Voice	v	3	3
13	Unused		1	1
14	Unused		1	1
15	Variant, Style	i	10	8

Tag translation

- Translate to the corresponding category in Russian (if obvious)
  - e.g., vocative → nominative; Pronominal clitics → pronouns, etc.
- Drop distinctions Russian does not make.
  - e.g., short adjectives do not distinguish case, verbs do not distinguish negation.
- Ignore rare tags.
- Some translations are not obvious:
  - Czech participles: *QW* (fem, sg OR neutr.pl) can be translated as Russian *FS* (fem,sg) or *NP* (neutr,pl), but Russian particples do not distinguish gender in plural (*XP*).

Script Modification

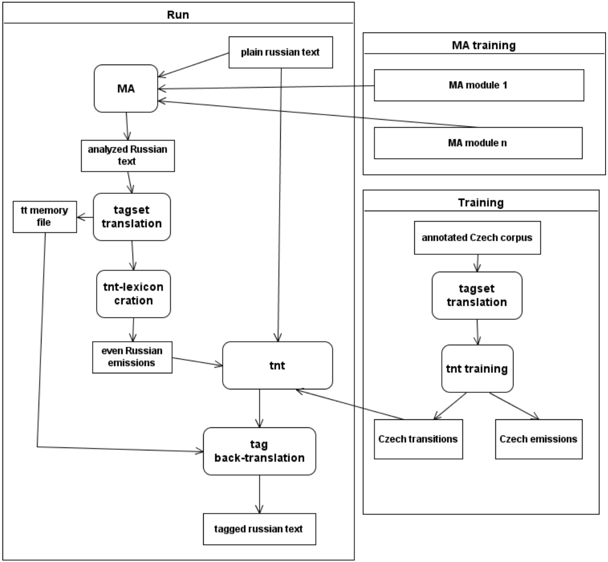
- Russian and Czech use different scripts
- Cannot use emissions directly
- Transliterate Russian, using Scientific Transliteration
  - e.g., it produces *š* for [ʃ] and *č* for [tʃ].
- Replace Czech characters not present in the transliterated Russian with their obvious (or most likely) counterparts.
  - e.g., long vowels are shortened ($\acute{a} \rightarrow a$), palatalization is expressed using the soft sign ($\check{n} \rightarrow n'$), etc.

Direct Tagger

Table: Direct Tagger: Czech tagger applied to Russian

tagger name			direct
		Scientific transliteration	Better transliteration
Unknown tokens (%)		59.0	55.3
All	Full tag:	44.9	48.1
	SubPOS	61.0	63.8
N	Full tag:	32.8	37.3
	SubPOS	84.0	81.1
A	Full tag:	20.7	31.7
	SubPOS	33.8	51.7
V	Full tag:	36.1	39.9
	SubPOS	44.6	48.1

Even Tagger


Run
plain russian text
MA
analyzed Russian text
tt memory file
tagset translation
tnt-lexicon cration
even Russian emissions
tnt
tag back-translation
tagged russian text
MA training
MA module 1
MA module n
Training
annotated Czech corpus
tagset translation
tnt training
Czech transitions
Czech emissions

Even Tagger: Results

Table: Tagging with evenly distributed output of Russian MA

tagger name		Direct	Even
transitions		Czech	Czech
emissions		Czech	uniform Russian MA
All	Full tag:	48.1	77.6
	SubPOS	63.8	91.2
N	Full tag:	37.3	54.4
	SubPOS	81.1	89.6
A	Full tag:	31.7	53.1
	SubPOS	51.7	86.9
V	Full tag:	39.9	90.1
	SubPOS	48.1	95.7

Approximating emissions

- Thus far, we used evenly distributed emissions, i.e. we lost some useful information
  - Identify source-target cognate pairs
  - Transfer the information about the source cognate word to the target cognate word

Cognates: Hypotheses

Cognate words

- will have similar morphological and distributional properties.
- are similar in form and this tendency is strong enough to be useful.

Cognates (cont.)

We are aware of the fact that

- Cognates could have departed in their meaning, and thus probably have different distributions.
  - *život* 'life' in Czech vs. *život* 'belly' in Russian, and *krásný* (adj.) 'nice' in Czech vs. *krasnyj* (adj.) 'red' in Russian.
- Cognates could have departed in their morphological properties.
  - *tema* 'theme', borrowed from Greek, is neuter in Czech and feminine in Russian.
- There are false cognates — unrelated, but similar or even identical words.
  - *dělo* 'cannon' in Czech vs. *delo* 'matter, affair' in Russian, *jel* [jɛl] 'drove' in Czech vs. *el* [jɛl] 'ate' in Russian, *pozor* 'attention' in Czech vs. *pozor* 'disgrace' in Russian, *ni* 'she$_{loc}$' in Czech vs. *ni* negative particle in Russian (corresponding to Czech *ani*).

Automatic cognate detection

- A variant of the edit distance where the cost of operations is dependent on the arguments:
  - Characters sharing certain phonetic features are closer than characters not sharing them (we use spelling as an approximation of pronunciation; E.g., *b* is closer to *p* than to, say, *j*.
  - Costs are refined based on some well-known and common language-specific phonetic-orthographic regularities. E.g.,
    - Russian *è* and *e* have zero distance from Czech *e*.
    - Czech *h* and *g* have zero distance from Russian *g* (in Czech, the original Slavic *g* was replaced by *h*, in Russian it was not).
    - The length of Czech vowels is ignored (in Russian, vowel length is not phonemic)
    - *y* and *i* are closer to each other than other vowels (modern Czech does not distinguish between them in pronunciation)

Cognates (cont.)

- Cognates are translated back to their original spelling.
- ED is affected by the number of arguments (characters) it needs to consider $\rightarrow$ normalize by word length.
- The list of cognates includes all Czech-Russian pairs of words whose distance is below a certain threshold.
- We require that the words have the same morphological features (except for the gender of nouns and the variant as they are lexical features).

Using cognates

- Map the Czech emission probabilities to Russian emissions.
  - Assume w_{cze} and w_{rus} are cognate words.
  - Let T_{cze} denote the tags that w_{cze} occurs with in the Czech training corpus.
  - Let $p(w_{cze}|t)$ be the emission probability of w_{cze}
  - Let T_{rus} denote tags assigned to w_{rus} by the morphological analyzer; $\frac{1}{|T_{rus}|}$ is the even emission probability.
  - Then, assign the new emission probability $p'(w_{rus}|t)$ to every tag $t \in T_{rus}$ (followed by normalization):

$$(1) \quad p'(w_{rus}|t) = \begin{cases} p(w_{cze}|t) + \frac{1}{|T_{rus}|} & \text{if } t \in T_{rus} \\ 0 & \text{otherwise} \end{cases}$$

Approximating transitions

- Czech transitions are a fairly good approximation of Russian transitions.
- Nevertheless, there's a drop in accuracy (especially for verbs), when compared to the native Russian transitions.
- Russify data.

Approximating transitions (examples)

Negation in Czech is expressed by the prefix *ne*-, whereas in Russian it is very common to see a separate particle (*ne*) instead:

(2) a. Nic **nerekl**.
nothing not-said
'He didn't say anything.' [Cz]

b. On ničego **ne** **skazal**.
he nothing not said
'He didn't say anything.' [Ru]

Approximating transitions (examples)

Reflexivization of verbs is expressed by a separate word in Czech, and by affixation in Russian.

(3) a. Filip **se** ještě neholí.
Filip REFL-CL still not-shaves
'Filip doesn't shave yet.' [Cz]

b. Filip esče ne breet+**sja**.
Filip still not shaves+REFL.SUFFIX
'Filip doesn't shave yet.' [Ru]

Approximating transitions (examples)

Even though auxiliaries and the copula are the forms of the same verb *být/byt'* 'to be', both in Czech and in Russian, the use of this verb is different in the two languages. For example, Russian does not use an auxiliary to form past tense:

(4) a. Já **jsem** psal.
I aux$_{1sg}$ wrote
'I was writing/I wrote.' [Cz]

b. Ja pisal.
I wrote
'I was writing/I wrote.' [Ru]

Russified transitions: examples

	Czech	Russian	
(5)	Já **bych** spal.	Ja **by** spal.	'I would sleep.'
	Ty **bys** spal.	Ty **by** spal.	'You.sg would sleep.'
	On **by** spal.	On **by** spal.	'He would sleep.'

Russified transitions: results

Table: Tagging Russian using Russified Czech transitions

tagger name		cognates	russified
transitions		Czech	Russified Czech
emissions		cognates	cognates
All	Full tag:	79.5	80.0
	SubPOS	92.2	92.3
N	Full tag:	57.3	57.1
	SubPOS	89.9	89.3
A	Full tag:	54.5	55.9
	SubPOS	86.9	86.9
V	Full tag:	90.6	92.7
	SubPOS	96.1	96.6

Russified transitions: discussion

- Russifications are language specific and therefore do not fit into our goal of developing a resource- and knowledge-light framework.
- The penalty for using Czech transitions is very small (although this might be different for other languages)
- Some improvements in transitions are the results of the tagset translation, which are part of the most basic tagger.

Tag decomposition

- Data sparsity problem (large tagset): with 1,000 tags there are $1,000^3$ potential trigrams.
- Decompose the tag into subtags to reduce the tagset
- We focus on six positions — POS (p), SubPOS (s), gender (g), number (n), case (c), and person (e). The selection of the slots is based on linguistic intuition.
- Train the tagger on the subtags
- Combine them

Combination of subtaggers

There are many possible formulas that could be used. E.g.,

(6) $\text{bestTag} = \text{argmax}_{t \in T_{MA}} \text{val}(t)$

where:

1. T_{MA} is the set of tags offered by MA
2. $\text{val}(t) = \sum_{k=0}^{14} N_k(t)/N_k$
3. $N_k(t)$ is the # of taggers voting for k-th slot of t
4. N_k is the total # of taggers on slot k

This formula means that the best tag is the tag that receives the highest average percentage of votes for each of its slots.

- No significant improvement in performance

Summary of results

		direct	even	cog	russif
emissions		cz	MA	cog	cog
transitions		cz	cz	cz	cz_{ru}
All	Full tag:	45.6	77.6	79.3	79.7
	SubPOS	62.3	90.4	**91.4**	**91.3**
N	Full tag:	36.7	59.6	61.2	62.1
	SubPOS	81.9	89.5	89.8	89.8
A	Full tag:	18.9	62.5	64.7	65.8
	SubPOS	36.1	86.5	86.8	86.8
V	Full tag:	44.1	93.0	93.2	**93.9**
	SubPOS	54.3	95.5	95.7	95.7

Comparisons with other tools

- Czech taggers (Hajic et al. 2001) – significantly better (4.84% error r.)
  - However, extensive lexicon (300K entries) with 1.5% recall error
  - Taggers trained and tested on the same language
- Xerox Russian Tagger – worse (but not a real evaluation)
  - Much smaller tagset (63 tags, collapsing some cases, ...)
  - Error rate comparison on 201 tokens of the testing corpus: Xerox tagger: 18%; our tagger: 8.5%;