Slavistik/Korpuslinguistik Regensburg

Finite-State Guesser

- Xerox's fsm software (Beesley & Karttunen 2003)
- 92 LEXICON units (continuation classes)
- regexes for orthographic normalization and phonological rules (18)
- small lexicon of stems and frequent function words
- If no stem in the lexicon fits, "guessed stems" inserted, following very basic phonotactic rules
- guesser vastly overgenerates, e.g.

начяти	начя +Guess+Part+Prt+Pass+Kf+Masc+Nom+Pl
начяти	начя +Guess+Verb+Inf
начяти	начят +Guess+Noun+Fem+Anim+Hum+Loc+Sg
начяти	начят +Guess+Noun+Fem+Anim+Hum+Nom+Du
начяти	начят +Guess+Noun+Fem+Anim+Hum+Dat+Sg

Slavistik/Korpuslinguistik Regensburg

Pure Guesser

Otecь	Verb:Sg:Aor:2P Verb:Sg:Aor:3P Noun:Nom:Sg:Masc:Anim:Hum Noun:Nom:Sg:Masc:Anim:NonHum	father
moj	PPos:Nom:Sg:Masc:1P	my
umerlъ	Verb:Sg:Masc:Prt:Pf:NonRefl	died
,	Punc	
a	Con	and
Svjatopolkъ	Part:Sf:Acc:Sg:Masc:Prt:Act Noun:Acc:Sg:Masc:NonAnim	S.
sěditь	Verb:Sg:Pres:3P Noun:Acc:Sg:Fem:Anim:NonHum Noun:Nom:Sg:Fem:NonAnim Noun:Acc:Sg:Fem:NonAnim Noun:Acc:Sg:Fem:Anim:Hum Adj:Sf:Nom:Sg:Masc Noun:Acc:Sg:Masc:Anim:Hum Noun:Acc:Sg:Masc:NonAnim Noun:Nom:Sg:Masc:NonAnim Noun:Nom:Sg:Masc:Anim:Hum Adj:Sf:Acc:Sg:Masc Noun:Nom:Sg:Masc:Anim:NonHum Noun:Acc:Sg:Masc:Anim:NonHum Noun:Nom:Sg:Fem:Anim:NonHum Noun:Nom:Sg:Fem:Anim:Hum	sits
Kyevě	Name:Loc:Sg:Masc:NonAnim	(in) Kiev

Slavistik/Korpuslinguistik Regensburg

Guesser with lexicon (best case)

Otecь	Verb:Sg:Aor:2P Verb:Sg:Aor:3P Noun:Nom:Sg:Masc:Anim:Hum Noun:Nom:Sg:Masc:Anim:NonHum
moj	PPos:Nom:Sg:Masc:1P
umerlъ	Verb:Sg:Masc:Prt:Pf:NonRefl
,	Punc
a	Con
Svjatopolkъ	Part:Sf:Acc:Sg:Masc:Prt:Act Noun:Acc:Sg:Masc:NonAnim Noun:Nom:Sg:Masc:NonAnim
sěditь	Verb:Sg:Pres:3P Noun:Nom:Sg:Fem:NonAnim Noun:Acc:Sg:Fem:NonAnim Noun:Acc:Sg:Fem:Anim:Hum Adj:Sf:Nom:Sg:Masc Noun:Acc:Sg:Masc:Anim:Hum Noun:Acc:Sg:Masc:NonAnim Noun:Nom:Sg:Masc:NonAnim Noun:Nom:Sg:Masc:Anim:Hum Adj:Sf:Acc:Sg:Masc Noun:Nom:Sg:Masc:Anim:NonHum Noun:Acc:Sg:Masc:Anim:NonHum Noun:Nom:Sg:Fem:Anim:NonHum Noun:Nom:Sg:Fem:Anim:Hum
Kyevě	Name:Loc:Sg:Masc:NonAnim

Slavistik/Korpuslinguistik Regensburg

Word Alignment (NRu–ORu normalized)

- statistical word alignment
  - general-purpose tools ignore the close relation between the aligned languages
  - Uplug (GIZA++) tested
  - potentially useful for aligning the orthographically non-normalized original and the normalized version — sentence boundaries?
- alternative approach: edit distance
  - for each ORu word, look for a perfect match in the NRu parallel segment
  - where there is no perfect match, select the modern word in the same segment which is minimally orthographically distinct from the ORu form
  - Levenshtein edit distance, adapted by
    - disregarding predictable orthographic variation
    - relativized to word length
    - introducing an upper threshold value for possible cognates

Slavistik/Korpuslinguistik Regensburg

Simple edit distance with obligatory matches

1	8	начяти	+Guess+Part+Prt+Pass+Kf+Masc+Nom+Pl ...	
начать				
1	9	старыми	+Guess+Adj+Lf+Neut+Ins+Pl ...	старыми
1	10	словесы	+Noun+Neut+NonAnim+Ins+Pl ...	словами
1	11	трудныхъ	+Guess+Adj+Lf+Neut+Loc+Pl ...	братья
1	12	повѣстій	+? ...	повести
1	13	о	+Guess+Verb+Aor+2P+Sg ...	о
1	14	пълку	+Guess+Noun+Fem+Anim+Hum+Gen+Du ...	ли
1	15	Игоревѣ	+Guess+Verb+Pres+1P+Du ...	Игоревом
1	16	,	+? ...	,
1	17	Игоря	+Guess+Verb+Aor+3P+Sg ...	Игоря
1	18	Святъславлича	+? ...	Святославича
1	19	?	+? ...	?2

Slavistik/Korpuslinguistik Regensburg

Improved edit distance

- "Yers" (ь, ъ) are disregarded
- ѣ/и and ѣ/е alternations do not count
- threshold relative to word length

Otecь	Otec	father
moj	moj	my
umerlъ	umer	died
,	,	,
a	a	*and*
Svjatopolkъ	Svjatopolk	S
sědiть	sidit	sit
Kyevě	Kieve	(in) Kiev
.	i	
izbivaja		beating-up
bratьju	brat'ej	brothers
svoju	svoich	his

Slavistik/Korpuslinguistik Regensburg

Tag projection (1)

- analyze w's and v's tags and transform them into a common tagset (properties of a tag object)
- simple transfer of NR tag where *no* OR tag could be guessed
- common case: use *part of* the NRu tag to select from guessed OR tags
- unification of part-of-speech and grammatical categories
- beware of ...
  - gender (syncretisms in declensions not yet related to animacy)
  - number (OR had a dual)
  - tense (OR had many past tenses – aorist, imperfect, pluperfect ...)
- tag projection removes implausible tags from the guessed ORu tags
- (guessed tags are preserved independently)

Slavistik/Korpuslinguistik Regensburg

Tag projection (2)

- The NRu corpus is not tagged perfectly — potential errors due to NRu syncretisms should be avoided.
- *Nominal gender and animacy* may be projected from NRu to ORu.
- *Number* may be projected with some proviso, i.e., Sg (NRu) => Sg (ORu), but Pl (NRu) => Pl or Du (ORu)
- *Unambiguous* noun tags in NRu show no systematic errors of the tagger; thus they may transfer case and number (Sg vs. non-Sg).
- Currently,
  - full projection into ORu for identical word forms without a ORu tag
  - some variability for names vs. nouns
  - rules hard-wired into program code, no separate rule language

Slavistik/Korpuslinguistik Regensburg

Output with projection rules

Otecь	Noun:Nom:Sg:Masc:Anim:Hum
moj	PPos:Nom:Sg:Masc:1P
umerlъ	Verb:Sg:Masc:Prt:Pf:NonRefl
,	Punc
a	Con
Svjatopolkъ	Noun:Acc:Sg:Masc:NonAnim
	Noun:Nom:Sg:Masc:NonAnim
sěditь	Verb:Sg:Pres:3P
Kyevě	Name:Loc:Sg:Masc:NonAnim
,	Punc
izbivaja	
bratьju	Noun:Ins:Sg:Neut:NonAnim
svoju	PPos:Gen:Pl:X
».	Punc

Slavistik/Korpuslinguistik Regensburg

Many problems for pure guessing + projection ...

I	Con	
sъbra	NumCard:Acc:X	|*should be Verb in Aorist*|
Jaroslavъ	Name:Nom:Sg:Masc:Anim:Hum	
	Name:Nom:Sg:Masc:Anim:NonHum	
varjagъ	Noun:Gen:Pl:Masc:Anim:NonHum	
	Noun:Gen:Pl:Masc:Anim:Hum	
tyjsačju	Noun:Acc:Sg:Neut:NonAnim	
,	Punc	
i	Con	
poide	Verb:Sg:Masc:Prt:Pf:NonRefl	|*should be Verb in Aorist*|
na	Prp:Acc	
Svjatopolka	Noun:Nom:Sg:Neut:NonAnim	
...		
Ne	Ptcl	
ja	PPer:Nom:Sg:1P	
počachъ	PPos:Loc:Pl:X:1P	|*should be Verb in Aorist*|
izbivati	Verb:Inf	Verb:Sg:Aor:2P	Verb:Sg:Aor:3P
bratьju	Noun:Gen:Pl:Masc:Anim	

Slavistik/Korpuslinguistik Regensburg

Syncretism in NRu noun paradigms

		Masc:Anim	*Masc:NonAnim*	*i-Class*	*Fem:Anim*	*Fem:NonAnim*	*Neut*
SG	NOM	*volk*	*les*	*put'*	*žena*	*zemlja*	*okno*
	GEN	*volka*	*lesa*	*puti*	*ženy*	*zemli*	*okna*
	DAT	*volku*	*lesu*	*puti*	*žene*	*zemle*	*oknu*
	ACC	*volka*	*les*	*put'*	*ženu*	*zemlju*	*okno*
	LOC	*volke*	*lesu (-e)*	*puti*	*žene*	*zemle*	*okne*
PL	NOM	*volki*	*lesy*	*puti*	*ženy*	*zemli*	*okna*
	GEN	*volkov*	*lesov*	*putej*	*žen*	*zemel'*	*okon*
	ACC	*volkov*	*lesy*	*puti*	*žen*	*zemli*	*okna*

- Anim => Pl:Acc = Pl:Gen; NonAnim => Pl:Acc = Pl:Nom
- Masc:Anim => Sg:Acc = Sg:Gen; Masc:NonAnim => Sg:Acc = Sg:Nom
- i-Class => Sg:Gen = Sg:Dat = Sg:Loc = Pl:Nom
- Fem => Sg:Dat = Sg:Loc; Fem => Sg:Gen = Pl:Nom

Slavistik/Korpuslinguistik Regensburg

Syncretism in ORu nominal paradigms

SG	NOM	*vьlkъ*	*konь*	*synъ*	*gostь*	*kamy*	*žena*	*zemlja*
	GEN	*vьlka*	*konja*	*synu*	*gostja*	*kamene*	*ženy*	*zemli*
	DAT	*vьlku*	*konju*	*synovi*	*gosti*	*kamenu*	*ženě*	*zemli*
	ACC	*vьlkъ*	*konь*	*synъ*	*gostь*	*kamenь*	*ženu*	*zemlju*
	LOC	*vьlcě*	*koni*	*synu*	*gostju*	*kamene*	*ženě*	*zemli*
	VOC	*vьlče*	*konju*	*synu*	*gostju*	*Ø*	*ženo*	*zemle*
DU	NOM=ACC	*vьlka*	*konja*	*syny*	*gosti*	*kamenu*	*ženě*	*zemli*
	GEN=LOC	*vьlku*	*konju*	*synovu*	*gostьju*	*kamenu*	*ženu*	*zemlju*
PL	NOM	*vьlci*	*koni*	*synove*	*gostьe*	*kamene*	*ženy*	*zemlja*
	GEN	*vьlkъ*	*konь*	*synovъ*	*gostьi*	*kamenь*	*ženъ*	*zemlь*
	ACC	*vьlky*	*koně*	*syny*	*gosti*	*kamene*	*ženy*	*zemlje*
	INS	*vьlky*	*koni*	*synъmi*	*gostьmi*	*kamene*	*ženami*	*zemljami*

Slavistik/Korpuslinguistik Regensburg

Unambiguous items in NRu

- Masc:Sg:Loc (excluding the "second locative"); Masc:Anim:Sg:Nom, Masc:Anim:Sg:Dat, Masc:Anim:Pl:Nom; Masc:NonAnim:Sg:Gen, Masc:NonAnim:Pl:Gen
- Fem:Sg:Acc; Fem:NonAnim:Pl:Gen
- Neut:Sg:Dat, Neut:Sg:Loc, Neut:Pl:Gen
- Masc:Sg:Ins, Neut:Sg:Ins
- Pl:Dat, Pl:Ins, Pl:Loc
- respective rule for each of the above tags:
  - if ORu guessed set contains a tag compatible with this tag, remove all tags containing a different case or number
  - (irrespective of ORu syncretisms)

Slavistik/Korpuslinguistik Regensburg

Informative syncretisms in NRu

- partial overlap in potential tags, e.g.
  - ORu: ({Sg:Gen}|...) & NRu: ({Sg:Gen}|{Anim:Sg:Acc})
    => ORu: {Sg:Gen}
  - ORu: ({Pl:Gen}|...) & NRu: ({Pl:Gen}|{Anim:Pl:Acc})
    => ORu: {Pl:Gen}

NB: We do not find Gen/Acc syncretisms in the Sg or Pl in early ORu

- ORu: ({Sg:Nom}|{Sg:Acc}|{Pl:Gen}|...)
  & NRu: ({Acc}|{Nom}) => ORu: ({Sg:Nom}|{Sg:Acc})

NB: The NRu tag is never more than two-way ambiguous.

Slavistik/Korpuslinguistik Regensburg

Perspectives

- Implemented
  - sentence alignment Nru/ORu-normalized; word alignment (cf. references)
  - guesser and small ORu lexicon
  - edit distance
  - tag projection rules
- Fine-tuning
  - word alignment: adding safe bets; hand-correction?
- Further gains
  - Bootstrapping an enlarged lexicon via word alignment
  - Extracting orthographic variants

Slavistik/Korpuslinguistik Regensburg

References

Gale, W. & Church, K. 1994. "A Program for Aligning Sentences in Bilingual Corpora", in: Armstong, S.(ed.). *Using Large Corpora*. MIT Press, 91-102.

Meyer, R. 2009. "Old wine in new wineskins? — Tagging Old Russian via projection from modern translations", submitted to *Russian Linguistics*.

Tiedemann, J. 2006. "ISA & ICA - Two Web Interfaces for Interactive Alignment of Bitexts", in: *Proceedings of LREC 2006*. Genova.

Thank you!

roland.meyer@sprachlit.uni-regensburg.de
http://www-slavistik.uni-r.de/subjekte