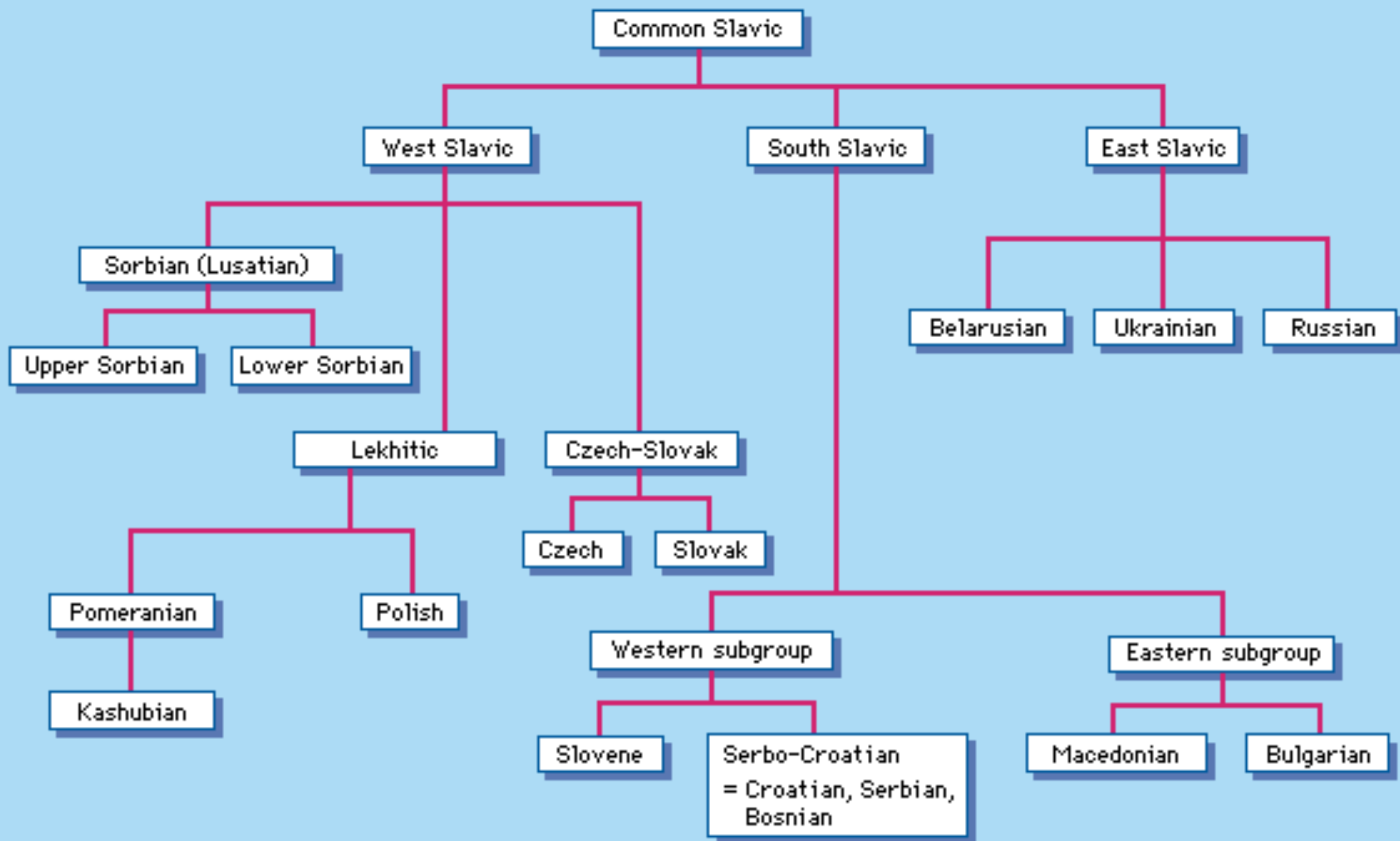


Serbo-Croatian

srpskohrvatski // хрватскосрпски

Arvind Srinivasan








Language Family

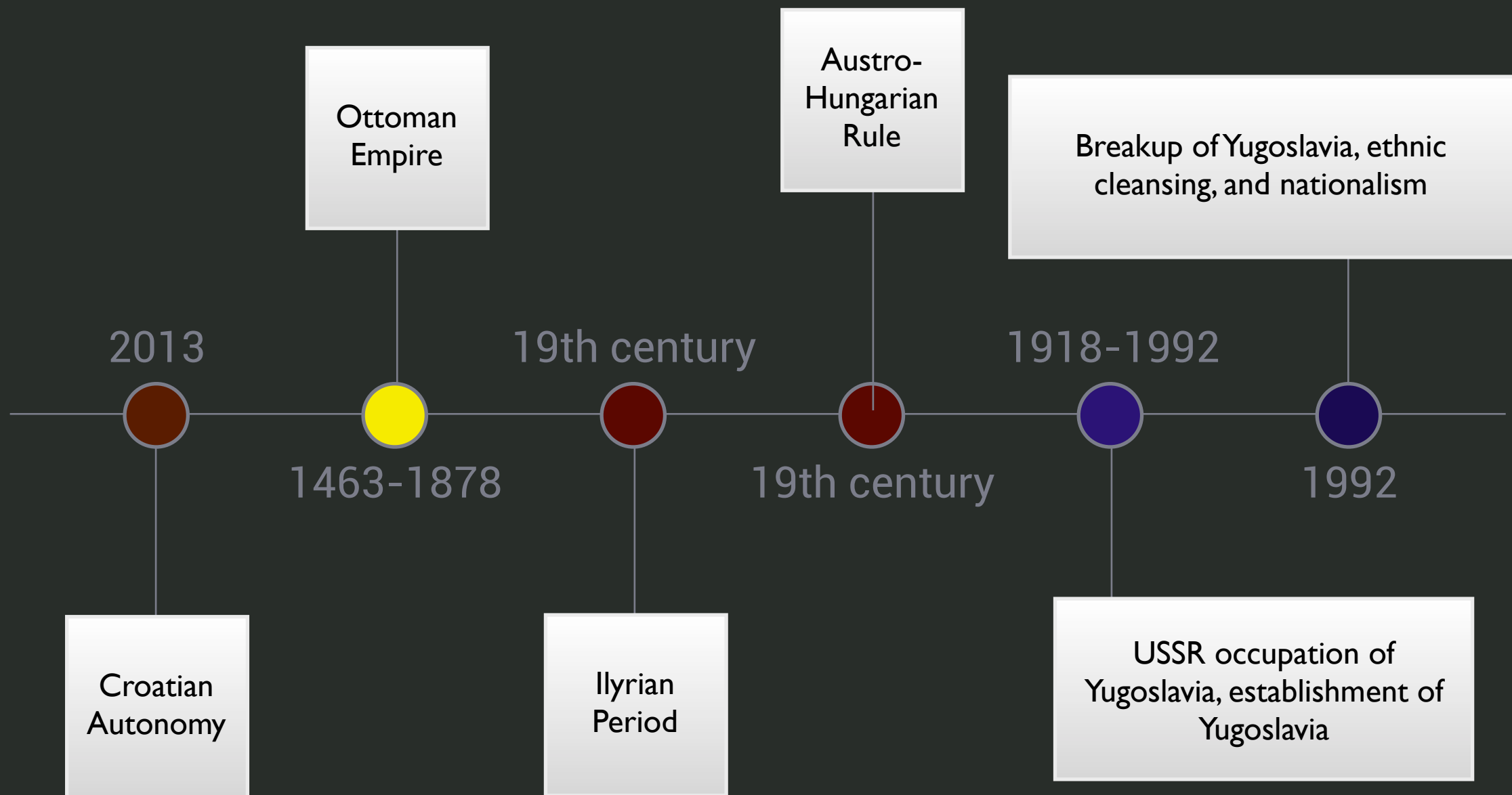
		
<ul style="list-style-type: none"> • 2m speakers • Latin majority, Cyrillic minority 	<ul style="list-style-type: none"> • 6m speakers • Only Latin Alphabet 	<ul style="list-style-type: none"> • 11m speakers • Cyrillic official, Latin colloquial
<ul style="list-style-type: none"> • januar • Allahimanet • naučnik 	<ul style="list-style-type: none"> • siječanj • Do viđenja • znanstvenik 	<ul style="list-style-type: none"> • januar • Doviđenja • naučnik

Loanwords

		
<ul style="list-style-type: none">• Majority Islam• Arabic, Turkish, Persian Loanwords• Middle Eastern Influence	<ul style="list-style-type: none">• Fewest loanwords, most standardized• Ilyrian period (late 19th century) codifies modern Croatian• vocabulary modification to "more soviet" words during USSR occupation• usually translate borrowed words	<ul style="list-style-type: none">• Old Serbian borrows heavily from Greek• Borrowed from Turkish during Ottoman rule• German + Italian during late 19th, early 20th centuries• 85% of new words post-1982 are from English• adopt transliterated versions of foreign words

http://www.serbianstudies.org/publications/pdf/Vol14_2_Gorup.pdf

Occupational History



Dialects

- Three major dialects, named for word meaning “what” : Kajkavian (kaj), Čakavian (ča), Štokavian (što)
 - dialects not mutually understandable
- Štokavian accents are mutually understandable (Ekavian, Ikavian, Ijekavian)
 - based on form of slavic vowel Jat
- Official documents will use Štokavian normal form (vrěme == {vrime, vreme, vrijeme})
- Transliteration understandable, but incorrect (BLEU penalty).

Dialects

Kajkavian (kaj)	Čakavian (ča)	Štokavian (što)
<ul style="list-style-type: none"> • 30% of Croatians (North Croatia) • spoken at home, Štokavian outside • most of historical Serbo-Croatian texts written in Kajkavian 	<ul style="list-style-type: none"> • oldest Croatian dialect • mostly dead, spoken by minority of Croatians (12%) • majority of literary Croatian pre-15th century 	<ul style="list-style-type: none"> • majority of public serbo-croatian • shared between Bosnian, Serbian, Croatian • closest to Church Slavonic
"Japek naš ki si v nebesaj"	n/a	"Oče naš, koji jesi na nebesima"

Orthography (Digraphia)

Table 33: The Serbo-Croatian Alphabet

letters				letters			
Croatian		Serbian		Croatian		Serbian	
capital	lower-case	capital	lower-case	capital	lower-case	capital	lower-case
A	a	А	а	L	l	Л	л
B	b	Б	б	Lj	lj	Љ	љ
C	c	Ц	ц	M	m	М	м
Č	č	Ч	ч	N	n	Н	н
Ć	ć	Ћ	ћ	Nj	nj	Њ	њ
D	d	Д	д	O	o	О	о
Dž*	dž*	Џ	џ	P	p	П	п
Đ†	đ†	Ђ	ђ	R	r	Р	р
E	e	Е	е	S	s	С	с
F	f	Ф	ф	Š	š	Ш	ш
G	g	Г	г	T	t	Т	т
H	h	Х	х	U	u	У	у
I	i	И	и	V	v	В	в
J	j	Ј	ј	Z	z	З	з
K	k	К	к	Ž	ž	Ж	ж

*Alphabetized in *Britannica* as *dz*. †Alternatively, *dj*.

- Serbian uses Cyrillic officially in Government documents, but many use Latin keyboards
- Mehmed Džemaludin Čaušević invents Arebica, not used post-WW2

arabica	latinica	arabica	latinica	arabica	latinica
آ	a	غ	g	ؤ	o
ب	b	ح	h	پ	p
چ	c	ای	i	ر	r
چ	č	ي	j	س	s
چ	ć	ق	k	ش	š
د	d	ل	l	ت	t
ج	dž	ل	lj	ؤ	u
خ	d	م	m	و	v
ه	e	ن	n	ز	z
ف	f	ن	nj	ژ	ž

Grammar - (Pro)nouns

- 3 genders, 7 cases (nominative, genitive, dative, accusative, vocative, locative, instrumental)
 - ex: N: dečak, G: dečaka, D: dečaku, ...
- Agreement for nouns, case, verb (noun endings are modified based on case + gender) unless they are special nouns (e-type or i-type)
- All noun modifications are standardized across Serbo-Croatian languages
- Optional pronoun drop
- Pronouns have two forms, longer and shorter (Mene == me, Tebe == te, etc)

Grammar – Verbs

- Gender affects verb conjugation
 - (M) : Ja sam radio vs (F) : Ja sam radila.
- Simple/ambiguous present tense (no present perfect, present continuous, etc.). Past tense formed by to be (biti) + perfect participle
- Also have 2 future tenses and a rarely used pluperfect, aorist, imperfect tense
 - less restrictive in Serbian over Croatian for formation of future tense (have additional allowed constructions + can merge the two verbs in Serbian)
- Mood denoted by morpheme affix (imperative, conditional, optative)
- Since there is no temporal descriptiveness in tense, verbs have aspect
 - many classes of prefixes that can make verbs perfective <-> imperfective
 - aspect can express several things, incl. whether an action is in progress, completed, iterative
- 3 verbs, hjeti, miti, imati, can be negated with a prefix + conjugation, i.e. (neću vs. ne ću)
 - can also attach pronoun to verb as “contraction”

Grammar (the rest)

- S-V-O word order, but non-restrictive. Lots of word-order permutations allowed, since declensions are highly expressive.
 - Man bites dog: "Čovjek grize psa"
 - Dog bites man: " Čovjeka grize pas"
- Adjectives have standardized endings based on case, number, and gender
- consonant-a-consonant ending adjectives can drop the a when attaching an ending (dobar can become dobri or dobari, but either is permissible)
- Capitalization similar to English, except pronouns are only capitalized in formal speech

SMT - Corpora

EVROTEKA

EU legislation
texts, parallel
Serbian-English

- 1428720 words
- <http://prevodjenje.seio.gov.rs/evroteka/index.php?jezik=engl>

hrenWaC

EU legislation
texts, parallel
Croatian-English

- 99,000 sentence pairs
- <http://www.nljubasic.net/resources/corpora/hrenwac/>
- other resources: <http://www.hnk.ffzg.hr/txts/mt4bratislava.pdf>

SETimes

Balkan news, aligned
in 10 languages, incl.
Bosnian, Serbian,
Croatian, English

- ~200,000 sentences, based on language pair
- <http://www.nljubasic.net/resources/corpora/setimes/>
- Original Paper: <http://xixona.dlsi.ua.es/~fran/publications/lrec2010.pdf>

hrWaC, srWaC

Monolingual
Serbian, Croatian,
Slovene scraped
from the web

- 1.2 bn words Croatian, 1bn words Slovene, WIP 1.2 bn Serbian
- <http://www.nljubasic.net/resources/corpora/{hrwac,slwac,srwac}/>
- Original Paper: <http://www.nljubasic.net/upload/ljubasic11-hrwac.pdf>

SMT Research

- Statistical Machine Translation of Serbian-English (Popović, Jovićić, Šarić, 2004)

Table 5: Examples of Serbian-English translations with and without transformations

to je mali grip , ništa ozbiljno .	⇒ transformations	to je mal- grip , ništa ozbilj- .
↓ Sr → En (baseline)		↓ Sr' → En
it is a touch of flu , nothing UNKNOWN-ozbiljno .		it is a small flu , nothing serious .
hajde da pogledamo neki izraz sa glagolom ``get'' .	⇒ transformations	hajde da pogleda- nek- izraz- sa glagol- ``get'' .
↓ Sr → En (baseline)		↓ Sr' → En
let us look at some expressions with UNKNOWN-glagolom ``get'' .		let us look at some expressions with verbs ``get''
svi su u isto vreme pokušavali da udju u autobus .	⇒ transformations	svi su u ist- vreme pokušav- da udj- u autobus .
↓ Sr → En (baseline)		↓ Sr' → En
everyone in as UNKNOWN-pokušavali time to in in in bus .		everyone in same time trying to come at the bus .

- First test of phrase based for S, E pair
- Assimil corpus (small: 3k sentences)
- Uses stemming to reduce morphological complexity, match to English complexity (pro-drop, gender)
- Stemming reduces error by 8%

Table 4: Translation error rates [%]
for English→Serbian

En→Sr	Develop		
	WER	PER	1-BLEU
Baseline	46.1	41.0	76.5
	Test		
	WER	PER	1-BLEU
Baseline	55.3	48.7	80.3

Table 2: Examples of reduced Serbian words

original	stem	English
mali	mal-	small (boy)
mala	mal-	small (girl)
malim	mal-	(with a) small (boy)
malom	mal-	(with a) small (girl)

Table 3: Translation error rates [%]
for Serbian→English

Sr→En	Develop		
	WER	PER	1-BLEU
Baseline	40.9	36.1	69.1
Stem	37.5	33.5	63.8
	Test		
	WER	PER	1-BLEU
Baseline	51.2	44.3	79.6
Stem	48.3	42.4	75.7

SMT Research, ctd.

- Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian-English Statistical Machine Translation (Popović, Vilar, Ney, 2005)
 - techniques for using monolingual data to improve SMT quality on small corpora (at the time, only had Assimil)
 - remove articles from English part of parallel data
 - train POS tagger with manually entered POS tags
 - reduce all inflected words to base forms in the Serbian vocabulary
 - negate "biti," "hteti", "imati" by adding prefix (to be, to have, to want)

Table 3: Translation error rates [%] for Serbian→English

Serbian → English		Development+Test		
Training Corpus	Method	WER	PER	1-BLEU
2.6k	baseline	45.6	39.6	70.0
2.6k	sr_base	43.5	38.2	68.9
2.6k	sr_base+v-pos	42.5	35.3	66.2
2.6k+phrases	baseline	46.0	39.6	69.5
2.6k+phrases	sr_base	44.6	39.1	70.2
2.6k+phrases	sr_base+v-pos	42.1	35.3	66.0
200	baseline	66.5	61.1	91.6
200	sr_base	63.2	58.2	90.3
200	sr_base+v-pos	63.3	56.2	88.5
200+phrases	baseline	65.2	59.5	90.2
200+phrases	sr_base	62.3	56.9	87.7
200+phrases	sr_base+v-pos	61.3	53.2	86.2

Table 6: Examples of Serbian-English translations with and without transformations

to je suvishe skupo .	⇒	to biti suvishe skup .	⇒	to SG3 biti suvishe skup .
	base forms		verb treatment	
↓ Sr → En (baseline)		↓ Sr' → En		↓ Sr'' → En
it is		it is		it is
too UNKNOWN skupo .		too expensive .		too expensive .
on ne igra .	⇒	on ne igrati .	⇒	on ne SG3 igrati .
	base forms		verb treatment	
↓ Sr → En (baseline)		↓ Sr' → En		↓ Sr'' → En
he he does not .		he do not play .		he does not play .
da , ali nemam	⇒	da , ali nemati	⇒	da , ali SG1 ne imati
mnogo vremena .	base forms	mnogo vreme .	verb treatment	mnogo vreme .
↓ Sr → En (baseline)		↓ Sr' → En		↓ Sr'' → En
yes , but I have		yes , but not		yes , but I have not got
much time .		much time .		much time .

SMT Research, ctd.

- BLEU Evaluation of Machine-Translated English-Croatian Legislation (Seljan, Vičić, Brkić, 2012)
 - survey of MT evaluation metrics by Croatian researchers
 - Includes Google Translate Bleu evaluation for Croatian baseline
 - GT is **BAD** at E - C translation

	Short sentences	Long sentences
Ref 1	0.2500	0.2009
Ref 2	0.1540	0.1539
Ref 3	0.1421	0.1498
Ref 1 & Ref 2	0.2984	0.2468
Ref 1 & Ref 2 & Ref 3	0.3186	0.2592

Table 7: BLEU scores with a single and multiple reference sets.

	Average number of errors per category					
# of sentences	Omissions	Surplus	Morphological	Lexical	Syntactic	Punctuation
100 short	0.27	0.27	1.24	0.73	0.5	0.09
100 long	0.59	0.61	3.28	1.19	1.17	0.37
200	0.43	0.44	2.26	0.96	0.84	0.23

Other Applicable Research

- Improved Statistical Machine Translation for Resource-Poor Languages Using Related Resource-Rich Languages (Nakov, NG 2009)
 - use various techniques (paraphrasing with a pivot language, transliteration, phrase table combination and augmentation) to improve translation to small-corpus languages given a large-corpus related language
 - mention that Bosnian / Croatian / Serbian can benefit from this approach with the higher availability of Croatian parallel corpora (EU)
-

References // Other Work

- <http://www.lmp.ucla.edu/Profile.aspx?LangID=46&menu=004>
- <http://www.lmp.ucla.edu/Profile.aspx?LangID=38&menu=004>
- <http://www.lmp.ucla.edu/Profile.aspx?LangID=189&menu=004>
- http://www.serbianstudies.org/publications/pdf/Vol14_2_Gorup.pdf
- <http://www.omniglot.com/language/articles/serbocroatian.htm>
- <http://www.sciencedirect.com/science/article/pii/S0024384103001323>
- http://www.eccess.eu/ECESS/Public_documents/Popovic_Stem+SuffixforSMT_LREC04.pdf
- <http://www.jstor.org/stable/pdfplus/415433.pdf?acceptTC=true>
- http://books.google.com/books?hl=en&lr=&id=_INjHgr3QioC&oi=fnd&pg=PR5&dq=serbo+croatian+linguistics&ots=fCbloQnHfL&sig=XOpolwE7eGoFbtr_x2igCfLHlmM#v=onepage
- <http://www.linguistics.ucla.edu/people/cschutze/sc2pclit.pdf>
- <http://www.ling.ohio-state.edu/~bjoseph/publications/1992balkanencyc.pdf>
- http://www.isca-speech.org/archive_open/specom_04/spc4_410.pdf
- http://delivery.acm.org/10.1145/1700000/1699682/p1358-nakov.pdf?ip=160.39.184.103&acc=OPEN&CFID=173660062&CFTOKEN=37655657&_acm_=1364477471_dbe8be6c9568dba97c9cd6911a7c4675
- <http://www ldc.upenn.edu/Catalog/byType.jsp>



Hvala

ХВАЛА