Genre-based BootCaT corpora for morphologically rich languages

Maja Miličević

University of Belgrade

BOTWU - Forlì, 24 June 2013

Overview

1. Objectives
2. Background
   - Genre and corpora
   - Cross-linguistic differences
3. Method
   - Preliminaries
   - Corpora creation procedure
4. Results
   - Details of the corpora
   - Summary of results
5. Conclusion

Objectives of the talk

Illustrate the use of BootCaT for creating corpora of Serbian:

Objectives of the talk

Illustrate the use of BootCaT for creating corpora of Serbian:

1. Main objective: Evaluate different methods for the creation of genre-based BootCaT corpora in a morphologically rich language

Objectives of the talk

Illustrate the use of BootCaT for creating corpora of Serbian:

1. Main objective: Evaluate different methods for the creation of genre-based BootCaT corpora in a morphologically rich language

   - Serbian is interesting for more than just its morphology...

Objectives of the talk

Illustrate the use of BootCaT for creating corpora of Serbian:

1. Main objective: Evaluate different methods for the creation of genre-based BootCaT corpora in a morphologically rich language

   - Serbian is interesting for more than just its morphology...

2. Additional objective: Evaluate the methods for the creation of genre-based BootCaT corpora in a bi-alphabetic writing system

Overview

Classification of corpora by genre

Classification of corpora by genre

Generally, of interest to computational and corpus linguists

$\rightarrow$ Automatic genre classification

Classification of corpora by genre

Generally, of interest to computational and corpus linguists

→ Automatic genre classification

Specifically, of interest to BootCaTters

→ Automatic creation of genre-based corpora

Classification of corpora by genre

Generally, of interest to computational and corpus linguists

→ Automatic genre classification

Specifically, of interest to BootCaTters

→ Automatic creation of genre-based corpora

A complex task, even for English!

Textual features used in genre detection

Textual features used in genre detection

Computationally complex measures

Different kinds of statistic measures on character, word and/or sentence level (function word counts, POS frequencies, character n-grams etc.)

Textual features used in genre detection

Computationally complex measures

Different kinds of statistic measures on character, word and/or sentence level (function word counts, POS frequencies, character n-grams etc.)

A simpler measure - **word n-grams**

$\rightarrow$ Sequences of n words

Textual features used in genre detection

Computationally complex measures

Different kinds of statistic measures on character, word and/or sentence level (function word counts, POS frequencies, character n-grams etc.)

A simpler measure - **word n-grams**

$\rightarrow$ Sequences of n words

Word n-grams have been shown to be good predictors of genre in English (Gries et al. 2011, Bernardini and Ferraresi 2013, Dalan 2012)

The role of inflectional morphology

English vs. morphologically richer languages

The role of inflectional morphology

English vs. morphologically richer languages

What works for English seems to work less well for Italian (cf. Bernardini and Ferraresi 2013)

→ Conjugated verbs make n-grams too "rigid"

What about Serbian?

What about Serbian?

A highly inflecting language

- 7-case declension system with 4 noun classes
- 7 verb classes conjugated for person, number, gender...

What about Serbian?

A highly inflecting language

- 7-case declension system with 4 noun classes
- 7 verb classes conjugated for person, number, gender...

Examples: *roman* 'novel', *knjiga* 'book'

	Masculine		**Feminine**	
	Sg	*Pl*	*Sg*	*Pl*
Nom	roman-ø	roman-i	knjig-a	knjig-e
Gen	roman-a	roman-a	knjig-e	knjig-a
Dat	roman-u	roman-ima	knji**z**-i	knjig-ama
Acc	roman-ø	roman-e	knjig-u	knjig-e
Voc	roman-u	roman-i	knjig-o	knjig-e
Ins	roman-om	roman-ima	knjig-om	knjig-ama
Loc	roman-u	roman-ima	knji**z**-i	knjig-ama

Genre classification in Serbian

Genre classification in Serbian

In automatic document classification byte-level n-grams have been shown to be good predictors of genre (Zečević and Utvić 2012)

Genre classification in Serbian

In automatic document classification byte-level n-grams have been shown to be good predictors of genre (Zečević and Utvić 2012)

- Problem for non-(computational) linguists: not very intuitive units
- They cannot be used in BootCaT

Genre classification in Serbian

In automatic document classification byte-level n-grams have been shown to be good predictors of genre (Zečević and Utvić 2012)

- Problem for non-(computational) linguists: not very intuitive units
- They cannot be used in BootCaT

Are word n-grams a reasonable alternative?

Genre classification in Serbian

In automatic document classification byte-level n-grams have been shown to be good predictors of genre (Zečević and Utvić 2012)

- Problem for non-(computational) linguists: not very intuitive units
- They cannot be used in BootCaT

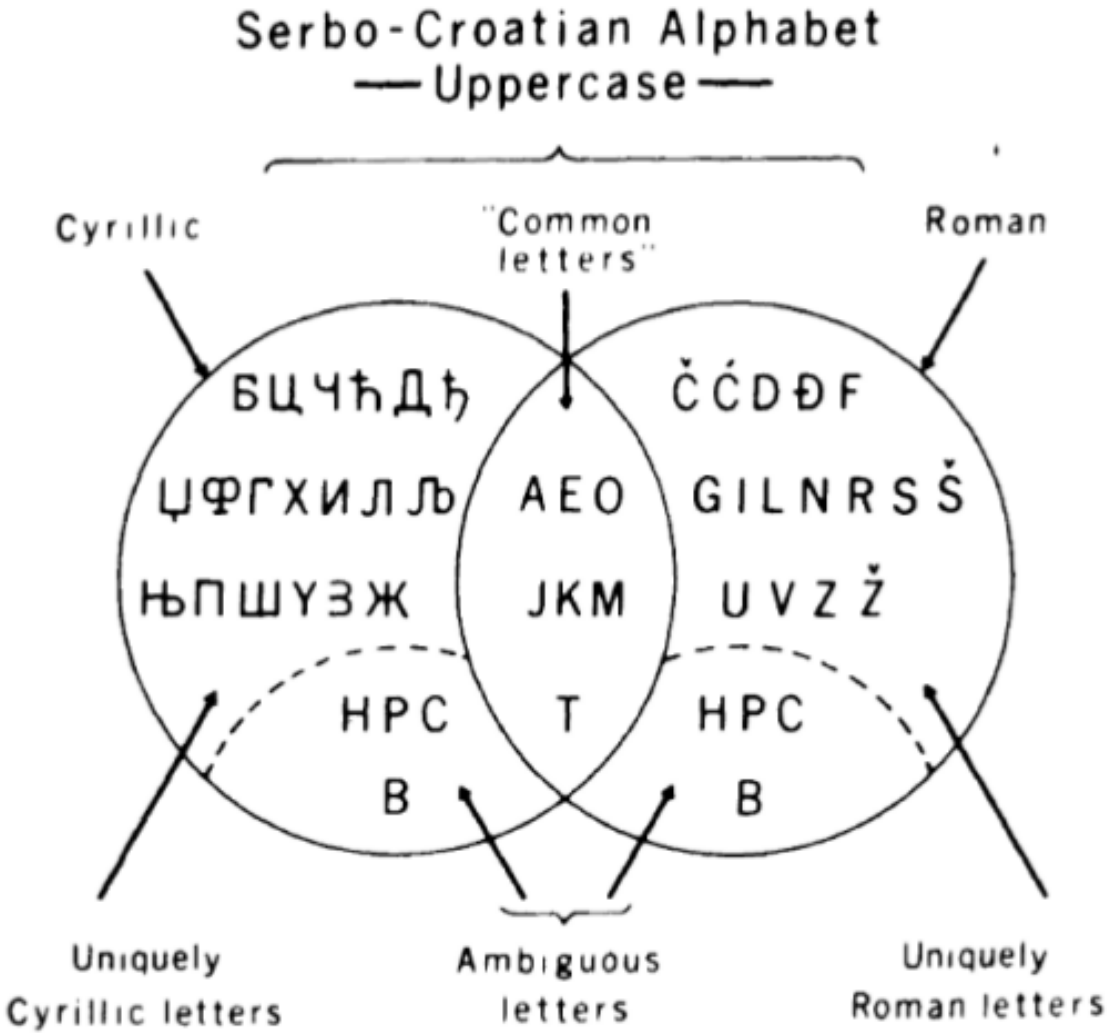
Are word n-grams a reasonable alternative?

Vitas et al. (2006) show that there is very little overlap for top frequency bigrams and trigrams between two narrative and two press corpora

→ Relevance for genre or sparse data indication?

The two alphabets of Serbian

The two alphabets of Serbian



Relevance of the two alphabets for genre

Relevance of the two alphabets for genre

Partial functional split between the Roman and the Cyrillic script

Relevance of the two alphabets for genre

Partial functional split between the Roman and the Cyrillic script

The Roman script is more present on the Internet

This needs to be kept in mind when using BootCaT for Serbian

Relevance of the two alphabets for genre

Partial functional split between the Roman and the Cyrillic script

The Roman script is more present on the Internet

This needs to be kept in mind when using BootCaT for Serbian

But, it can also be seen as an opportunity!

Relevance of the two alphabets for genre

Partial functional split between the Roman and the Cyrillic script

The Roman script is more present on the Internet

This needs to be kept in mind when using BootCaT for Serbian

But, it can also be seen as an opportunity!

Corpora of Serbian tend to be entirely in the Roman script

→ Non-transliterated web corpora could be an efficient ways for obtaining data on the use of the two alphabets for different topics or genres

Overview

The choice of genre

The choice of genre

Recipes

The choice of genre

Recipes

A highly schematic and formulaic (sub)genre

Potentially useful for translators and L2 learners

The choice of genre

Recipes

A highly schematic and formulaic (sub)genre

Potentially useful for translators and L2 learners

Closely related to a single topic (food)!

A recipe example

Retro salata od koka kole

Sastojci

- 180g želatina sa ukusom višnje (bez šećera)
- 1 šolja kipuće vode
- 300ml koka kole
- 1/2 l kompota od višanja ili konzerva sa višnjama u soku
- 1konzerva (250g ananasa) naseckanog na kockice
- 1 šolja naseckanih pekan oraha ili običnih (sirovi ili pečeni)

Korisnik: zoe1 | Broj ocena: 0 | Prosečna ocena: (0,00)

Tweet 0 Like 570 Pin it +1 0

Priprema

Stavite želatin u veliku činiju i prelijte ga kipućom vodom pa mešajte dok se ne rastopi. Umešajte koka kolu. Stavite mešavinu u frižider dok se ne zgusne, ali da ne bude potpuno čvrsta, oko 30 minuta. Proveravajte i mešajte svakih 3 do 5 minuta. Treba da se zgusne, ali ne da bude sasvim čvrsta. Stavite višnje sa sokom u blender. Uključite i isključite nekoliko puta, da se isitne. Kada se žele stegne umešajte iseckane višnje sa sokom, ananas sa sokom i orahe. Sipajte u kalup. Ostavite u frižideru da se potpuno stegne, može i preko noći. Istresite salatu iz kalupa na tanjir za posluživanje. Ako nemate kalup, možete salatu ostaviti u velikoj činji. Izvrnite je na tanjir za serviranje kad se stegne.

The choice of features to be studied

The choice of features to be studied

Previous studies looked at keywords and n-grams of different lengths

The choice of features to be studied

Previous studies looked at keywords and n-grams of different lengths

Which n-grams?

The choice of features to be studied

Previous studies looked at keywords and n-grams of different lengths

Which n-grams?

- Trigrams work well for English (Gries et al. 2011, Dalan 2012, Bernardini and Ferraresi 2013), but appear to be problematic for languages like Italian (Bernardini and Ferraresi 2013)
- Bigrams are a good choice because they capture some syntactic and semantic context, while not being too infrequent (Crossley and Louwerse 2007, Louwerse and Crossley 2006)
- Successful genre classification in the Italian corpus la Repubblica is based on non-lemmatised unigrams (Baroni et al. 2004)

The choice of features to be studied

Previous studies looked at keywords and n-grams of different lengths

Which n-grams?

- Trigrams work well for English (Gries et al. 2011, Dalan 2012, Bernardini and Ferraresi 2013), but appear to be problematic for languages like Italian (Bernardini and Ferraresi 2013)
- Bigrams are a good choice because they capture some syntactic and semantic context, while not being too infrequent (Crossley and Louwerse 2007, Louwerse and Crossley 2006)
- Successful genre classification in the Italian corpus la Repubblica is based on non-lemmatised unigrams (Baroni et al. 2004)

$\rightarrow$ **It was decided to look at all of them, separately for the Roman and Cyrillic scripts**

Seeds selection

Seeds selection

Based on a **manual/semi-automatic corpus of recipes** (200,608 words)

Seeds selection

Based on a **manual/semi-automatic corpus of recipes** (200,608 words)

The steps involved in the corpus creation included:

- Manual selection of websites containing recipes (in Roman script)
- Download of the selected pages with BootCaT (using the `site:` operator, Bernardini et al. cf. 2010)
- Manual cleaning of the corpus

Seeds selection

Based on a **manual/semi-automatic corpus of recipes** (200,608 words)

The steps involved in the corpus creation included:

- Manual selection of websites containing recipes (in Roman script)
- Download of the selected pages with BootCaT (using the `site:` operator, Bernardini et al. cf. 2010)
- Manual cleaning of the corpus

Keyword and n-grams were obtained from the manual corpus using AntConc; top **50** were used in all cases

Seeds selection

Based on a **manual/semi-automatic corpus of recipes** (200,608 words)

The steps involved in the corpus creation included:

- Manual selection of websites containing recipes (in Roman script)
- Download of the selected pages with BootCaT (using the `site:` operator, Bernardini et al. cf. 2010)
- Manual cleaning of the corpus

Keyword and n-grams were obtained from the manual corpus using AntConc; top **50** were used in all cases

Keywords were extracted based on an *ad hoc* reference corpus (1,584,920 words) containing narrative and newspaper texts*

*The largest portion of the reference corpus was prepared by D. Vitas, M. Utvić and C. Krstev, with the help of students in the Department of Information Science, University of Belgrade

Tuple creation

Tuples of the following lengths were created:

- For keywords and unigrams n=**5**
- For bigrams n=**4**
- For trigrams n=**3**

Tuple creation

Tuples of the following lengths were created:

- For keywords and unigrams n=**5**
- For bigrams n=**4**
- For trigrams n=**3**

The Cyrillic versions of the corpora were based on the same tuples (transliterated using `preslovljavanje.com`)

Tuple creation

Tuples of the following lengths were created:

- For keywords and unigrams n=**5**
- For bigrams n=**4**
- For trigrams n=**3**

The Cyrillic versions of the corpora were based on the same tuples (transliterated using `preslovljavanje.com`)

There was no manual editing in the phases of seeds selection and tuple creation

Examples of keyword and unigram tuples

Examples of keyword and unigram tuples

Keywords

1 pecite kašika fil stavite
stavite 2 sitno rerni šećera
umutiti meso kašika ulja pomešajte
jaja fil rerni kašika ostaviti
minuta mleka ulje šećera kuvajte

Examples of keyword and unigram tuples

Keywords

1 pecite kašika fil stavite
stavite 2 sitno rerni šećera
umutiti meso kašika ulja pomešajte
jaja fil rerni kašika ostaviti
minuta mleka ulje šećera kuvajte

Unigrams

ga kašike ili još ostavite
od ulja oko umutiti s
stavite preko od vode ne
priprema staviti umutiti pa za
od priprema i u kuvajte

Examples of bigram and trigram tuples

Bigrams

“i na” “ostaviti da” “i pecite” “se ohladi”
“i kuvajte” “15 minuta” “g šećera” “da se”
“na kraju” “u prahu” “belog luka” “za pečenje”
“minuta na” “kada se” “100 g” “30 minuta”
“100 g” “u pleh” “na pari” “i ostavite”

Examples of bigram and trigram tuples

Bigrams

“i na” “ostaviti da” “i pecite” “se ohladi”
“i kuvajte” “15 minuta” “g šećera” “da se”
“na kraju” “u prahu” “belog luka” “za pečenje”
“minuta na” “kada se” “100 g” “30 minuta”
“100 g” “u pleh” “na pari” “i ostavite”

Trigrams

“posolite i pobiberite” “u podmazan pleh” “čokolade za kuvanje”
“papirom za pečenje” “u zagrejanoj rerni” “i beli luk”
“da se ne” “rerni zagrejanoj na” “u frižider na”
“u frižider na” “u podmazan pleh” “oko 20 minuta”
“minuta u rerni” “rerni zagrejanoj na” “na sobnoj temperaturi”

Creation and evaluation of corpora

Creation and evaluation of corpora

20 tuples were created for each seed list

20 pages per tuple were downloaded

No Internet domain limitations were imposed

Creation and evaluation of corpora

20 tuples were created for each seed list

20 pages per tuple were downloaded

No Internet domain limitations were imposed

Identical procedures were followed for the two scripts

→ **A total of 8 corpora were created**

Creation and evaluation of corpora

20 tuples were created for each seed list

20 pages per tuple were downloaded

No Internet domain limitations were imposed

Identical procedures were followed for the two scripts

→ **A total of 8 corpora were created**

Precision-based evaluation
→ A sample of 50 texts from each corpus was manually checked to establish whether it belonged to the target genre

Overview

Size and relevance of obtained corpora

Size and relevance of obtained corpora

	Roman alphabet			
	Keywords	*Unigrams*	*Bigrams*	*Trigrams*
N URLs	314 (338)	339 (367)	324 (343)	321 (363)
N words	167,605	204,266	191,521	265,382
Relevant URLs	90%	72%	90%	88%

Size and relevance of obtained corpora

	Roman alphabet			
	Keywords	*Unigrams*	*Bigrams*	*Trigrams*
N URLs	314 (338)	339 (367)	324 (343)	321 (363)
N words	167,605	204,266	191,521	265,382
Relevant URLs	90%	72%	90%	88%

	Cyrillic alphabet			
	Keywords	*Unigrams*	*Bigrams*	*Trigrams*
N URLs	234 (239)	308 (314)	88 (90)	51 (51)
N words	161,290	216,006	524,389	395,366
Relevant URLs	68%	50%	80%	92%

Findings on genre

The evaluation procedure has shown that:

Findings on genre

The evaluation procedure has shown that:

- Keywords are a reliable method for the Roman script, but not for the Cyrillic one

Findings on genre

The evaluation procedure has shown that:

- Keywords are a reliable method for the Roman script, but not for the Cyrillic one
- Unigrams give the lowest precision rate in both scripts

Findings on genre

The evaluation procedure has shown that:

- Keywords are a reliable method for the Roman script, but not for the Cyrillic one
- Unigrams give the lowest precision rate in both scripts
- Bigrams and trigrams work well in both scripts - regardless of the morphological complexity of Serbian

Differences between the two scripts

Some differences between the Roman and Cyrillic scripts were found:

Differences between the two scripts

Some differences between the Roman and Cyrillic scripts were found:

- More even-sized texts obtained for the Roman script; these texts tend to be from websites dedicated to recipes

Differences between the two scripts

Some differences between the Roman and Cyrillic scripts were found:

- More even-sized texts obtained for the Roman script; these texts tend to be from websites dedicated to recipes
- Substantial variation found for the Cyrillic script; several entire books are downloaded (mostly from `scribd.org`) $\rightarrow$ larger corpora; in general, the texts in Cyrillic seem to be more "narrative"

Overview

Genre-based corpora and inflectional morphology

Genre-based corpora and inflectional morphology

The study did not show a major advantage of n-grams over keywords for recipes corpora in Serbian, at least for the Roman script (cf. English)

Genre-based corpora and inflectional morphology

The study did not show a major advantage of n-grams over keywords for recipes corpora in Serbian, at least for the Roman script (cf. English)

Bigrams and trigrams gave good results (in both scripts) despite the rich morphology of Serbian

Genre-based corpora and inflectional morphology

The study did not show a major advantage of n-grams over keywords for recipes corpora in Serbian, at least for the Roman script (cf. English)

Bigrams and trigrams gave good results (in both scripts) despite the rich morphology of Serbian

In addition to the morphological properties of the language, the nature of the genre under study appears to be very important

Genre-based corpora and inflectional morphology

The study did not show a major advantage of n-grams over keywords for recipes corpora in Serbian, at least for the Roman script (cf. English)

Bigrams and trigrams gave good results (in both scripts) despite the rich morphology of Serbian

In addition to the morphological properties of the language, the nature of the genre under study appears to be very important

$\rightarrow$ **Different procedures might be needed for different genres**

Ideas for follow up analyses and future work

Ideas for follow up analyses and future work

Several possible lines of development:

Ideas for follow up analyses and future work

Several possible lines of development:

- Refining the evaluation procedure
  - Counting the number of words in the relevant portion of the corpus rather than the number of relevant URLs/texts?
  - Have a closer look at topic variety among the recipes that were extracted - e.g. food vs. beverages /medications?

Ideas for follow up analyses and future work

Several possible lines of development:

- Refining the evaluation procedure
  - Counting the number of words in the relevant portion of the corpus rather than the number of relevant URLs/texts?
  - Have a closer look at topic variety among the recipes that were extracted - e.g. food vs. beverages /medications?
- Complementing the analysis
  - Repeat the procedure for a less conventional genre and/or a genre with greater variation in topic
  - Repeat the procedure for another morphologically rich language, e.g. Italian

Thank you!

m.milicevic@fil.bg.ac.rs

References

Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, and M. Mazzoleni (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. *Proceedings of LREC 2004*.

Bernardini, S. and A. Ferraresi (2013). Old needs, new solutions. Comparable corpora for language professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum, and P. Fung (Eds.), *BUCC Building and Using Comparable Corpora*. Dordrecht: Springer.

Bernardini, S., A. Ferraresi, and F. Gaspari (2010). Institutional academic English in the European context: a web-as-corpus approach to comparing native and non-native language. In A. L. López and C. J. Rosalía (Eds.), *Professional English in the European Context: The EHEA Challenge*, pp. 27–53. Bern: Peter Lang.

Crossley, S. A. and M. M. Louwerse (2007). Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics 12*, 453–478.

Dalan, E. (2012). Costruzione automatica di corpora orientati al genere e fraseologia: Il caso delle guide web in inglese degli Atenei europei. MA thesis.

Gries, S. T., J. Newman, and C. Shaoul (2011). N-grams and the clustering of registers. *Empirical Language Research Journal 5*.

Louwerse, M. M. and S. A. Crossley (2006). Dialog act classification using n-gram algorithms. In G. Sutcliffe and R. Goebel (Eds.), *Proceedings of the 19th International Florida Artificial Intelligence Research Society (FLAIRS)*, pp. 758–763. Menlo Park, CA: AAAI Press.

Vitas, D., G. Pavlović-Lažetić, and C. Krstev (2006). About word length counting in Serbian. In P. Gryzbek (Ed.), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*, pp. 301–317. Dordrecht: Springer.

Zečević, A. and M. Utvić (2012). An authorship attribution for Serbian. In Z. Budimac, M. Ivanović, and M. Radovanović (Eds.), *BCI-LOCAL 2012. Local Papers of the Balkan Conference in Informatics*, pp. 109–112. Novi Sad: University of Novi Sad.