Comparing named entities in different ELTeC collections

Diana Santos

Belgrade training school, 23 March 2022

1 Getting to know the collections

We have three different collections with named entities in level 2. In the dataframes that we are going to read, for each work we have how many named entities in total, and also for each kind, as you can check by running summary on them.

```
ne_srp<-read.table("tableSerbian.R", header=TRUE)
summary(ne_srp)
ne_slv<-read.table("tableSlovenian.R", header=TRUE)
summary(ne_slv)
ne_por<-read.table("tablePortuguese.R",header=TRUE)
summary(ne_por)
```

Note: according to Tomaž Erjavec, who created level 2 for Slovenian and added named entity annotation, for comparison purposes it is better not to use the category `deriv.per`, so we start by adjusting the Slovenian NE dataframe:

```
ne_slv$NE<-ne_slv$NE-ne_slv$deriv.per
```

We have then for each level 2 collection the metadata, which we are going to read for all in a loop

```
list <- c("por","nor","slv","srp","deu","hun","fra","eng")
for (i in list) {
  filename <- paste0("col_", i)
  wd <- paste0("meta", i, ".R")
  assign(filename, read.table(wd,header=TRUE))
}
summary(col_eng) # to test if they were read successfully
```

Because we will soon also use proper names, let us do the same for yet another set of files including proper names per work.

```
for (i in list) {
  filename <- paste0("prop_", i)
  wd <- paste0("proprios", i, ".R")
  assign(filename, read.table(wd,header=TRUE))
}
summary(prop_fra)
```

We are now going to some dataframe manipulation, so that we join/merge the information in the several tables. For example, we add the meta-information to the dataframe which includes named entities, because we want more detailed information:

```
slovenian<-merge(ne_slv,col_slv,by.x=c("Ficheiro"),by.y=c("file"))
portuguese<-merge(ne_por,col_por,by.x=c("Ficheiro"),by.y=c("file"))
serbian<-merge(ne_srp,col_srp,by.x=c("Ficheiro"),by.y=c("file"))
```

The reason for using by.x and by.y is that we want that the two columns be considered the same, although having originally different names.

Given that for Serbian so far only 65 novels have been annotated with named entities, we select that subset first, calling it arbitrarily `serbian2`:

```
serbian2<-subset(serbian,PERS!=0)
```

Then we want to compare data across languages, and across differen works. So we create a new column for relative number of named entities per number of words, called `relne`.

```
slovenian$relne<-slovenian$NE/slovenian$words
portuguese$relne<-portuguese$NE/portuguese$words
serbian2$relne<-serbian2$NE/serbian2$words
```

This allows us to see the distribution of NEs depending on metadata parameters. Try

```
boxplot(slovenian$relne~slovenian$canon)
boxplot(slovenian$relne~slovenian$gender)
boxplot(slovenian$relne~slovenian$size)
boxplot(slovenian$relne~slovenian$period)
```

See e.g. the distributions per period in the three languages, in Figure 1.

This may be confusing, because the three pictures have different y-axes. A better way of visualizing would require the same y-axis.

```
boxplot(portuguese$relne~portuguese$period,ylim=c(0,0.1), \\
  ylab="Number of NE per number of words",xlab="20-year period")
boxplot(slovenian$relne~slovenian$period,ylim=c(0,0.1), \\
  ylab="Number of NE per number of words",xlab="20-year period")
boxplot(serbian2$relne~serbian2$period,ylim=c(0,0.1), \\
  ylab="Number of NE per number of words",xlab="20-year period")
```

Figure 2 provides the new drawings.

2 Comparing proper names in 8 collections

Let us now see if we can widen our investigation by looking not at named entities, but proper names.

We start by doing the corresponding merges. With the next commands, we had meta information to the dataframe with proper names. Since both columns representing the file are called `file`, it is enough to specify `by` (and not `by.x` and `by.y`:

```
french<-merge(prop_fra,col_fra,by=c("file"))
french$relprop<-french$proper/french$words
boxplot(french$relprop~french$period)
```

(Do this for all collections.)

But why not add all files?

```
all<-rbind(hungarian,english,german,norwegian,french,portuguese3, \\
  serbian3,slovenian3)
all_clean<-all[all$file!="file",]
all_clean$lang <-as.factor(sub( "^([A-Z][A-Z][A-Z]*).*", "\\1", \\
  all_clean$file, perl=TRUE))
summary(all_clean)
```

Now we can observe the similarites or differences across all collections:

```
boxplot(all_clean$relprop~all_clean$lang)
boxplot(all_clean$relprop~all_clean$gender+all_clean$lang)
```

3 Check correlation between proper names and named entities

But is the number of proper names in a work a good proxy for the number os named entities?

There are two things that seem to go against this hypothesis:

- We (the NE-subgroup) suggested that professions and demonyms should also be considered "named entities". Or better, part of a semantic light annotation of literary texts. SO, the Serbian and Portuguese collections, which followed this suggestion, should have more named entities than proper names (all those professions and demonyms counted as "named entities" and were not proper names)

- There is some difference in how expressions with more than one word are encoded in the different collections. More specifically, some collections assign the proper name classification to only the words that are separately a proper name, see French:

  ```
  <w pos=’PROPN’ lemma=’GeneviÃ¨ve’>GeneviÃ¨ve</w>
  <w pos=’ADP’ lemma=’de’>de</w>
  <w pos=’PROPN’ lemma=’Brabant’>Brabant</w>
  ```

  while others assign all the members of the expression the category proper name, see Portuguese:

  ```
  <w pos="PROPN" lemma="Henrique_de_Souzellas"  msd="M S">Henrique</w>
  <w pos="PROPN" lemma="Henrique_de_Souzellas" msd="M S">de</w>
  <w pos="PROPN" lemma="Henrique_de_Souzellas" msd="M S">Souzellas</w>
  ```

  This will give the second kind of languages (possibly only Portuguese and Norwegian) a surplus of proper names.

Let us therefore check whether we can use proper names as a proxy for named entities, with a correlation analysis in the three languages which have both proper names and named entities:

```
allport <- merge(portuguese,portuguese3,by.x=c("Ficheiro","words", \\
  "date","gender","size","canon","period"),by.y=c("file","words",\\
  "date","gender","size","canon","period"))
plot(allport$relne,allport$relprop)
cor(allport$relne,allport$relprop)
  [1] 0.9251557
```

You should repeat this for the other two languages (remember to only use the part of Serbian collection which has named entities).

I include here the three plots in Figure 3, that show that Slovenian is indeed the language where proper nouns and named entities have stronger correlation (0.9927722). Serbian has the least correlation (0.8999835), most probably because, in addition to demonyms and professions, it does not compensate with a surplus of proper noun tags, as Portuguese.

```
par(mfrow=c(1,3))
plot(allport$relne,allport$relprop, main="Portuguese")
plot(allslov$relne,allslov$relprop, main="Slovenian")
plot(allserb$relne,allserb$relprop, main="Serbian")
```

In any case, all correlations are high enough to justify using proper noun counts.

4 Further comparing of named entities

But what about comparing different types of named entities?

In order to merge the data about the three collections, we have to make sure that they contain the same information. Since Slovenian has a different set of NE categories from Serban and Portuguese, we have to reduce all dataframes to those which are common, namely people, places and organizations.

```
unifport<-subset(allport,TRUE,c(1:10, 15:18))
unifsrp<-subset(allsrp,TRUE,c(1:10, 15:18))
colnames(allslv)[8]<-"PERS"
colnames(allslv)[9]<-"PLACE"
colnames(allslv)[10]<-"ORG"
unifslv<-subset(allslv,TRUE,c(1:10, 13:16))
allne<-rbind(unifport,unifslv,unifsrp)
allne$lang <-as.factor(sub( "^([A-Z][A-Z][A-Z]*).*", "\\1", \\
  allne$Ficheiro, perl=TRUE))
summary(allne)
```

We can then compare the three languages as far as these three kinds of named entities are concerned, and especially if we create columns displayng the relative number of persons (palces, organisations) per work.

```
boxplot(allne$NE~allne$lang)
allne$relPERS<-allne$PERS/allne$words
allne$relPLACE<-allne$PLACE/allne$words
allne$relORG<-allne$ORG/allne$words
boxplot(allne$relPERS~allne$lingua)
boxplot(allne$relPERS~allne$lingua+allne$canon)
boxplot(allne$relPERS~allne$lingua+allne$period+allne$gender, las=2)
```

And all other comparisons that one might be interested in doing.

5 Napoleon in European literature

Get the dataframe napoleon.R and observe it.

```
napoleon <- read.table("napoleon.R", header=TRUE)
summary(napoleon)
```

Add language information based on the name of the files.

```
napoleon$lang <-as.factor(sub( "^([A-Z][A-Z][A-Z]*).*", "\\1", \\
  napoleon$file, perl=TRUE))
summary(napoleon)
```

Compute the total number of occurrences, per language, in an array called `addition`$_n$

```
addition_n<-vector()
list <- c("POR","NOR","SLV","SRP","DEU","HU","FRA","ENG")
for (i in list) {
  addition_n[i] <- sum(napoleon[napoleon$lingua==i,]$palavra)
}
barplot(addition_n, main="Napoleon in 8 European literatures: amount of referenc
#pie(addition_n)
```

To compute the number of works that contain reference to Napoleon, simply count all files with number of words (in column `palavra`) different from 0.

```
barplot(table(napoleon[napoleon$palavra!=0,]$lingua), main="Napoleon \\
  in 8 European literatures: number of works citing him")
```

6 Christmas in European literature

Follow the exact same procedure for Christmas.

```
christmas <- read.table("christmas.R", header=TRUE)
summary(christmas)
christmas$lang <-as.factor(sub( "^([A-Z][A-Z][A-Z]*).*", "\\1", \\
  christmas$file, perl=TRUE))
summary(christmas)
addition_c<-vector()
list <- c("POR","NOR","SLV","SRP","DEU","HU","FRA","ENG")
for (i in list) {
  addition_c[i] <- sum(christmas[christmas$lingua==i,]$palavra)
}
```

```
barplot(addition_c, main="Christmas in 8 European literatures: amount \\
  of references")
#pie(addition_c)
barplot(table(christmas[christmas$palavra!=0,]$lingua), main="Christmas \\
  in 8 European literatures: number of works citing it")
```

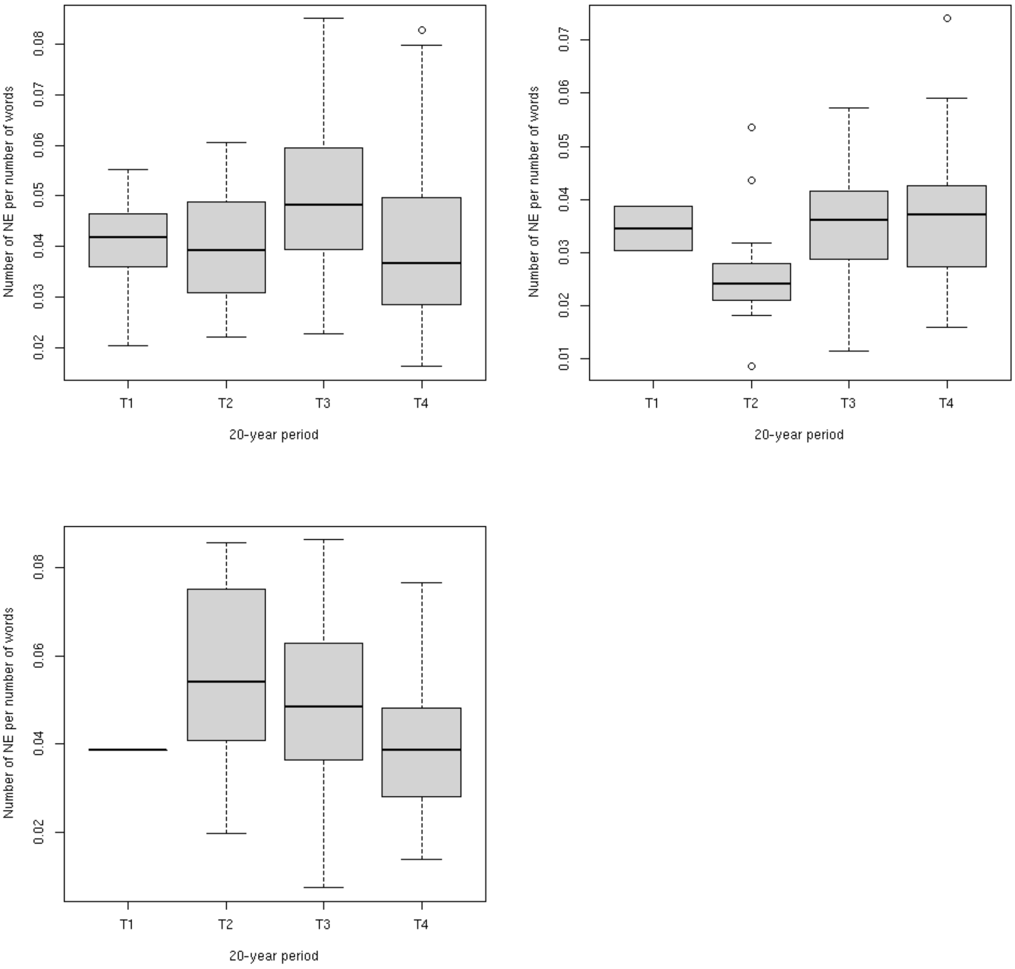

Figure 1: The distribution in Portuguese, Slovenian and Serbian per period, independently

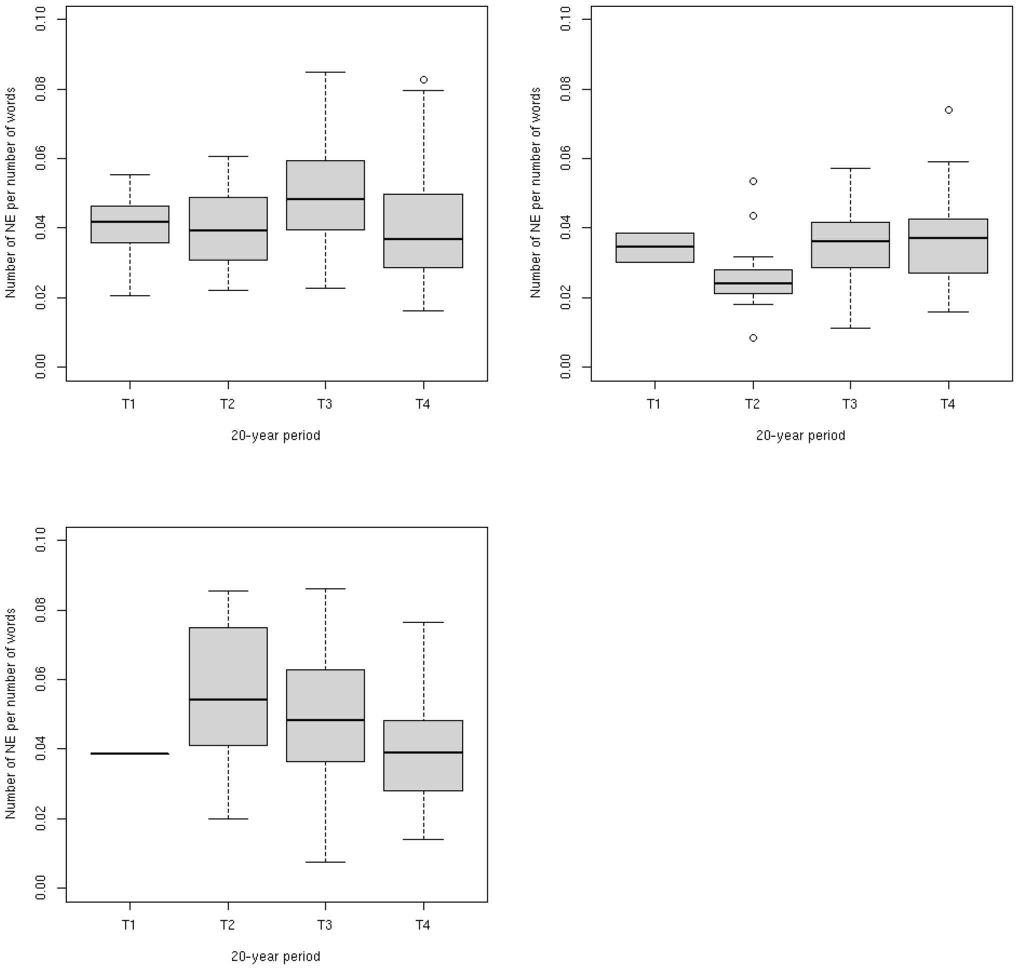

Figure 2: The distribution in Portuguese, Slovenian and Serbian per period, with the same y-axis

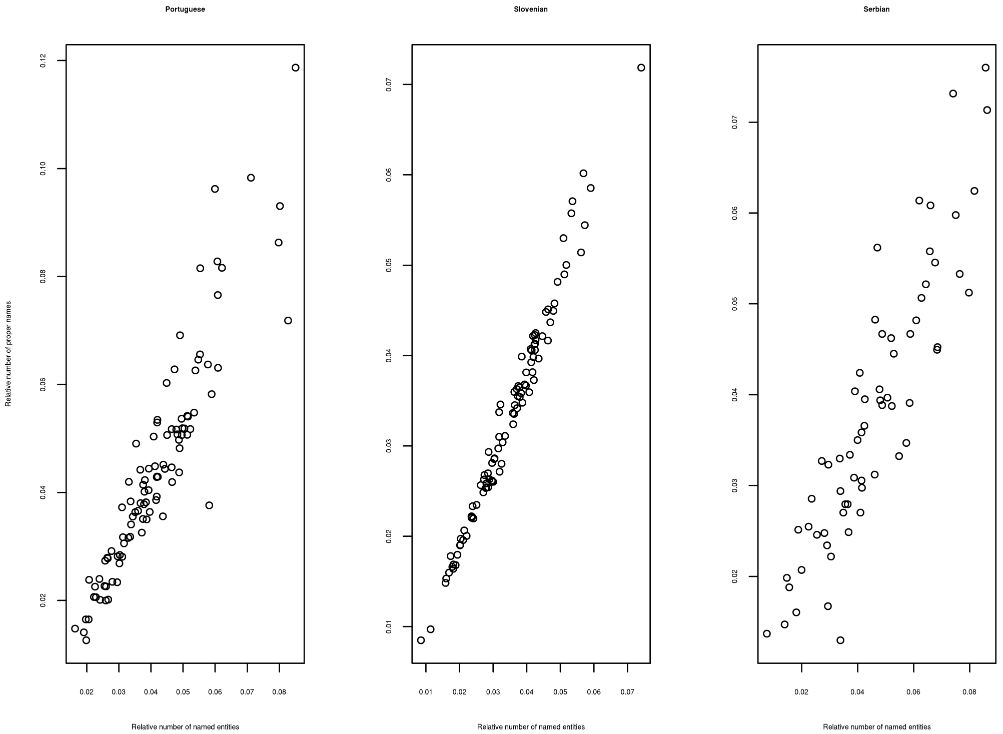

Figure 3: Comparison for the three languages