Subtitles as a Corpus

Natalia Levshina
F.R.S. – FRNS, Université catholique de Louvain
Belgium

Outline

1. What are subtitles?
2. Are subtitles ‘good’ data for linguists?
3. Let’s ParTy!

Subtitles today

- A type of audiovisual translation (dubbing, surtitling, voice-over, etc.)
- Films, TED talks, TV, games, clips, etc.
- Original (e.g. for the hearing impaired) and translations, sometimes in more than one language (e.g. Belgium)
- Spatiotemporal restrictions: how many lines and characters, how long the duration, etc. due to low reading speed and multitasking pressure

Linguistic features

- Speech reduction (up to 40%, e.g. Woody Allen's films)
  - elimination of what is irrelevant
  - reformulation of what is relevant in a concise form
- Depends on the genre and context (relevance)
- Verbal phrases:

  be going to do smth → 'will do smth'

  have smth fixed → 'fix smth'

  have a feeling → 'feel'

  have a lot of money → 'be rich'

  Let me help you… → 'I'll help you'

Outline

1. What are subtitles?
2. Are subtitles ‘good’ data for linguists?
3. Let’s ParTy!

Can we rely on subtitles?

- It’s an empirical question.
- Let’s compare them with different ‘natural’ registers of spoken and written language and see how and to what extent subtitles are different.
- Today, the focus is on English…

Registers

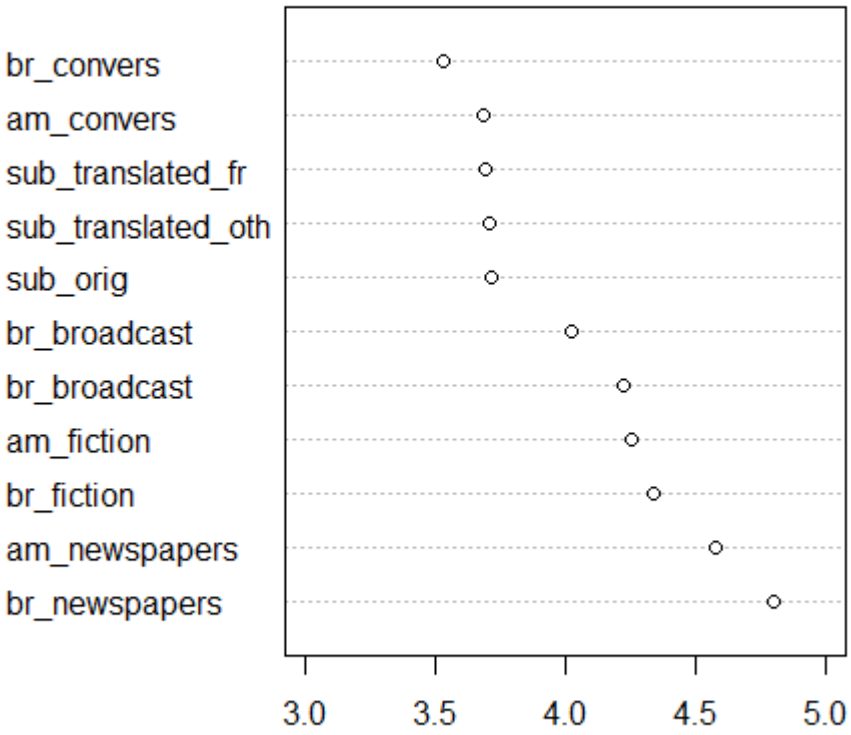
- Subtitles of films originally in English (different genres, according to IMDB)
- Subtitles of films originally in French
- Subtitles of films originally in other languages (14 languages)
- AmE: informal conversations (Santa Barbara corpus of spoken English)
- AmE: transcripts of broadcast TV and radio programmes (COCA)
- AmE: newspapers
- AmE: fiction
- BrE: informal conversations (BNC)
- BrE: transcripts of broadcast TV and radio programmes (BNC)
- BrE: national newspapers
- BrE: fiction

Total: > 2M words

Methods

- average word lengths
- correlations between wordform frequencies in each register
- Principal Component Analysis (similar to Biber's multidimensional analysis)

Average word lengths



The words in subtitles are on average NOT shorter than words in informal conversations!

Correlations

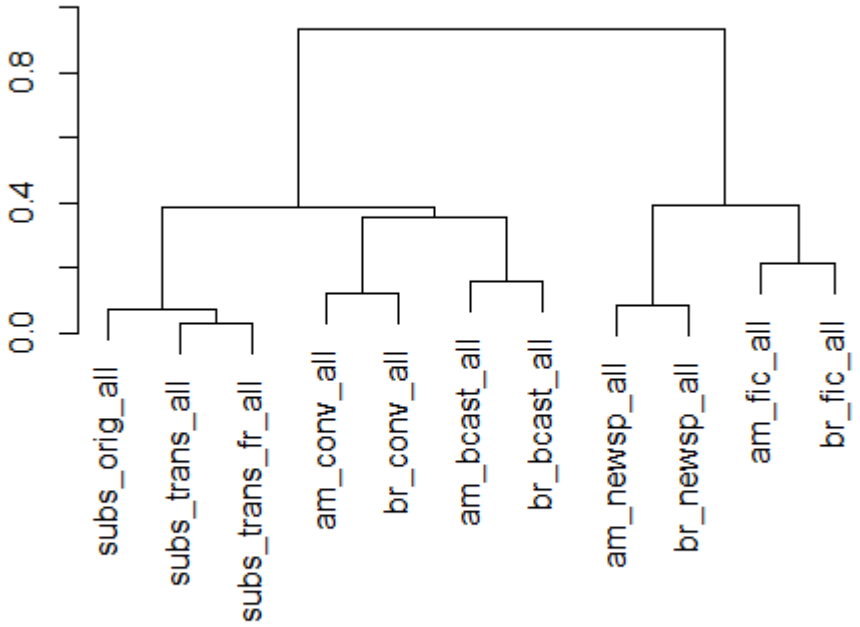
- The subtitles in original English are the most similar to the translated subtitles, of all registers.
- From all other varieties (not subtitles), the original English subtitles are the most similar to American and British informal conversations.

  → films subtitles are a special register of spoken language
- The subtitles translated from French have an almost perfect correlation with the subtitles translated from other languages.

  → lack of systematic evidence of translationese
- No substantial differences between American and British English is observed, although some lexical and grammatical markers (e.g. colour vs. color) suggest that the subtitles are mostly in American English.

  → one can neglect the geographic differences (at least, in grammar)

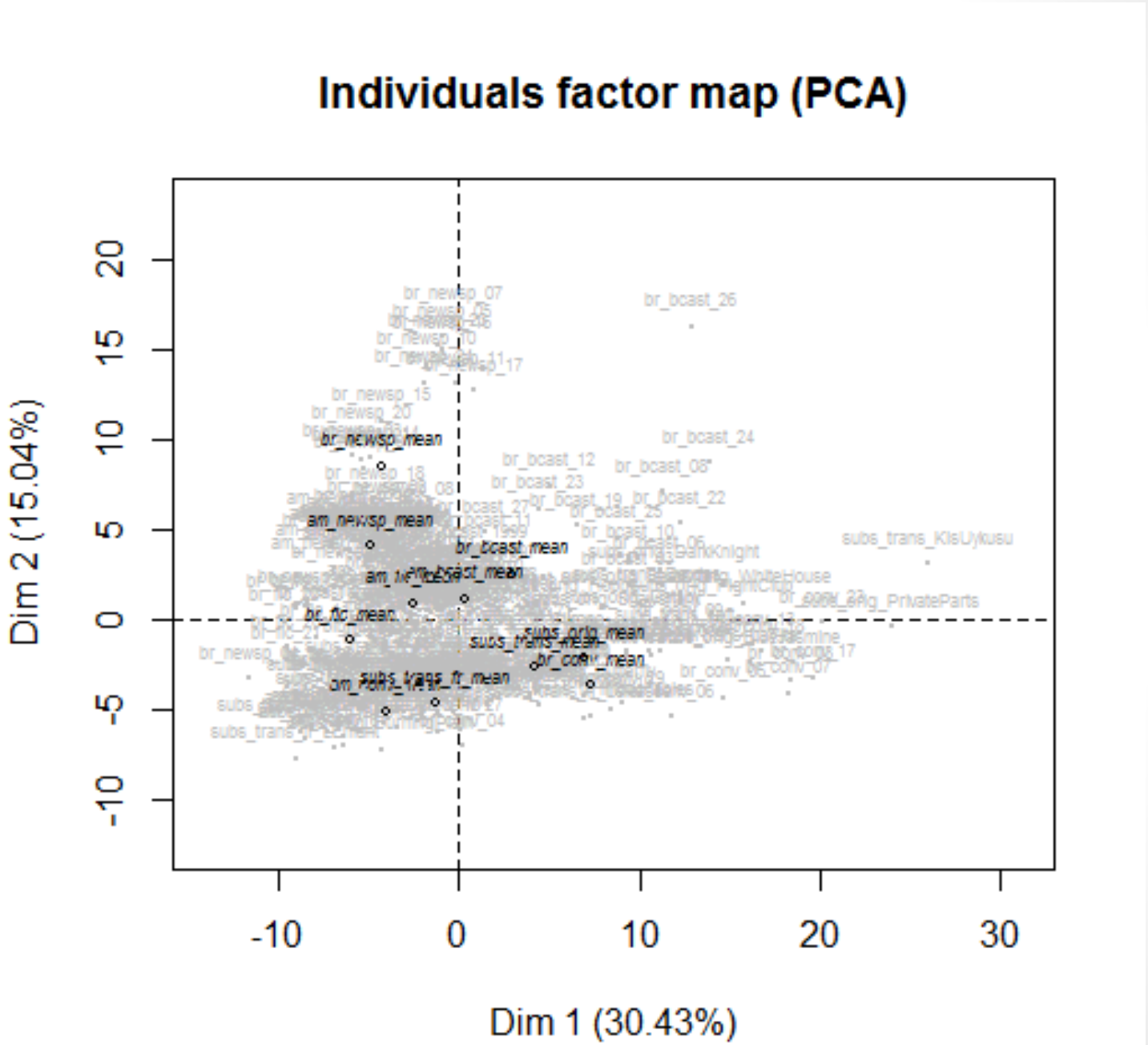
A clustering based on correlations

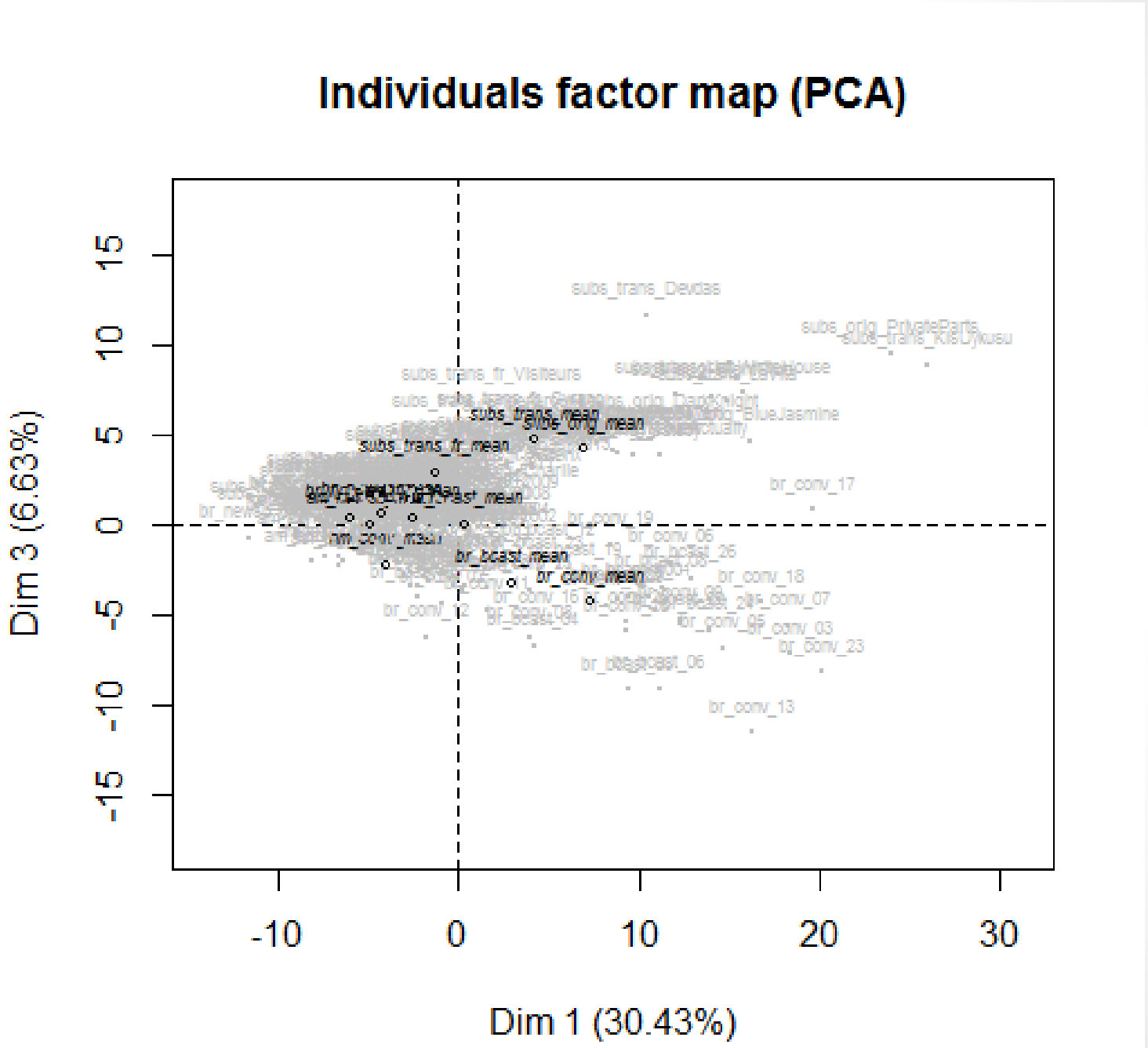


Based on 150 most frequent word forms.
The same result is observed if one takes all word forms.

Principle Component Analysis

- Based on 150 most frequent wordforms (the, my, do, can, not, ’m, that, etc.) – mostly function words.
- Similar to factor analysis, classical method for multidimensional register analysis, as in Biber (1988).
- PCA allows one to see which wordforms are responsible for similarities and differences between the registers.


Individuals factor map (PCA)
Dim 2 (15.04%)
Dim 1 (30.43%)
-10
-5
0
5
10
15
20
-10
0
10
20
30
br_newsp_mean
am_newsp_mean
br_bcast_mean
br_tlc_mean
subs_orig_mean
br_conv_mean
br_bcast_26
br_bcast_24
br_bcast_12
br_bcast_08
br_bcast_23
br_bcast_22
br_bcast_10
subs_trans_KlsUykusu
br_conv_PrivateParts


Individuals factor map (PCA)
Dim 3 (6.63%)
Dim 1 (30.43%)
15
10
5
0
-5
-10
-15
-10
0
10
20
30
subs_trans_Devdas
subs_orig_PrivateParts
subs_trans_fr_Visiteurs
subs_trans_mean
subs_orig_mean
subs_trans_fr_mean
br_conv_17
br_conv_19
br_conv_05
br_bcast_mean
br_conv_mean
br_conv_18
br_conv_07
br_conv_03
br_conv_23
br_bcast_06
br_conv_13

Dimension 2

Positive: of, in, the, as, by, from, an, which, has…

Negative: you, n’t, ’m, do, go, why, did, me, oh, your, here…

Interpretation: information vs. involved interaction (cf. Biber 1988)

Dimension 3

- Positive: let, my, me, your, man, him, here, why, too, ’m…
- Negative: erm, ’ve, well, they, er, mm, then, put, got, yeah…

- Interpretation: subtitles imitate spontaneous speech, but not exactly. Pause fillers and hesitation markers are frequently omitted, maybe because they do not require translation (similar to gestures and body language)? Another explanation is that these markers may not be used in films as often as in everyday language(?)

Conclusions

1. Subtitles are a special register of 'spoken' language, which is very close to real informal conversations.
2. The main difference between the registers lies in the presence or absence of pause fillers and hesitation markers.

   → If you study these phenomena, subtitles are probably not the best choice!

3. No systematic differences between American and British English is detected.

   → If you study regional variation, that could be a problem!

4. Translationese effects are not observed at large. However, a finer-grained inspection of a specific construction might detect some biases.

   → It's an empirical question for every specific phenomenon one wants to study.

Outline

1. What are subtitles?
2. Are subtitles ‘good’ data for linguists?
3. Let’s ParTy!

ParTy corpus

- Collected for cross-linguistic comparisons of more than two languages
- A small fraction (for test purposes) is available at www.natalialevshina.com
- Film subtitles + TED subtitles (in the future)
- More fun than existing massively parallel corpora (the Bible, EU law, Europarl), more languages

Sources

- Online: opensubtitles.org, podnapisi.net, subscene.com
- .srt format: time and text + formatting (optionally)
- Time precision: one millisecond
- Example of a caption:

```
2
00:03:55,439 --> 00:03:57,336
I had the craziest dream last night.
```

A QUIZ!

Extract 1

17
00:07:51,356 --> 00:07:54,567
I need your clothes, your boots,
and your motorcycle.

18
00:08:03,868 --> 00:08:06,162
You forgot to say 'please'.

19
00:09:06,973 --> 00:09:08,099
Take it.

20
00:09:34,751 --> 00:09:37,628
You can't leave and take
the man's wheels, son.

Extract 2

20
00:02:39,232 --> 00:02:42,715
Your brother represented
a significant investment.

21
00:02:42,750 --> 00:02:45,749
We'd like to talk to you
about taking over his contract.

22
00:02:45,784 --> 00:02:48,348
And since your genome is identical to his,

23
00:02:48,383 --> 00:02:50,946
you could step into his shoes.

24
00:02:51,546 --> 00:02:53,745
So to speak.

Extract 3

1789
02:28:40,320 --> 02:28:42,450
I mean, if I had my way,

1790
02:28:42,950 --> 02:28:47,370
you'd wear that goddamn uniform
for the rest of your pecker-sucking life.

1791
02:28:48,870 --> 02:28:49,920
But I'm aware that ain't practical.

1792
02:28:50,000 --> 02:28:53,040
I mean, at some point,
you're going to have to take it off.

1793
02:28:53,170 --> 02:28:54,210
So,

1794
02:28:55,170 --> 02:28:58,970
I'm going to give you
a little something you can't take off.

1795
02:29:23,820 --> 02:29:25,950
You know something, Utivich?

1796
02:29:26,330 --> 02:29:29,590
I think this just might be my masterpiece.

Extract 4

2
00:03:55,439 --> 00:03:57,336
I had the craziest dream last night.

3
00:03:59,779 --> 00:04:01,686
I was dancing the White Swan.

4
00:04:03,444 --> 00:04:06,650
It was different choreography, though. It was more like the Bolshoi's.

5
00:04:10,664 --> 00:04:12,291
It was the prologue,

6
00:04:12,326 --> 00:04:15,113
when Rothbart casts his spell.

Extract 5

1
00:00:51,084 --> 00:00:55,612
On September 3, 1973...

2
00:00:55,655 --> 00:01:00,820
a blue fly capable of flapping
70 beats a minute...

3
00:01:00,860 --> 00:01:03,886
landed on St. Vincent Street
in Montmartre.

4
00:01:07,734 --> 00:01:11,864
At that moment, on
a restaurant terrace nearby...

5
00:01:11,905 --> 00:01:14,271
the wind magically made
two glasses dance unseen...

6
00:01:14,307 --> 00:01:18,073
on a tablecloth.