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.
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!
I need your clothes, your boots, and your motorcycle.

You forgot to say 'please'.

Take it.

You can't leave and take the man's wheels, son.
Your brother represented a significant investment.

We'd like to talk to you about taking over his contract.

And since your genome is identical to his, you could step into his shoes.

So to speak.
I mean, if I had my way, you'd wear that goddamn uniform for the rest of your pecker-sucking life. But I'm aware that ain't practical. I mean, at some point, you're going to have to take it off. So, I'm going to give you a little something you can't take off. You know something, Utivich? I think this just might be my masterpiece.
I had the craziest dream last night.

I was dancing the White Swan.

It was different choreography, though. It was more like the Bolshoi's.

It was the prologue,

when Rothbart casts his spell.
On September 3, 1973...

a blue fly capable of flapping 70 beats a minute...

landed on St. Vincent Street in Montmartre.

At that moment, on a restaurant terrace nearby...

the wind magically made two glasses dance unseen...

on a tablecloth.