UNI FR

UNIVERSITÄT BERN

Using parallel corpora

Sandrine Zufferey

COST Training School

January 2016

The example of Europarl

- Corpus with the minutes of **debates** at the **European Parliament**.
  - each deputy **speaks in their mother tongue**
  - each **statement** is **transcribed** and **translated** into the other official languages
- The **number** of **official languages increased** over the years:
  - 1958: 4 languages (Dutch, French, German and Italian)
  - 1973: + 2 (Danish and English)
  - 1981: + 1 (Greek)
  - 1986: + 2 (Portuguese and Spanish)
  - 1995: + 2 (Finnish and Swedish)
  - 2004: + 9 (Czech, Estonian, Hungarian, Latvian, Lithuanian, Maltese, Polish, Slovak and Slovene)
  - 2007: + 3 (Bulgarian, Romanian and Irish)
  - 2013: +1 (Croatian)

Europarl for research (Koehn, 2005)

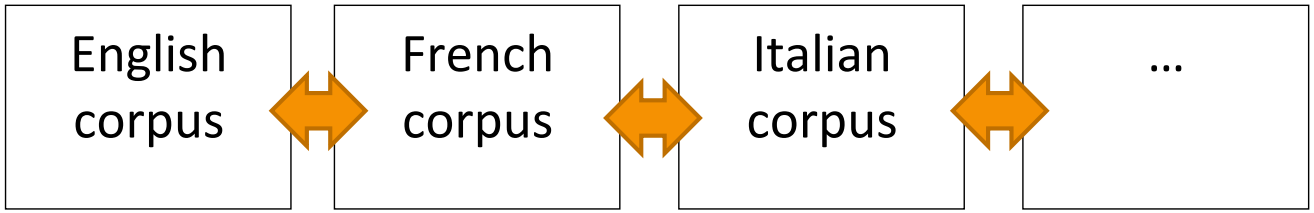
- Collection of **statements** in one file per **language** and per **day**.
- **Sentence aligned** for a number of language pairs.
  - http://www.statmt.org/europarl/
  - latest release (version 7) in May 2012
  - includes 21 languages
- The sub-corpora of several languages contain up to over **50 million words**, with **important cross-linguistic variations**.
  - **54 million words** in English, Spanish and French
  - **7 million words** in Polish, **10 million words** in Romanian

An example of sequence from Europarl

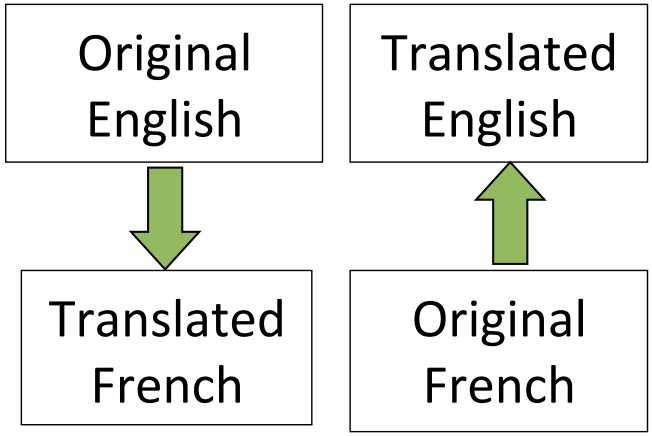
<SPEAKER ID=5 **LANGUAGE="DE"** NAME="Graefe zu Baringdorf">
Mr President, I do not want to talk about the content, but about BSE. The current debate is about the fact that gelatine is not safe. And we should like you to tell us...<P>
(The President cut the speaker off)
<SPEAKER ID=6 **LANGUAGE="ES"** NAME="Gutiérrez Díaz">
Mr President, there must be a mistake in the information you have been given. I am against the wording of this amendment and I have personally told Mr Santini. I understand the value he gives to the amendment but, in the explanatory statement - at the bottom of page 9 - the limits are well explained and I think it would be excessive to introduce it into the text, by a procedure that we think is too far reaching, without its being discussed beforehand in the committee.
<SPEAKER ID=7 **LANGUAGE="IT"** NAME="Santini">
Mr President, after an exchange of views with the rapporteur, I should like to withdraw this amendment, which was perhaps badly worded and has become even less clear in translation. Since, on the other hand, the report is extremely coherent and straightforward, I withdraw the amendment to avoid confusion.<P>
(Parliament adopted the resolution)
<SPEAKER ID=8 **missing** NAME="Posselt">
Mr President, I have given my agreement to Mr Cars' excellent report, although I have serious problems with the Council regulation, because I am bound to say that reconstruction in the Federal Republic of Yugoslavia is absurd, since nothing has been destroyed there, and large groups of refugees have come only out of Kosovo.
[…]

Directional corpora from Europarl

Comparable corpora in Europarl



Bi-directional parallel corpora in Europarl



Directional parallel corpora

<SPEAKER ID=35 LANGUAGE="EN" NAME="Cox"> Madam President, if the vote records correctly how my Group voted I shall not, and cannot, object to that. If your ruling is that I cannot give an explanation of vote, I accept that but with reservations.

<SPEAKER ID=35 LANGUAGE="EN" NAME="Cox"> Madame la Présidente, si le procès-verbal reflète correctement le vote de mon groupe, je n'ai et n'aurai aucune objection à formuler. Si votre décision est que je ne puis pas donner d'explication de vote, je l'accepte, mais avec certaines réserves.

<SPEAKER ID=40 LANGUAGE="EN" NAME="Simpson">
Madam President, first of all I should like to thank Mr Koch for his report which has, at its heart, the issue of transport safety. The report looks at the issue of harmonising the examination requirements for safety advisors working in the areas of transportation of dangerous goods by road, rail and inland waterway. I congratulate him on his excellent report.

<SPEAKER ID=40 LANGUAGE="EN" NAME="Simpson">
Madame la Présidente, je voudrais avant tout remercier M. Koch de son rapport dans la question de la sécurité des transports occupe une place centrale. Il envisage l'harmonisation du niveau des exigences applicables à l'examen des conseillers à la sécurité pour le transport de marchandises dangereuses par route, par rail ou par voie navigable. Je le félicite de son excellent rapport.

Translation spotting of ‘since’

Source	Target
Will we speak with one voice when we go to events in the future **since** we now have our single currency about to be born?	Parlerons-nous d’une seule voix lorsque nous n arriverons aux événements futurs, **puisqu**’à présent notre monnaie unique est sur le point de voir le jour ?
In East Timor an estimated one-third of the population has died **since** the Indonesian invasion of 1975.	Au Timor Oriental, environ un tiers de la pulation est décédée **depuis** l’invasion ndonésienne de 1975.
C'est beaucoup trop et leur nombre devrait être très sensiblement diminué, **car** il s'agit d'auta de féodalités.	This is far too many, and the number needs to be siderably reduced, **since** it is nothing more than a feudal system.
Monsieur le Président, **comme** je suis un élu bordelais, je croyais que vous me donniez la parole pour répondre à mon collègue bavarois au sujet du vin de Bordeaux.	Mr. President, I thought, **since** I represent the leaux area, that you were giving me the floor so that I could answer my Bavarian colleague on the subject of Bordeaux wine.

Uses of translation spotting

- Identify the range of **translation equivalents.**
  - the results of translation spotting cover a **broader range of equivalences** than the lists provided by bilingual dictionaries
  - **not limited** to **similar parts of speech** (paraphrases, syntax, etc.)
  - includes **implicit relations**
- Identify the **discourse relation** for **ambiguous connectives**.
  - *since* is ambiguous between temporal and causal relations, but in French the two senses are communicated by different connectives
  - worthwhile technique because sense annotation is hard

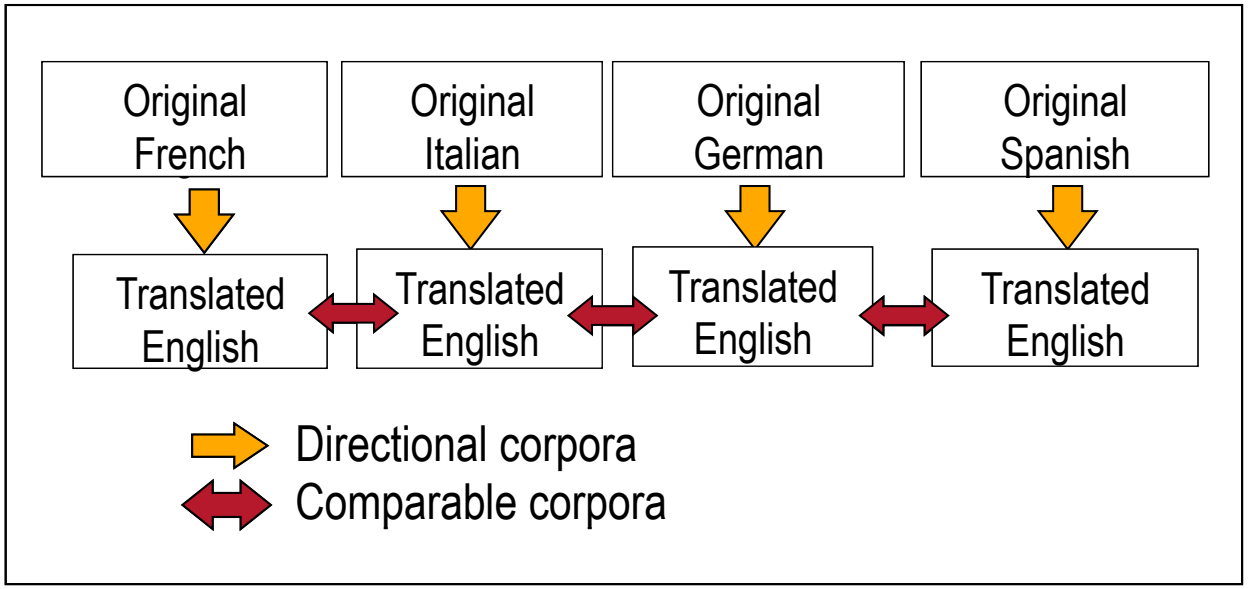
Defining translations

- Translations represent a **specific genre**.
  - 'translationese' (Gellerstam 1996)
  - 'third code' (Baker 2003)
- **Typical features** of translations
  - translations are **simpler** than original texts (Laviosa-Braithwaite 1996)
  - the **items** that are **unique** in the target system are **under-represented** in translations (Tirkkonen-Condit 2000)
  - translations are **more explicit** than original texts due to an increase of **cohesion markers** (Blum-Kulka 1986)

Testing the explicitation hypothesis

- Focus on the number of added **causal connectives** in translations.
  - causal connectives are particularly volatile in translation
- Comparisons between several **source languages**.
  - Italian, Spanish, German, English, French
- Comparison between two **target languages.**
  - English and French
- Comparison between several **connectives**.
  - in **French**: *parce que, car, puisque, étant donné que*
  - in **English**: *because, since, given that*

Comparable corpora in Europarl

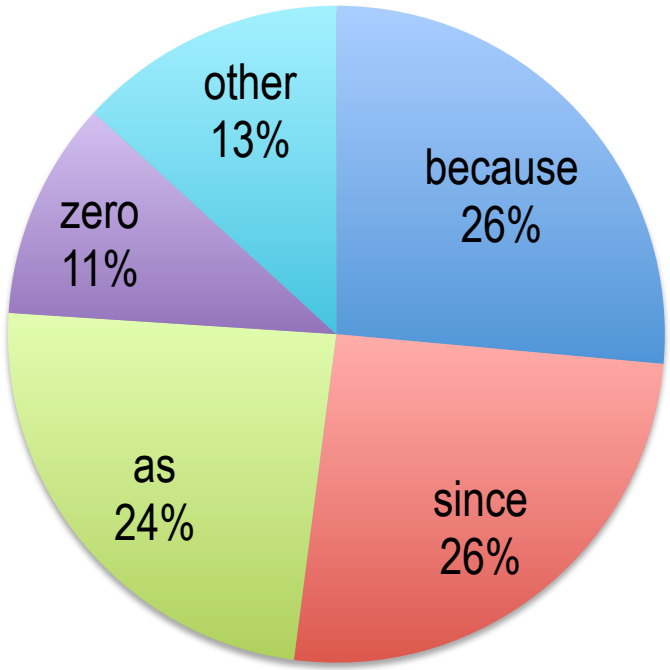


- In translated **English** and in translated **French**, extraction of **200 occurrences** of each causal connective in all sub-corpora.
- Translation spotting of **equivalents** in **source corpora.**

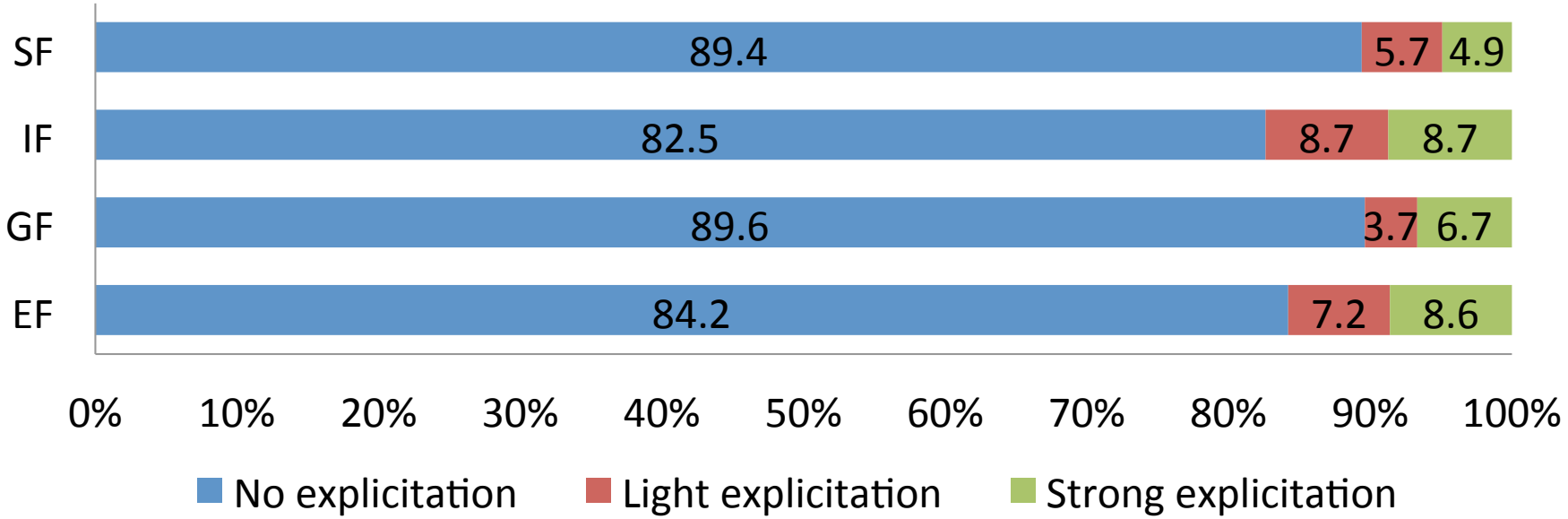
Backward translation spotting

- What does “puisque” translate?

Equivalents in English source texts	Nbr.
because	32
since	31
as	29
zero	13
gerund	9
given that	4
if	3
as well as	1
whilst	1
while	1

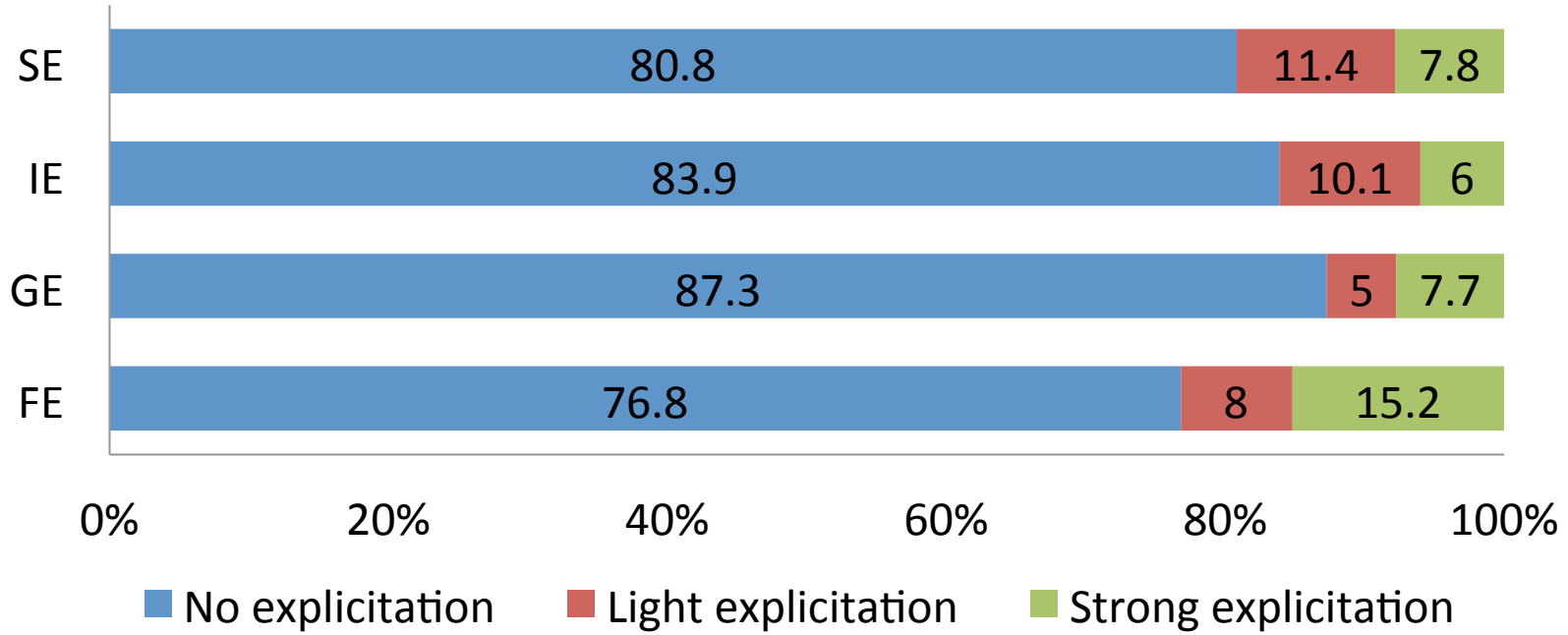


The influence of source languages



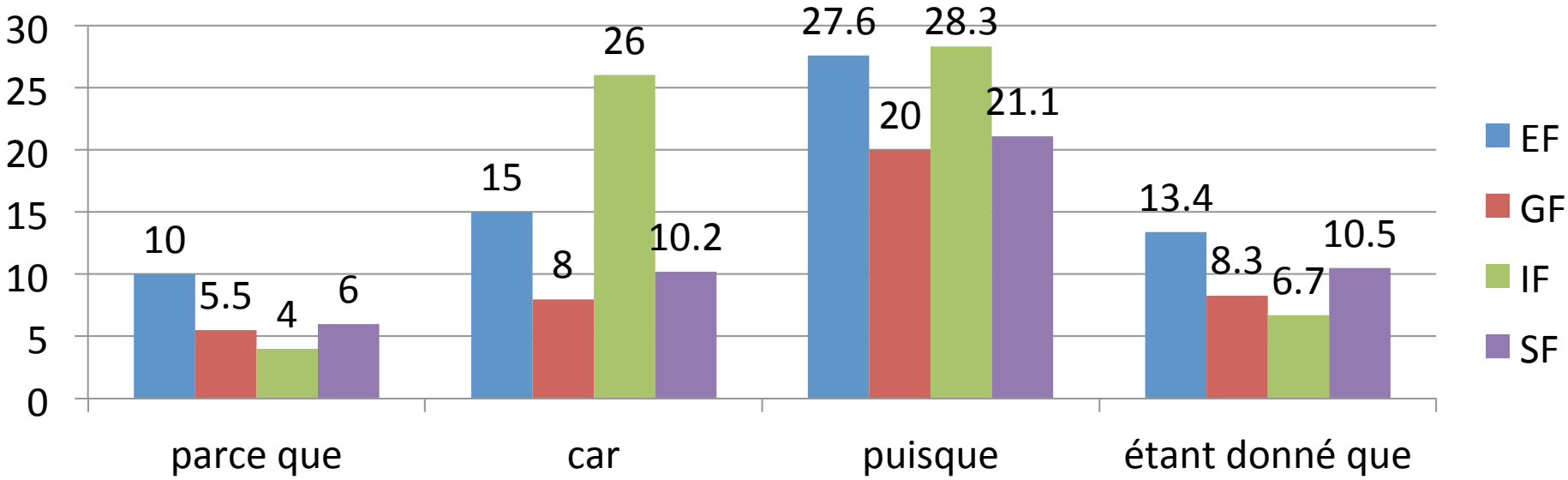
- **No significant differences** between **source languages**.
  - χ^2 = 3.67, df = 6, and p = 0.68

Comparison with English as a target language



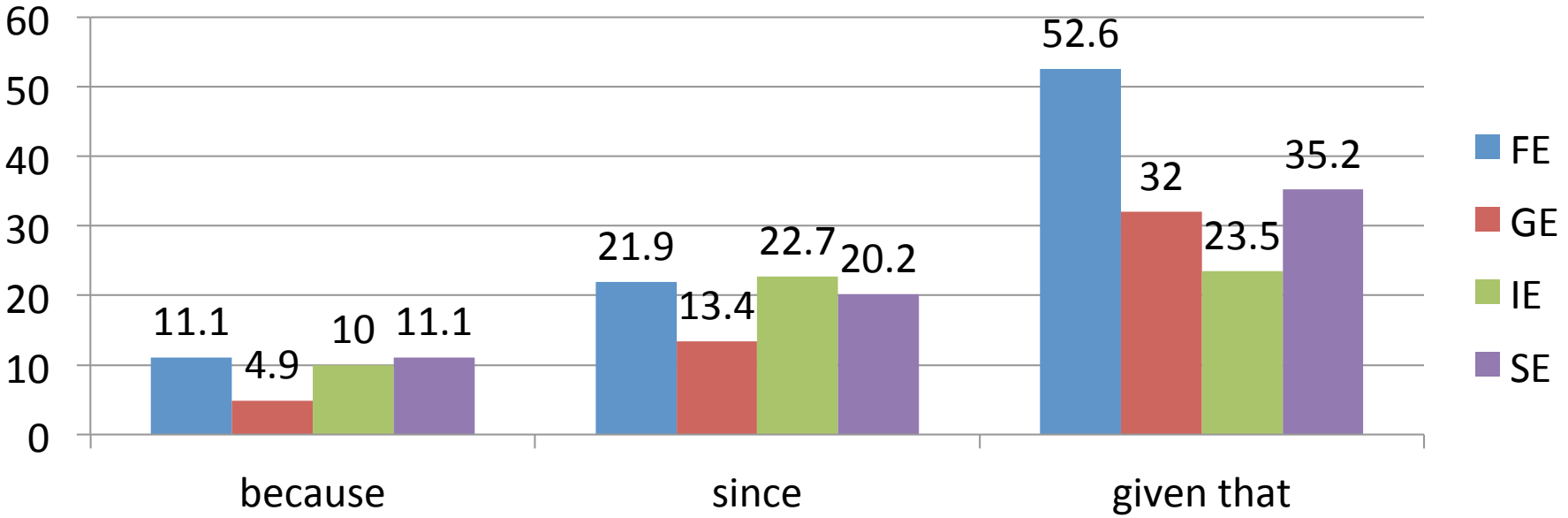
- **No significant differences** between **source languages** either.
  - $\chi^2 = 8.8$, df = 6, and $p = 0.18$
- **No significant difference** between the **two target languages** for any of the source languages.

The influence of connectives (FR)



- A significant difference between **all connectives** was found for **all source languages**.
- An analysis of standardized residuals reveals an **overuse** of *puisque* and an **underuse** of *parce que.*

The influence of connectives (EN)



- Again, a significant difference between **all connectives** was found for **all source languages**.
- An analysis of standardized residuals reveals an **overuse** of *given that* and an **underuse** of *because*.

Connectives and explicitation

- To **summarize**:
  - the connectives that trigger most cases of explicitation are *puisque* in French and *given that* in English
  - the connectives that trigger the least cases of explicitation are *parce que* in French and *because* in English.
- These results can be related to the **semantic profile** of these **pairs of connectives**.
  - the connectives that **trigger explicitation** convey **subjective** causal relations and introduce **given** information (shared between speaker and hearer)
  - the connectives that **don't trigger explicitation** have opposite profiles: they convey **objective** relations and introduce **new** information

Take home message

- **Translation corpora** represent a **valuable resource** for the multilingual study of discourse relational devices (DRDs).
- Large **multilingual corpora** can be used as **comparable** and **directionnal** corpora by spliting the data in different ways.
- **Translation spotting** is a simple way to use **multingual data** to shed light on DRD **senses** and their **realization**.
- **Translated data** is a **specific genre** and should not be treated as equivalent to source language data.

Plan for the afternoon lab session

- **Schedule**
  - three activities of about **an hour** each
  - work in **small groups** for **30 minutes on each activity**
  - then **30 minutes** of **reporting** and **discussion for each activity**
- **Three activities**
  - **Identifying DRDs** in English written and spoken data
  - **Annotating discourse relations** in four different dimensions
  - **Analyzing** the **results** of **translation spotting** in English as a target language with data from two different registers

References

- Cartoni B. & Meyer T. (2012). Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of LREC 2012*, 2132-2137, Istanbul, Turkey.
- Ilisei L., Inkpen D., Corpas Pastor G. & Mitkov R. (2010). Identification of Translationese: A Machine Learning Approach. In Gelbukh, A. (Ed), Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, 503-511.
- Koehn P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, September 13-15, 79-86, Phuket, Thailand.
- Noël D. (2003). Translations as evidence for semantics: An illustration. *Linguistics* 41(4):757-785.
- Zufferey S. & Cartoni B. (2014). A multifactorial analysis of explicitation in translation. Target 26(3): 361-384.