

Data and other challenges

November 29, 2023

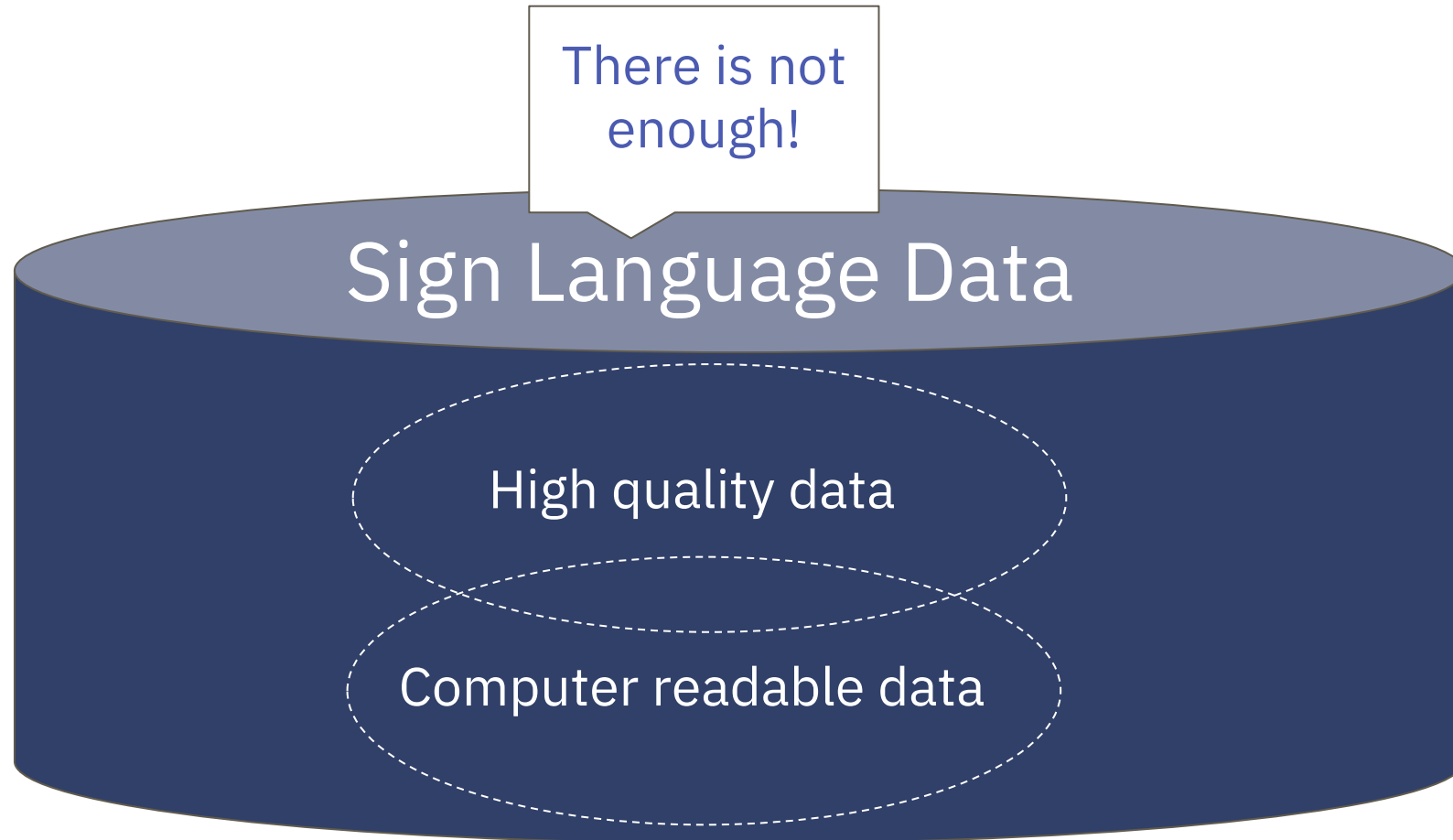
Introduction

- The data bottleneck
- Collecting data
 - Perspective gap (expectations management)
 - “Native signer” (metadata)
- Processing data
 - Manual labor + suboptimal tools/software
- Conclusions



Caro Brosens
linguistic researcher

Sign Language Data Bottleneck



Collecting Sign Language Data |

What data to use?

The ideal sign language data...

...is representative of the whole community

- gender,
- age,
- regional variation,
- registers,
- additional disabilities,

... features highly proficient signers

...meets all requirements for all research domains (linguistics, deaf studies, NLP, computer sciences, ...)

- amount of cameras + angles,
- amount of signers on screen,
- topics covered,
- types/detail of annotations,
- eye tracking
- motion capture
- ...

Perspective gap

Deaf people & signers

= high quality data

! there is not enough



Computer scientists & MT experts

= big quantity of data

! just not always available

! big variation in sign proficiency

The myth of the Native Signer

Native speaker: language of the home, community and formal education.

Native signer (Miller et al, 2015; Hauser et al, 2018)

- prelingual deafness / deaf from birth
- loss over 85db in the better ear
- sign language is the preferred/main form of communication
- exposure to sign language since birth through deaf signing parents
- be educated in (some form of) sign in special education

Metadata – signer

- **Auditory status:** deaf, hard-of-hearing, hearing
- **Age of occurrence:** from birth/pre-lingually, as a child, as an adult, ...
- **Primary/preferred language**
- **Educational background:** special education, regular education, ...
- **Age of SL acquisition:** < 3, as a child, as an adult, ...
- **Familial background:** deaf/hearing, (non)signers, ...

Metadata – general context

- **Type of video:** social media post, call, research report, ...
- **Type of language:** spontaneous, interpretation/translation, prompted, ...
- **Register:** formal, informal, vulgar, ...
- **Source language**
- **Target language**
- **Available parallel data:** none, subtitles, voice interpretation, translation
- **Target audience (of the message):** signers, hearing individuals, students, ...

Example



Vacancy VGTC



Interpreted newscast

Whose data is it?

- (temporary) collaborations within the community
- flexible and ever-changing roles
- natural evolution of the community

Who should collect data?

The collector should:

- be familiar with the language community
- know the language themselves
- have (or be able to win) the trust of the community members

Processing Sign Language Data |

Sign Language Data

= not machine readable

- What kinds of “enrichments” are needed?
 - translations, glosses, phonetic descriptions, non-manual markers (e.g. eyegase)...
- Who should process the data?
- What is the most efficient way to do this?
 - Barriers regarding software

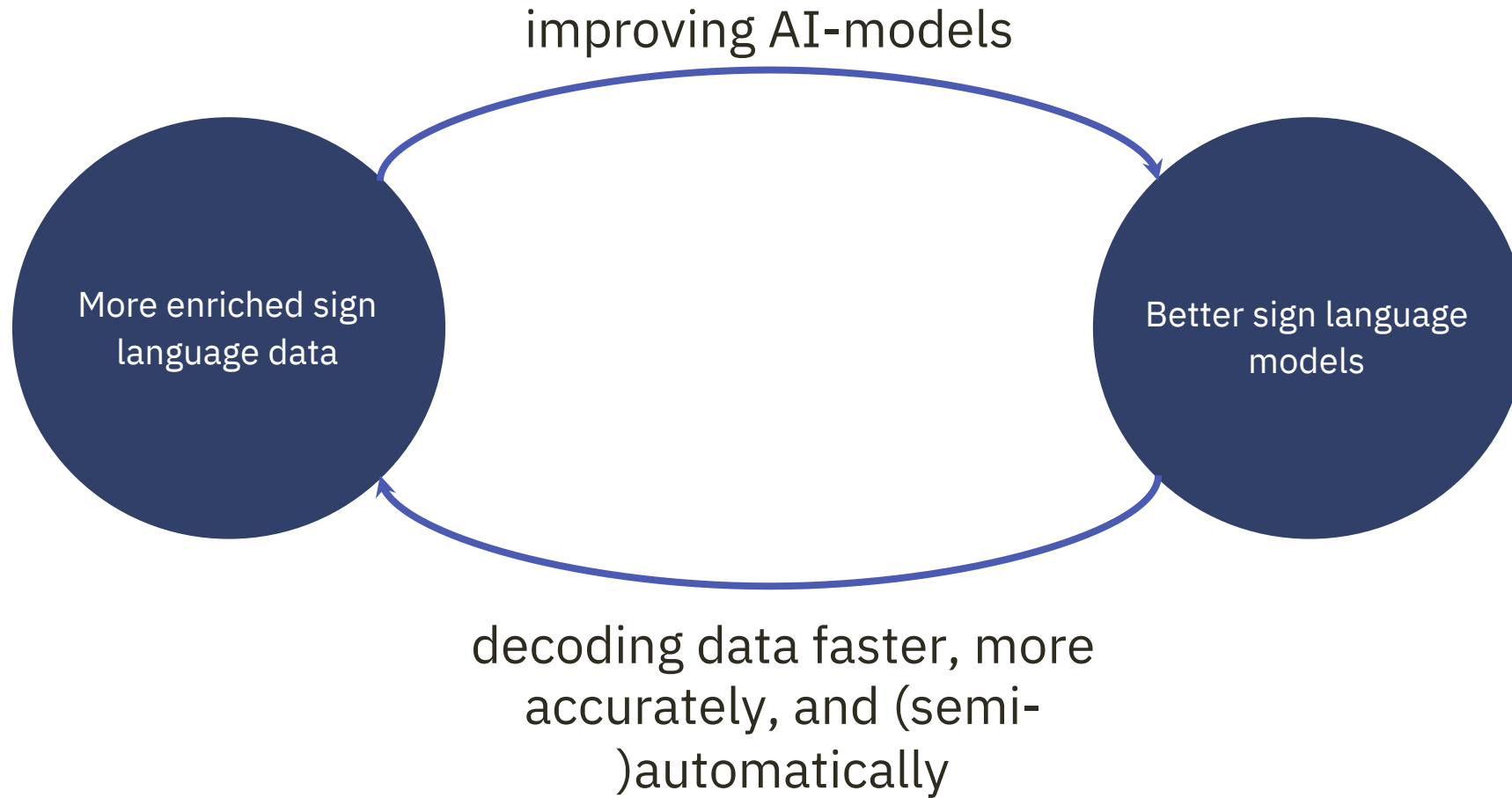
! inconsistencies

Eudico Linguistic Annotator – ELAN

The screenshot displays the ELAN software interface for the file 'Elan - HSG1.eaf'. The main window is divided into a video view and a timeline view. The video view shows two men sitting on chairs against a blue background, engaged in a conversation. The timeline view below the video shows a series of annotations for the video segment. The current time is 00:00:17.470, and the selection range is 00:00:17.370 - 00:00:17.470. The timeline includes various linguistic annotations such as 'RH ID gloss', 'LH ID gloss', 'Dependent variabl', '1-handed/2-hande', 'Grammatical functi', 'HS preceding target', 'HS following target', 'Genre of text which', and 'Summary'. The 'LH ID gloss' table is also visible on the right side of the interface.

Nr	Annotation	Begin Time	End Time	Duration
1	SAME	00:00:15.140	00:00:15.330	00:00:00.190
2	SCOTLAND	00:00:15.330	00:00:15.630	00:00:00.300
3	HOLD-FLAG	00:00:17.040	00:00:17.270	00:00:00.230
4	PT:loc	00:00:17.270	00:00:17.590	00:00:00.320
5	G	00:00:17.590	00:00:17.890	00:00:00.300
6	PEOPLE-LOOK	00:00:19.380	00:00:19.640	00:00:00.260
7	NORTH	00:00:19.650	00:00:19.860	00:00:00.210
8	IRELAND	00:00:19.870	00:00:20.310	00:00:00.440
9	SAME	00:00:20.320	00:00:20.550	00:00:00.230
10	PEOPLE-LOOK	00:00:20.550	00:00:20.930	00:00:00.380
11	PT:pro1	00:00:24.420	00:00:24.610	00:00:00.190

Cycle



Conclusions |

How to resolve the data bottleneck?

- Expanding the amount of enriched data
 - Better technical tools!
- Expanding the amount of high quality data
 - Community media?

Thank you!

For more information: caro.brosens@vgtc.be