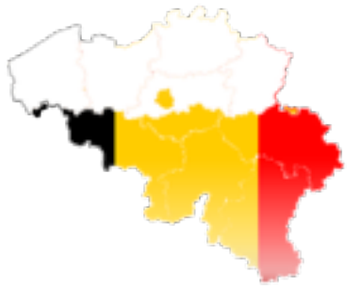
USING SIGN LANGUAGE CORPORA AS BILINGUAL CORPORA FOR DATA MINING CONTRASTIVE LINGUISTICS AND COMPUTER-ASSISTED ANNOTATION

LAURENCE MEURANT, F.R.S.-FNRS & UNIVERSITY OF NAMUR, BELGIUM
ANTHONY CLEVE, UNIVERSITY OF NAMUR, BELGIUM
ONNO CRASBORN, RADBOUD UNIVERSITY, THE NETHERLANDS

laurence.meurant@unamur.be anthony.cleve@unamur.be o.crasborn@let.ru.nl

LREC 2016 - 7th Workshop on the Representation and Processing of Sign Languages: Corpus mining
28 May 2016, Portorož (Slovenia)

CORPUS LSFB



- 150 hours of video
- 12 hours glossed
- 3 hours translated
- 104,000 left and right hand glosses; 2,500 types
- Paragraph-level translations

CORPUS NGT

- 72 hours of video
- 15 hours glossed and translated
- ±100,000 signs (150,000 left and right hand glosses; 3,400 types)
- Sentence-level translations

SIGN LANGUAGE – SPOKEN LANGUAGE COMPARISON

L'Égypte possède l'un des meilleurs réseaux électriques du monde en développement, ses services pouvant desservir pratiquement toute la population. *ofid.org*

L'accélération actuelle des hausses de salaires et les tensions provenant du vif développement de la demande intérieure constituent une difficulté pour le maintien de la stabilité des prix. *ecb.europa.eu*

Nous leur devons dès lors quarante-cinq années de retard en matière de développement social et de croissance économique. *europarl.europa.eu*

- Relying on corpus-based contrastive linguistics' methodology
- SL corpora as bilingual corpora

FRENCH EQUIVALENTS OF LSFB AUTREFOIS [IN THE PAST, FORMERLY]

AUTREFOIS

[IN THE PAST, FORMERLY]

- Idiom
- Tense (imperfect)
- 'Avant,' + imperfect

- Oui, bien sûr. Ça n'a pas toujours été facile. Je suis sourde, née dans une famille d'entendants.

  [Yes, of course. **It wasn't always** easy. I am deaf and born in a hearing family.]

- Avant, je travaillais là-bas.

  [I used to work there.]

- Comment se passaient les fêtes de Noël dans ta famille?

  [How did the Christmas days look like in your family?]

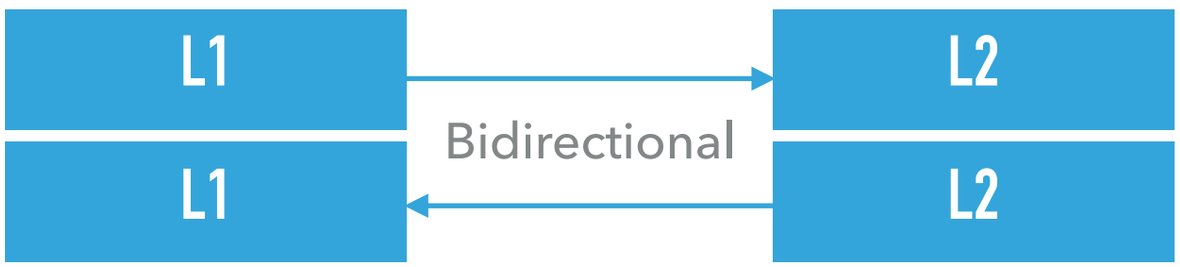
CONTENT

1. SL corpora as **translation** corpora
2. **Aligning** the bilingual data
3. **Exploiting** the bilingual properties of SL data

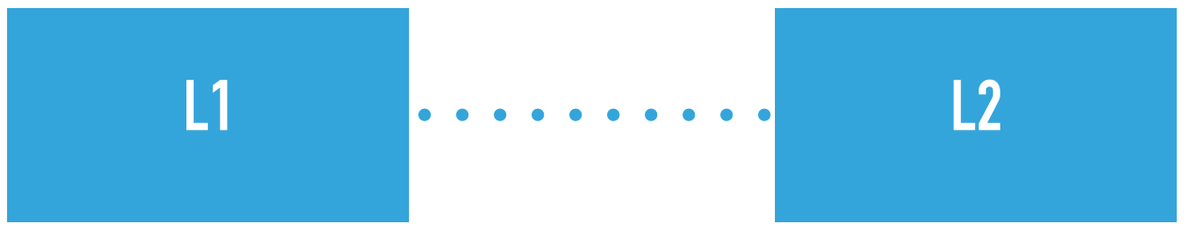
1. SL CORPORA AS TRANSLATION CORPORA

MULTILINGUAL CORPORA

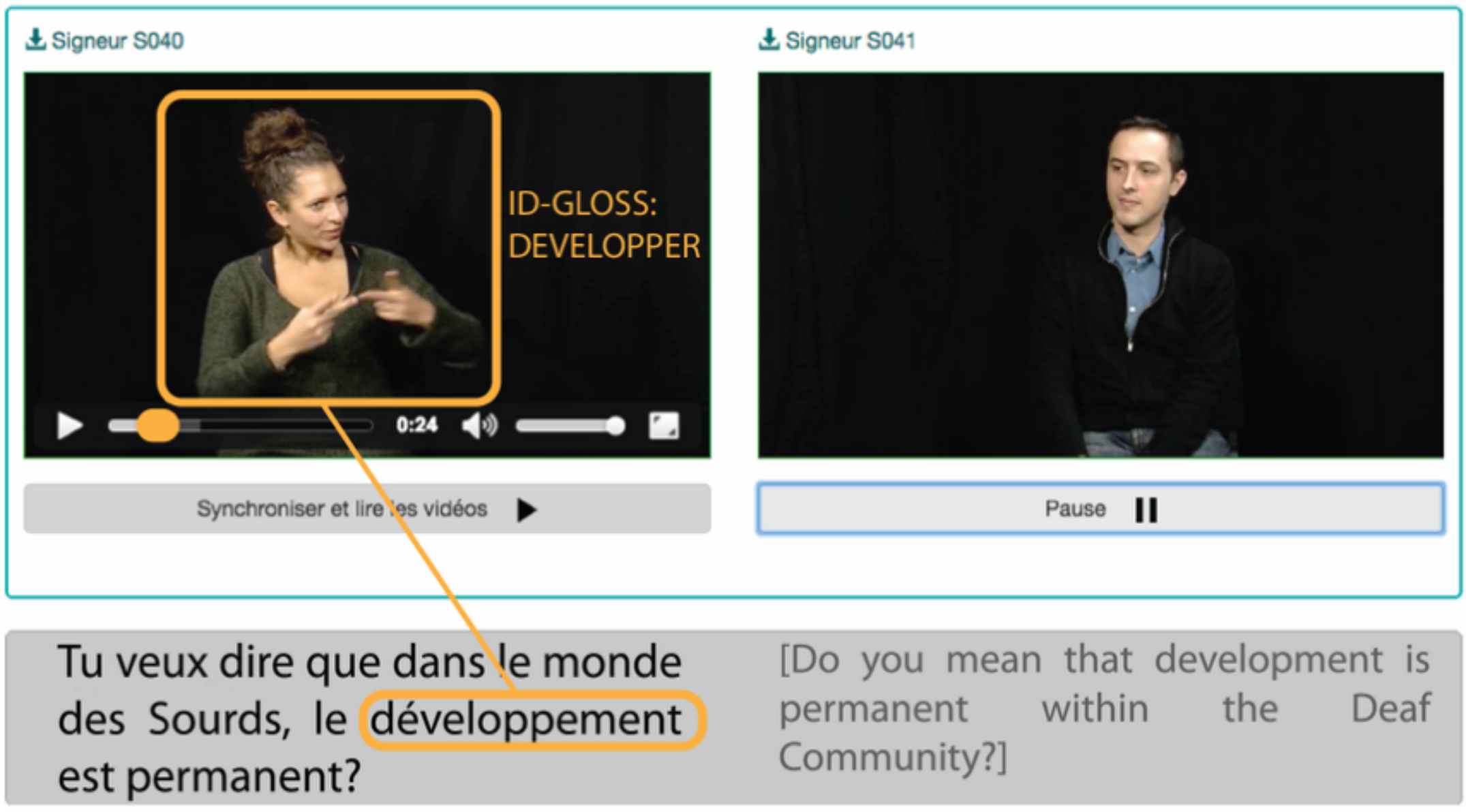
Translation corpora



Comparable corpora



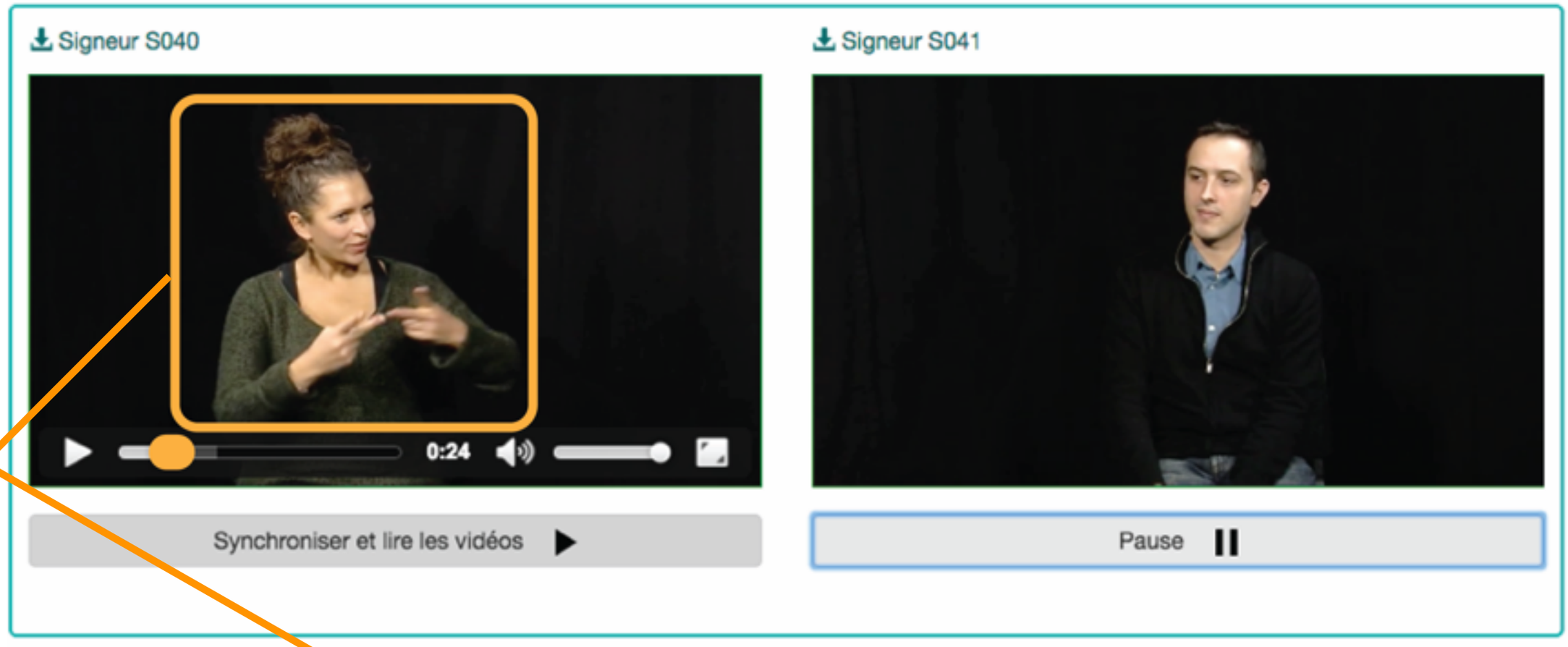
SL CORPORA ARE BILINGUAL



2. ALIGNING THE BILINGUAL DATA

CHALLENGE

DIRE
MONDE
SOURD
TOUJOURS
LA
DEVELOPPER
TOUJOURS



Tu veux dire que dans le monde des Sourds, le développement est permanent?

[Do you mean that development is permanent within the Deaf Community?]

AVAILABLE DATA

VIDEOS AND ANNOTATION FILES

Glosses (tokens)
Time codes

DIRE - MONDE - SOURD - TOUJOURS - LA - DEVELOPPER - TOUJOURS

[SAY - WORLD - DEAF - ALWAYS - THERE - DEVELOP - ALWAYS]

TRANSLATIONS

Tu veux dire que dans le monde des Sourds, le développement est permanent?

[Do you mean that in the Deaf world, development is permanent?]

Aligned in the annotation files

- LSFB: paragraph-level
- NGT: sentence-level

LEXICAL DATABASE

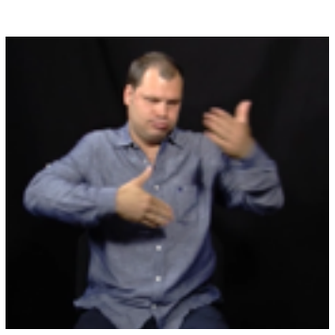
DEVELOPPER développer, développement

Sign types

Translation equivalents (manually added)

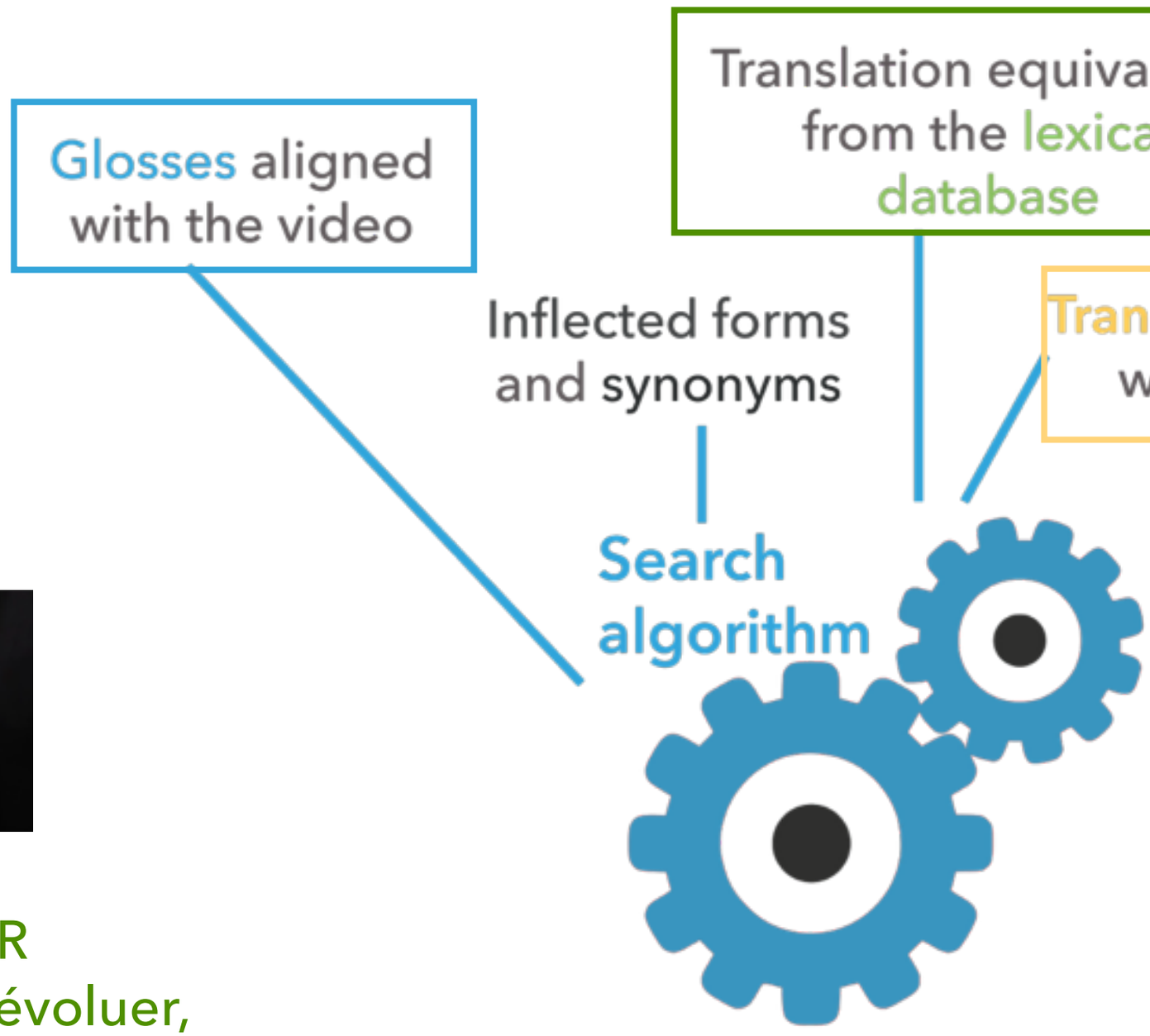
+ EXTERNAL RESOURCES: DICTIONARIES OF INFLECTED FORMS AND OF SYNONYMS

METHOD



AVANCER
avancer, évoluer, développer, fonctionner

progressait



CURRENT RESULTS OF THE MATCHING

one gloss matching different words

	Tokens	Exact match	Inflected form	(Inflected) Synonym	No match	Total matches	One-to-many	Many-to-one
LSFB	18,770	9,668	3,596	4,764	6,592	18,028	3,211	5,800
NGT	33,942	18,969	3,226	*	14,972	22,195	2,794	7,443

several glosses matching the same word

SING-WORD STABLE PAIRS?

UNEXPECTED MEANINGS?

REPENTIONS?

EXAMPLE: ONE-TO-MANY MATCHING

*Oui, c'est frappant. Pour passer à **autre** chose… Quand une personne invente un **nouveau** signe, je ne le remarque pas. (Session 03 Tâche 06)*

Id Annot	Libele	Mot Clés	Type Match	Mot Trouvé
48S008D35	NOUVEAU.FLASH	nouveau, recommencer	EXACT	nouveau
48S008D35	NOUVEAU.FLASH	nouveau, recommencer	EXACTSYNONYM	autre

CURRENT AND FUTURE WORK

- Review and validate a sample of the results

Id Annot	Libele	Mot Clés	Type Match	Mot Trouvé	Valide?	Traduction	Commentaire
24S003D0	PT:PRO1	je, moi, me	NULL			Mon nom en langue des signes est Hyppolite, comme ceci [LUNETTES-C]. Avant, j'étais à l'école à Liège puis à l'école de Woluwé.	Valeur : Même signe Expression NULL Enregistrer Même signe : Expression : Matching manuel :

- Use the reviewing process to improve the matching algorithm
- Combine results with statistics on colocations
- Extract examples in context for each meaning of each sign

3. EXPLOITING THE BILINGUAL PROPERTIES OF SL DATA

LEXICON: MEANINGS OF SIGNS IN CONTEXT

IN THE LEXICAL DATA BASE

AUSSI
[ALSO]

aussi [also], même [same], comme [as]

→ **manually encoded**

FROM THE LSFB/FR BILINGUAL DATA

AUSSI
[ALSO]

aussi, même, comme, aussi que, autant, comme si, égal, également, en plus, même chose, parce que [because], tout aussi, tout comme, vu que [given that], en quelque sorte [kind of], plus, / [no lexical equivalent]

→ **automatically detected**

AUSSI [ALSO]

- [WORLD - DEAF - DS:small - ALSO - TRUST - WHAT]

- Vu que le monde de la surdité est si petit, comment les gens pouvaient-ils avoir confiance en moi?

  [Given that the Deaf world is so small, how could people trust me?]

- [ALSO - IMAGINE - PT:POSS - CHILD - KID - HOW - ACCESSIBILITY - IMAGINE]

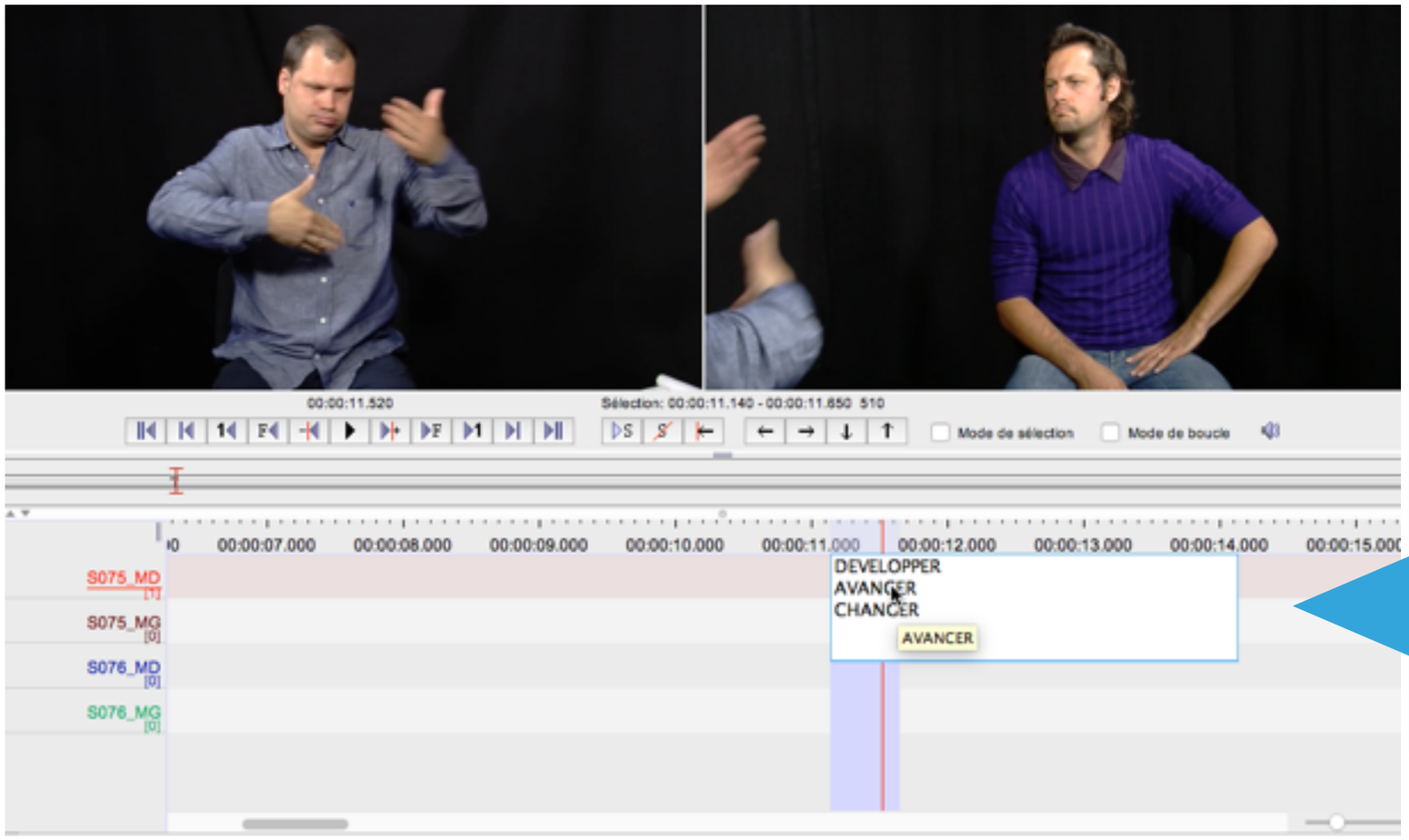
- En quelque sorte, ça me permet d'en tirer des leçons sur la façon dont je veux promouvoir l'accessibilité pour mes enfants.

  [In some way, it makes me learn about the way I want to promote accessibility for my kids.]

EQUIVALENTS OF NON-LEXICAL ELEMENTS

LSFB/NGT	FR/NL
? ←	passive forms
left vs. right oppositions →	?
? ←	prepositions
partly lexicalized signs →	?
? ←	discourse markers

SUGGESTIONS TO THE ANNOTATOR WHILE ANNOTATING



suggested glosses for one meaning

(from the enriched list of meanings in context)

+ suggestions based on the text of the previously translated video ?

LIMITS

The relevance of the results relies

- on the annotation and the translation quality

The creation of a bidirectional translation corpus would mitigate the impact of the translation quality

- and then on the quantity of annotated and translated data

In order to develop automatisation, we still depend on the laborious fully-manual annotation

FORESEEN IMPROVEMENTS

Linguistic description

New insights on SL by comparison with SpL

Direct comparison of signed and spoken grammar

Annotation

Computer support for the manual process of annotation

Literacy

Resources based on sign to acquire a higher level of SpL

Bilingualism

Resources to improve the bilingual competence of teachers and learners of SL

Material for interpreters, trainers and SL teachers

Support to assist translation, reading, writing tasks

FORESEEN IMPROVEMENTS

Literacy
Resources based on sign to acquire a higher level of SpL

Pupils from the bilingual « co-enrollment » program in Namur (Belgium)

Thank you!

This study would not have been possible without the signers who participated to the LSFB and the NGT corpus projects, nor without the help of Maxime Gobert and Eric Bernagou from the University of Namur. We are grateful to them.

L. Meurant and A. Cleve are supported by the Incentive Research Progamme of the University of Namur. O. Crasborn is supported by NWO grant 360.70.500 ‘Form-Meaning Units’.

BIBLIOGRAPHY

B. Altenberg et al., editors. (2002). Lexis in contrast: corpus-based approaches . John Benjamins Publishing.

Börstell, C., Mesch, J., and Wallin, L. (2014). Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In Onno Crasborn, et al., editors, Beyond the manual

channel. 6th Workshop on the Representation and Processing of Sign Languages , pages 7–10, Reykjavik. ELRA.

Bourdaillet, J., Huet, S., Langlais, P., and Lapalme, G. (2010). TransSearch: from a bilingual concordancer to a translation finder. Machine Translation , 24:241–271.

Onno Crasborn, editor. (2007). Identifying sentences in signed languages , volume 10-2 of Special issue of Sign Language and Linguistics . John Benjamins Publishing Company, Amsterdam; Philadelphia.

Deville, G., Dumortier, L., Meurisse, J.-R., and Miceli, M. (2013). Ressources lexicales: Contenu, construction, utilisation, ´evaluation. In N. Gala et al., editors, Ressources lexicales pour l´aide `a l´apprentissage des langues , volume 30, pages 291–312. John Benjamins.

Fenlon, J., Denmark, T., Campbell, R., and Woll, B. (2007). Seeing sentence boundaries. In Onno Crasborn, editor, Special issue of Sign Language & Linguistics , volume 10-2, pages 177–200. John Benjamins Publishing Company, Amsterdam; Philadelphia.

Gellerstam, M. (1996). Translations as a source for crosslinguistic studies. Lund studies in English , 88:53–62.

Gilquin, G. (2000). The integrated contrastive model: Spicing up your data. Languages in Contrast , 3:95–123.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, et al., editors, Languages in Contrast.

Text-based cross-linguistic studies , volume 88 of Lund Studies in English , pages 37–51. Lund University Press.

Johansson, S. (2007). Seeing through multilingual corpora. Language and Computers , 62:51–71.

Johnston, T. and Schembri, A. (2010). Variation, lexicaliztation and grammaticalization in signed languages. Langage et société , 131:19–35.

Johnston, T. (2010). From archive to corpus: transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics, 15:106–131.

Linguee. (2015). http://www.linguee.com.

Ormel, E. and Crasborn, O. (2012). Prosodic correlates of sentences in signed languages: A Literature Review and Suggestions for New Types of Studies. Sign Language Studies, 12(2):109–145.

Crasborn, O. and Zwitserlood, I. and Ros, J. (2008). The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen, ISLRN 175-346-174-413-3.

Meurant, L. (2015). Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB). Laboratoire de langue des signes de Belgique francophone (LSFB-Lab). FRS-F.N.R.S et Université de Namur.