easier

intElligent Automatic
SIgn languagE tRanslation

NLP for sign languages:

How to index sign language resources?

Sam Bigeard
University of Hamburg
Institute of German Sign Language and Communication of the Deaf

Also presenting the work of Maria Kopf, Marc Schulder and Thomas Hanke

17/10/2022

About me

- Linguistics bachelor at Paris 3, NLP master at INALCO
- PhD at Lille 3 on mining social media for medication misuse
- Post-doc at IECL/ATILF in 2020 under Marianne Clausel and Mathieu Constant
  - Text mining on publications of Cancéropôle Est to highlight their research
- After that, wanted to work on another domain for a change and do more linguistics -> sign language research

Summary

1. A brief overview of sign languages
2. EASIER project
3. harmonising corpora
4. harmonising lexicons with Wordnet

easier

Sign languages:

The basics

Sign languages: the basics

Sign languages are natural languages

- Many SL exist
- They are not a word-to-word translation of a spoken language. They have their own grammar and vocabulary.
- They have synonyms, local variants, slang, wordplay, etc
- They are part of the culture of a minority
  - not a problem to be solved
  - complete, functional languages
  - often butchered
- They have a lot of variation (18 different “September” in a corpus of German sign language)

Sign languages: the basics

https://www.sign-lang.uni-hamburg.de/meinedgs/html/1246329_en.html

Examples of:

- Anonymisation
- A sentence of 2 signs translated to English as 11 words in 2 sentences
- Signing space used to describe locations, pronouns, time

Sign languages: the basics

- Phonetic features:
  - Manual signs have 4 articulators: shape, orientation, position, movement
  - There are 2 hands, they can have different configurations, and a movement relative to each other
  - Besides manual signs, there is mouth gestures, mouthings, facial expression, body orientation

- Use of the signing space for grammar

Obstacles to NLP

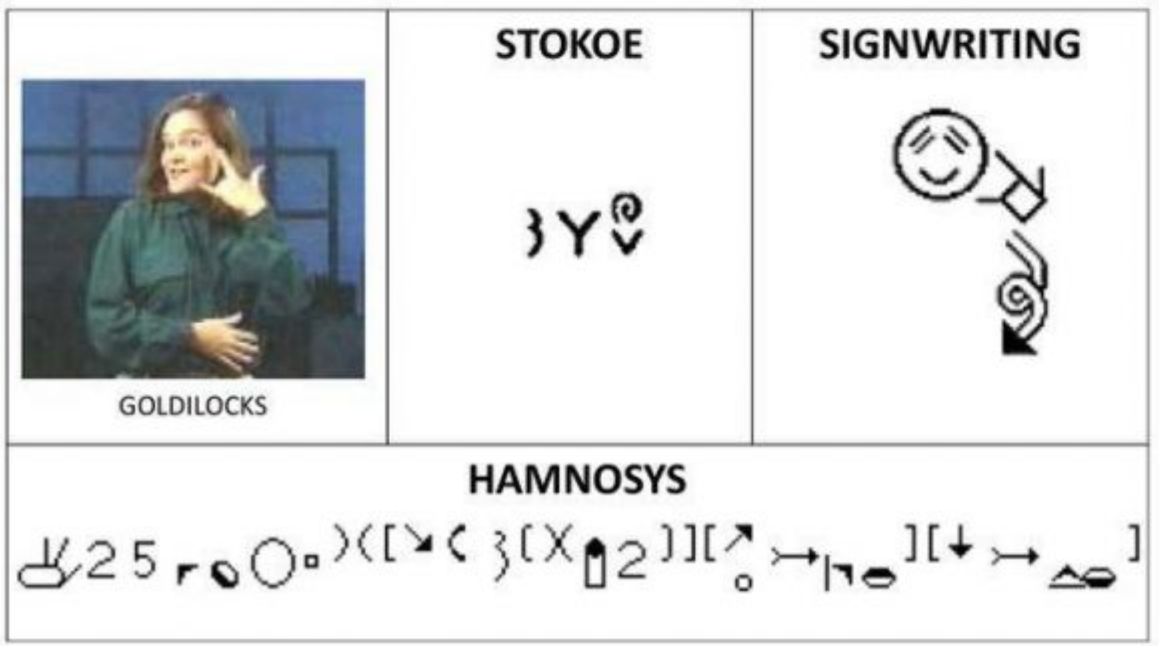
- Share issues with under-resourced spoken languages: no standardised writing, no lemmatisation software, few corpora, few dictionaries, corpora need costly manual annotation
- Requires video: more technically difficult to do well (motion blur, angles), more expensive, difficult to store and share, anonymisation is not possible

Transcription: Phonetic transcription

Stokoe: only encodes manuals. no unicode

Signwriting: can encode non-manuals, easy to learn, used by some signers, limited unicode needs software

Hamnosys: no unicode but can be typed with special font, can encode non-manuals



Used by linguists, rarely signers

Each can have variation

No automatic transcription software

Transcription: Gloss-ID transcription

[DGS corpus (dog)]

[BSL signbank (imagine)]

[DGS corpus (DOWNWARDS1^)]

- Sign-level, not phonetic
- Easier, quicker to use
- Need to establish a vocabulary, decide lemmatisation…
  - 2 distinct signs / same sign with flexion
- Vocabularies are different between projects, even for universal signs like pointings

BSL Corpus	PT:PRO1SG
Corpus FinSL	OS:me
Corpus NGT	PT-1hand:1
DGS Corpus	ICH1
POLYTROPON	ΕΓΩ
SSLC	PRO1
PJM Corpus	WSKAZ: 1 (JA)

easier

EASIER Project and my work

EASIER project

- Horizon 2020 project over 2021-2023
- Goal: Create a framework for sign-to-spoken, spoken-to-sign, and sign-to-sign translation
- Partners from France (LISN Univ Paris Sud), Germany (Univ Hamburg), UK, Greece, the Netherlands, Switzerland and Belgium
- Involves the European Union of the Deaf and deaf researchers
- 7 sign languages and 7 spoken languages of Europe
    - british, german, swiss german, french, greek, dutch, italian

EASIER project

Avatar Paula

[Dimou et al, 2022]

[video]

EASIER project

Fact: There is not enough corpora to train machine translation on any European sign language

- DGS corpus: 64 000 sentence pairs
- BSL corpus: 6 000 sentence pairs
- English-German standard translation corpus: 150 000 000 pairs

The project's research question: European SL are similar enough that we can combine corpora, have machine translation produce grammatical sentences, and simply substitute vocabulary to translate to each SL

My task:

- Harmonise corpora so they can be used together as training data
- Index vocabularies so they can be used for word-to-word translation

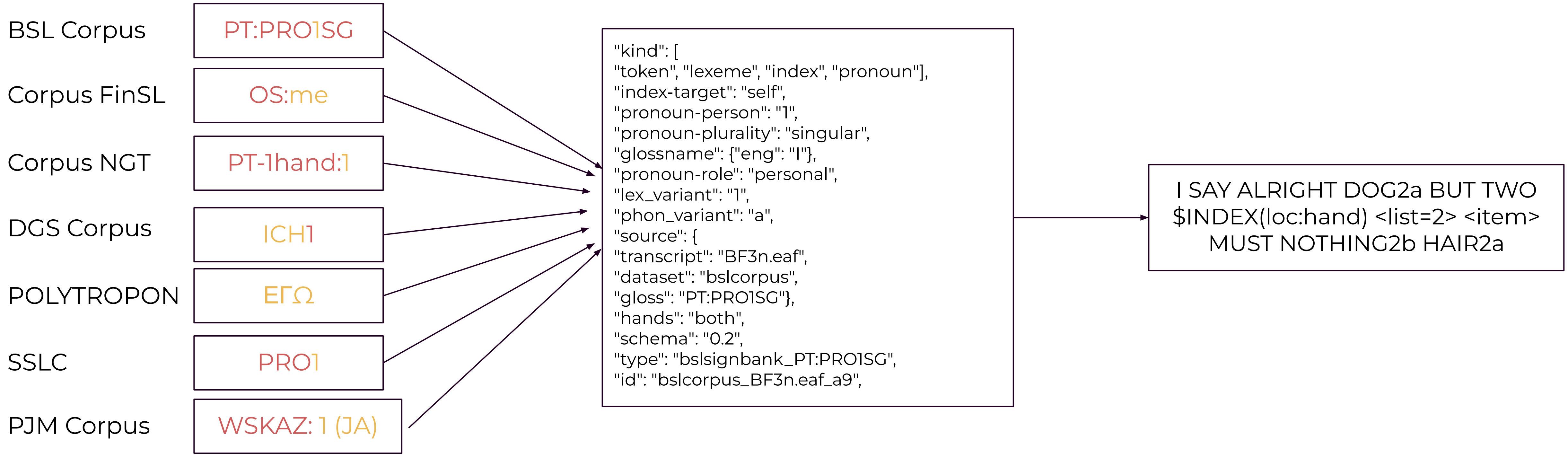
Harmonizing Corpora

Harmonising corpora

Compared available annotation conventions for over 20 aspects in 17 corpora : report by Maria Kopft

- Gloss id for common signs (pointing, fingerspelling, numbers, etc)
- Structure
  - ELAN or ilex
  - tiers structure (tiers names, separate tier for each hand, etc)
- Aspects missing (mouthings)

Corpus Annotation Harmonisation



BSL Corpus — PT:PRO1SG

Corpus FinSL — OS:me

Corpus NGT — PT-1hand:1

DGS Corpus — ICH1

POLYTROPON — ΕΓΩ

SSLC — PRO1

PJM Corpus — WSKAZ: 1 (JA)

```
"kind": [
"token", "lexeme", "index", "pronoun"],
"index-target": "self",
"pronoun-person": "1",
"pronoun-plurality": "singular",
"glossname": {"eng": "I"},
"pronoun-role": "personal",
"lex_variant": "1",
"phon_variant": "a",
"source": {
"transcript": "BF3n.eaf",
"dataset": "bslcorpus",
"gloss": "PT:PRO1SG"},
"hands": "both",
"schema": "0.2",
"type": "bslsignbank_PT:PRO1SG",
"id": "bslcorpus_BF3n.eaf_a9",
```

I SAY ALRIGHT DOG2a BUT TWO $INDEX(loc:hand) <list=2> <item> MUST NOTHING2b HAIR2a

original annotation

JSON **interchange format**
detailed format
machine readable
not very human readable

EASIER notation
more compact
human- and machine-readable
intermediate representation within the project

EASIER Notation

Signs can be more or less specified

- ○ CINEMA1A GO-THERE2A -> precise sign
- ○ CINEMA GO-THERE -> any sign with this gloss
- ○ 03032252-n GO-THERE

Can express important sign language phenomena

Examples:

- ○ <affect=smile:50><mouth=cinema> CINEMA GO-THERE
- ○ <list=3><item> FLOUR <item> MILK <item> SUGAR </list>
- ○ <ground>TABLE <figure>PLATE(loc:2) <figure>PLATE(loc:3a) <figure>PLATE(loc:3b) </ground>

Machine and human readable, human editable, cross-language

Can encode syntax, morphology, mouthings, affect, scopes and strength

Harmonizing Lexical Data using Wordnet

What’s in a SL lexicon ?

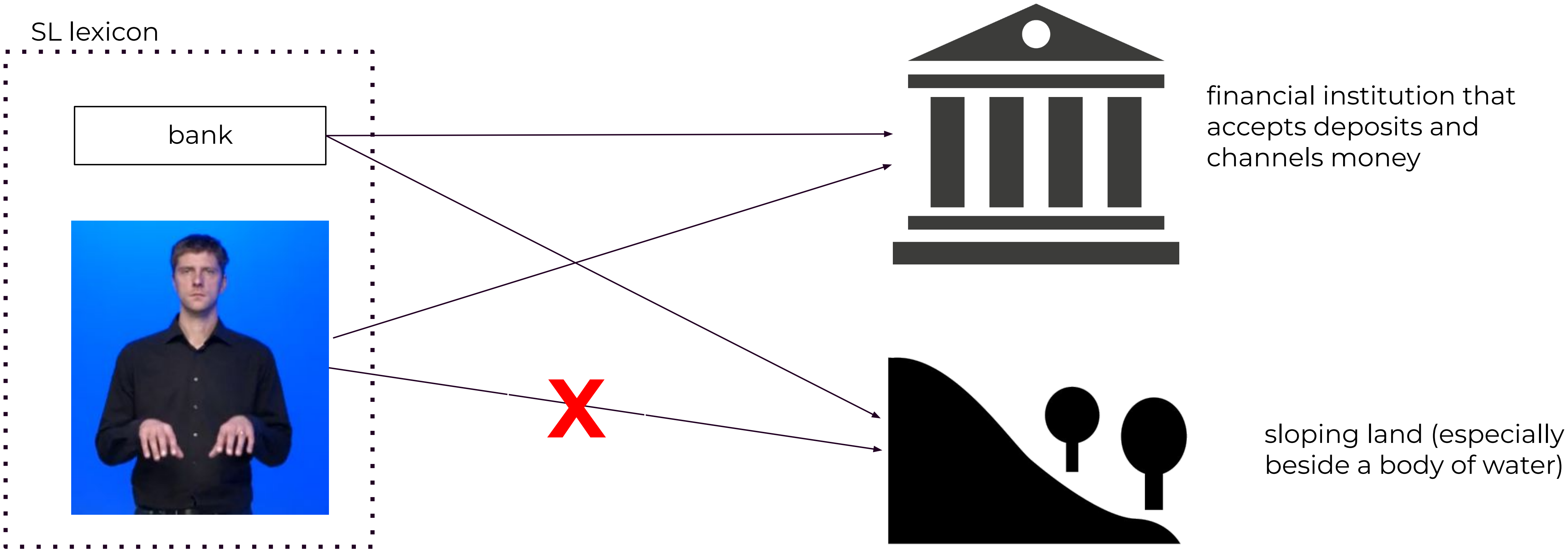
https://bslsignbank.ucl.ac.uk/dictionary/words/imagine-1.html

https://www.sign-lang.uni-hamburg.de/meinedgs/types/type13739_en.html

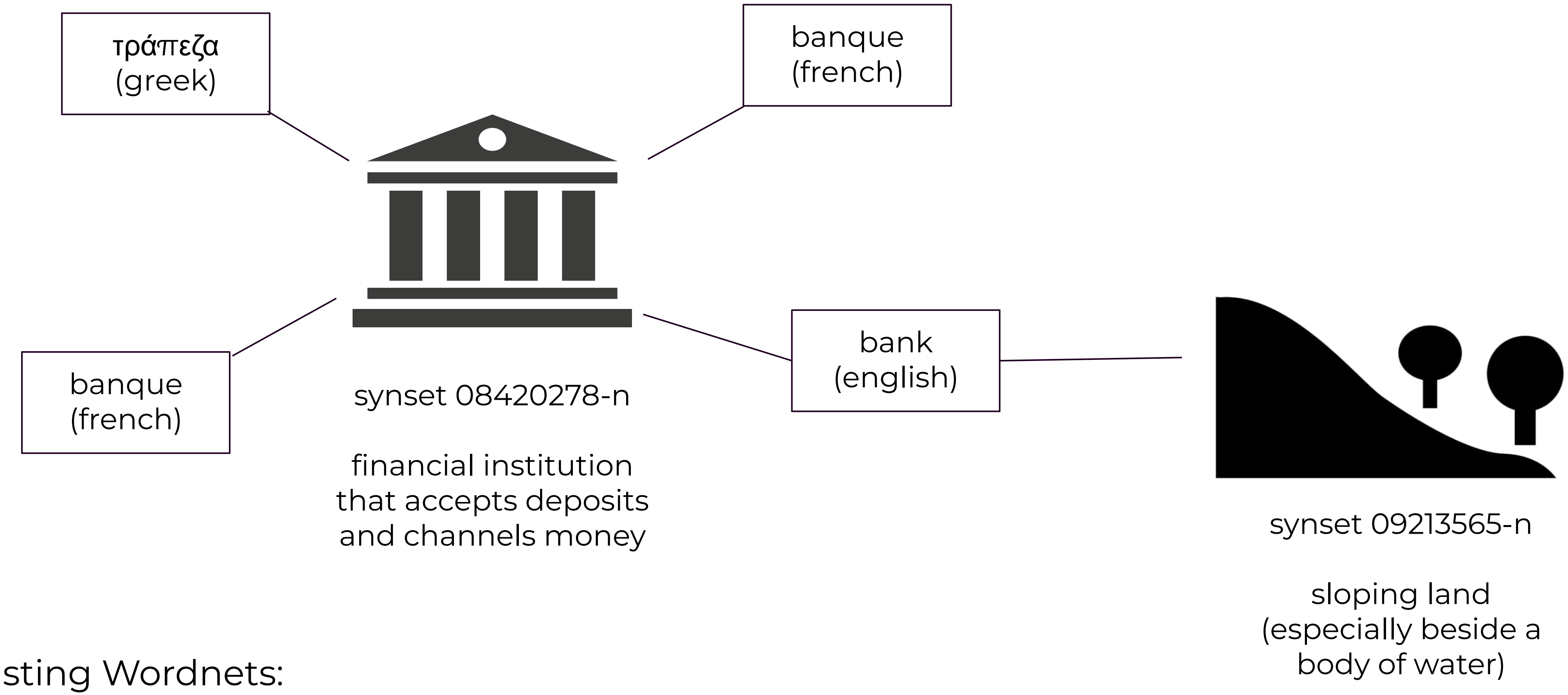
Sense typically described by gloss id / keywords in spoken language, open to mistranslation

Harmonizing lexical data

Goal: make the sense searchable across languages and disambiguated


SL lexicon
bank
X
financial institution that accepts deposits and channels money
sloping land (especially beside a body of water)

Multilingual Wordnet



τράπεζα (greek)

banque (french)

banque (french)

bank (english)

synset 08420278-n

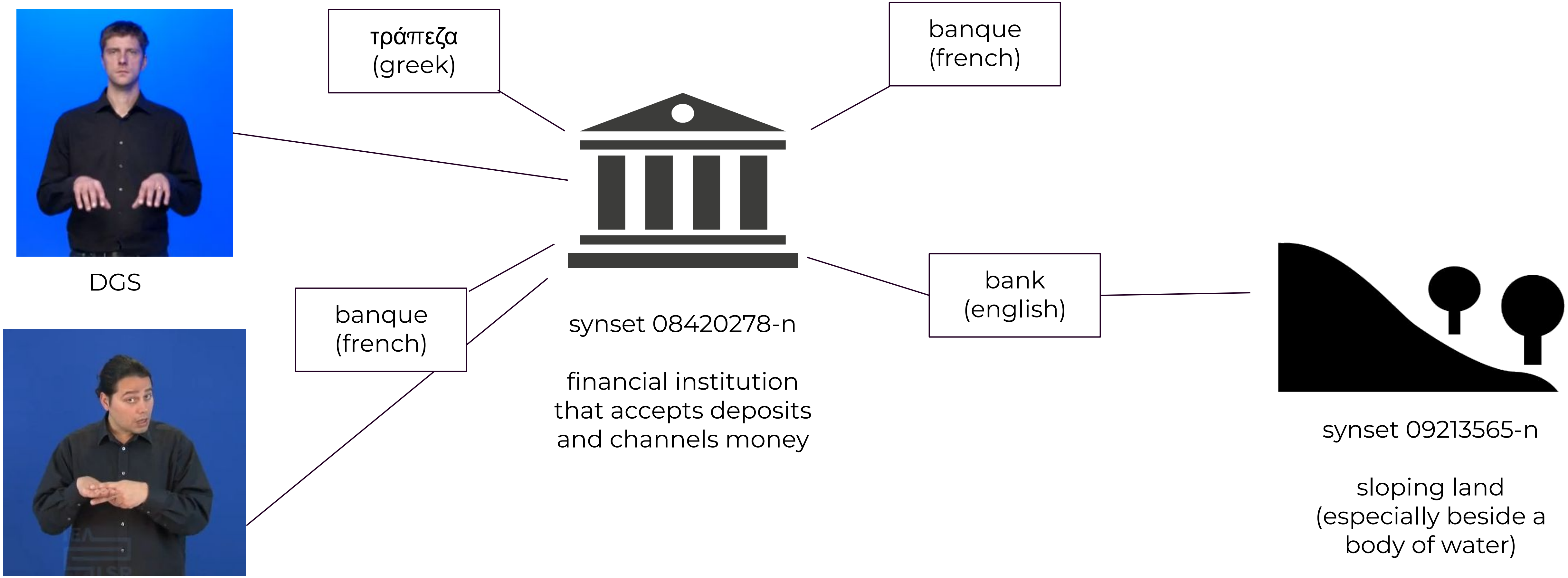
financial institution that accepts deposits and channels money

synset 09213565-n

sloping land (especially beside a body of water)

Based on existing Wordnets:
Princeton's Wordnet, Miller 1995, Fellbaum 1998
Open Multilingual Wordnet, Bond and Foster 2013

Multilingual Wordnet



DGS

GSL

τράπεζα
(greek)

banque
(french)

banque
(french)

bank
(english)

synset 08420278-n

financial institution
that accepts deposits
and channels money

synset 09213565-n

sloping land
(especially beside a
body of water)

Traps

- Wordnet/OMW properties working against us
  - too fine-grained meanings: more manual work. freeze - go
  - correct meanings absent (grammatical words) me
  - low quality wordnet words in other languages (auto-translation)
    - “lumière”: having relatively few calories. WOLF Sagot and Darja 2008

- variability in annotators’ acceptability judgement
  - dialect variation
  - +- strict
  - not enough annotators to double annotate

Need to define and limit the tool’s purpose

Purpose and limits

The goal defines design choices

- Sense inventory and granularity
- Precision/recall
- Unusual meanings (domain-specific, named entities, dated, slang...)
- Semantic links: Wordnet VS ontology VS dictionary
- etc

Wordnet of SLs' purpose

Usage:

- Be used in machine translation in everyday settings
- Link lexicons by sense from several languages

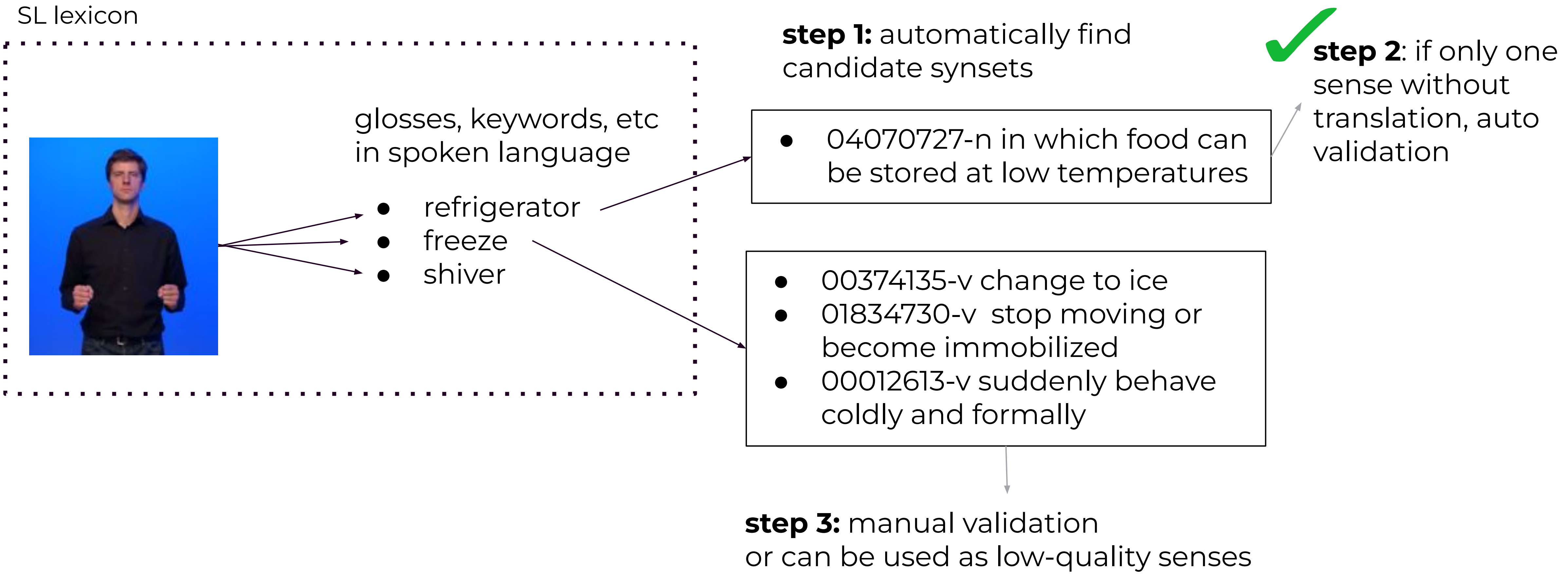
Functions:

- Sense indexation
- Disambiguation
- Simple word-by-word translation

Design choices

- Sense inventory and granularity
  - -> Only need enough precision to avoid mistranslation
- Precision/recall
  - -> Better to show a partially wrong translation than nothing, but should have a warning/confidence displayed
- Unusual meanings
  - -> Only some can be encountered in an everyday setting
- Importance of semantic links
  - -> Hypernyms can be used as alternative

Process: semi-automatic word matching



SL lexicon

glosses, keywords, etc in spoken language

- refrigerator
- freeze
- shiver

step 1: automatically find candidate synsets

- 04070727-n in which food can be stored at low temperatures

- 00374135-v change to ice
- 01834730-v stop moving or become immobilized
- 00012613-v suddenly behave coldly and formally

✓ **step 2**: if only one sense without translation, auto validation

step 3: manual validation
or can be used as low-quality senses

Manual validation

Not enough native signers. How to make the job easier and faster?

- Use non-native signers
  - corpus examples
  - good-enough signer to know what they don’t know
  - keep natives for what’s left
- Interface: Use it yourself. Be responsive
- Give examples
- Give only the best candidates, remove noise, but:
- Giving low-quality candidates is better than nothing: easier to decide if a sense is correct or not than imagining them.
- Sort senses: used in other SLs ; frequency

Manual validation

Excluded unlikely senses:

- named entities except if only one sense
- domain-specific Ex: “reflection”
  - “The phenomenon of a propagating wave (light or sound) being thrown back from a surface” domain: optics, physics
  - “The image of something as reflected by a mirror (or other reflective material)” normal sense

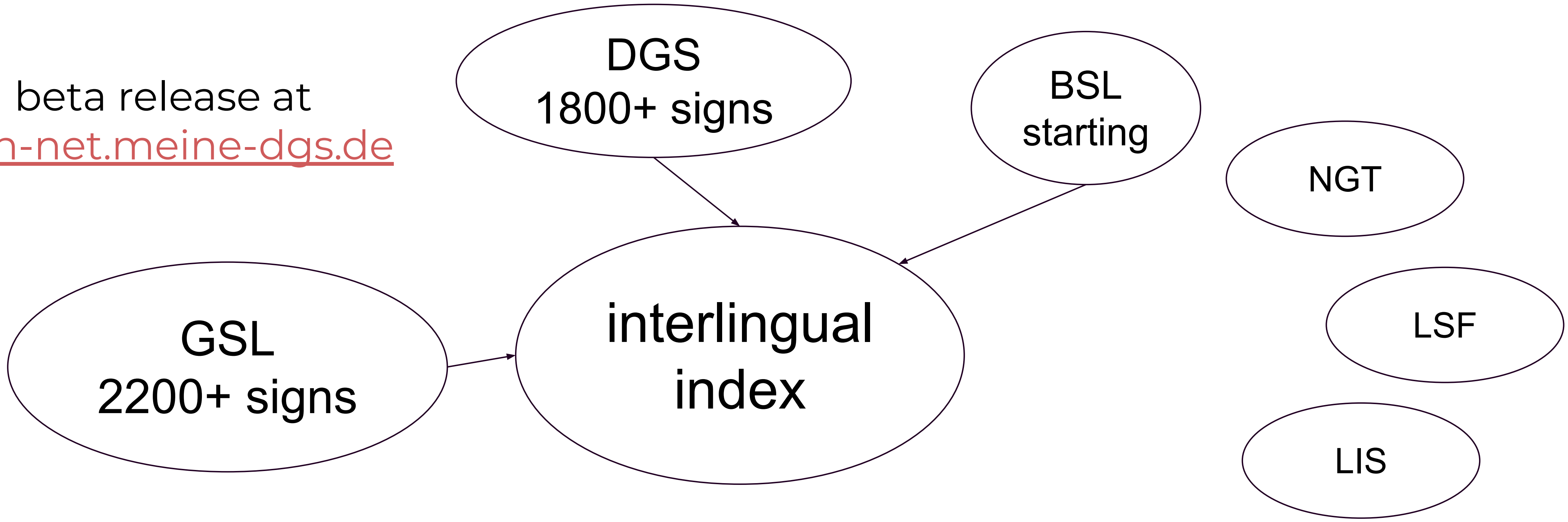
New synsets

created synsets for common signs:

- question words
- pronouns
- basic pointing

Current state

beta release at
sign-net.meine-dgs.de



Perspectives

- Use corpus for word embeddings
- Cluster similar meanings
- Recognise identical forms, to index per form

easier

intElligent Automatic
SIgn languagE tRanslation

THANKS

EASIER project is funded by the EU's Horizon2020 Programme under Grant Agreement number 101016982

PROJECT-EASIER.EU

@EASIERPROJECT