# GeoQuery-LSFB: A French Belgian Sign Language Corpus with Procedural Semantic Annotations

**Liesbet De Vos***, **Laurence Meurant**‡, **Paul Van Eecke**†◇, **Katrien Beuls***◇

*Faculté d'Informatique, Université de Namur, Namur, Belgium
{liesbet.devos, katrien.beuls}@unamur.be

†Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium
paul@ai.vub.ac.be

‡NaLTT & LSFB-Lab,Université de Namur, Namur, Belgium
laurence.meurant@unamur.be

**Abstract**

Procedural semantic representations describe the meaning of natural language expressions in terms of computer programs that can be evaluated against images, databases, knowledge graphs or other external resources. While resources annotated with procedural semantic representations already exist for a variety of spoken languages, such resources are still lacking entirely for signed languages. In this paper, we introduce GeoQuery-LSFB as a signed language extension to the multilingual GeoQuery corpus. Concretely, we have complemented each procedural semantic annotation from the original corpus with a corresponding French Belgian Sign Language (LSFB) expression that was phonetically transcribed from video recordings following the HamNoSys convention and annotated with French ID-glosses. The GeoQuery-LSFB corpus constitutes a new resource for a low-resource language and offers for the first time the possibility to study, from an onomasialogical perspective, a signed language along a diverse variety of spoken languages.

**Keywords:** datasets for low resource languages, French Belgian Sign Language, multilingual corpora, NLP datasets, procedural semantics

## 1. Introduction

Procedural semantic representations describe the meaning of natural language expressions in terms of computer programs that are compositionally structured and can be evaluated by a machine (Woods, 1967; Johnson-Laird, 1977; Woods, 1981; Winograd, 1972; Woods, 2010). A defining property of procedural semantics is that the evaluation of semantic representations involves their grounding in some kind of 'world model'. Such a model can range from a database or knowledge graph, through a quantitative or qualitative simulation, to the real world as perceived through a robot's sensory system. The grounded and compositional nature of procedural semantic representations is of great interest to a variety of natural language understanding tasks, including database querying (Zelle and Mooney, 1996; Kwiatkowski et al., 2010; Berant et al., 2013; Liang, 2016; Dong and Lapata, 2016; Cheng et al., 2019), visual question answering and dialogue (Andreas et al., 2016; Johnson et al., 2017b; Hudson and Manning, 2019; Verheyen et al., 2023), and robot instruction (Bollini et al., 2013; Misra et al., 2016; van Trijp et al., 2024).

While a number of resources that annotate natural language expressions with procedural semantic representations already exist today, most of these resources exclusively include English data (Hemphill et al., 1990; Zelle and Mooney, 1996; Kuhlmann et al., 2004; Zettlemoyer and Collins, 2005; Chen and Mooney, 2008; Tasse and Smith, 2008; Johnson et al., 2017a; Hudson and Manning, 2019; Nevens et al., 2024). For other spoken languages, resources that come with procedural semantic annotations exist much more scarcely, but some corpora are available for German (Jones et al., 2012; Gross et al., 2018), Chinese (Lu and Ng, 2011), Spanish, Japanese and Turkish (Wong and Mooney, 2006), Greek and Thai (Jones et al., 2012), and Indonesian, Swedish and Farsi (Susanto and Lu, 2017). All non-English resources, with the exception of Gross et al. (2018), extend the original English GeoQuery corpus (Zelle and Mooney, 1996) with translations, resulting in a parallel corpus that includes a typologically rather diverse selection of languages. When it comes to signed languages however, no corpora annotated with procedural semantic representations exist to date.

In this paper, we introduce the GeoQuery-LSFB corpus as a signed language extension to the multilingual GeoQuery corpus. Concretely, this new resource complements each procedural semantic annotation from the original corpus with a correspond-

◇Joint last authors.



*Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages*, pages 100–112

ing French Belgian Sign Language (LSFB) expression. Based on video recordings that feature a native LSFB signer, the expressions were annotated with French ID-glosses (Johnston, 2008, 2010) and phonetically transcribed following a time-aligned, multilinear extension to the Hamburg Notation System (HamNoSys) convention (Hanke, 2004). Apart from the 250 utterances in the original parallel corpus, we also present an augmented version of the corpus that covers 4519 utterances and thereby better fits today's data-intensive processing methods.

The GeoQuery-LSFB corpus constitutes a significant contribution to the linguistic resource landscape in three main respects: first of all, it is the first resource that aligns signed language expressions with procedural semantic representations. The corpus thereby facilitates research into language processing technologies that could so far only be developed for spoken languages. Second, it adds valuable corpus data to the limited resources that are currently available to support the study of LSFB. Finally, the parallel and semantically annotated nature of the corpus offers for the first time the possibility to study, from an onomasiological perspective, a signed language along a diverse variety of spoken languages.

The GeoQuery-LSFB corpus was released under the GNU General Public License 2.0 and is available for download at https://doi.org/10.5281/zenodo.19220833.

## 2. Background and Related Work

### 2.1. French Belgian Sign Language Resources

LSFB (*Langue des signes de Belgique francophone*) is used by the deaf and hard-of-hearing community within the French-speaking Community of Belgium. The total number of LSFB users (including deaf, hearing and hard-of-hearing users) is estimated at 20,000 of which 4,600 are first language users[1].

As a sign language, LSFB is produced and comprehended through the visual-gestural modality. This means forms are produced using articulators of the upper body (e.g. hands, head, face, shoulders) and comprehended using the visual perception system (Meier, 2002, 2012). This contrasts with spoken languages, which are produced in the aural-oral modality. Although both types of languages share structural properties, they also differ in some regards. A first point of divergence is the degree of iconicity between linguistic forms and their meanings. Iconicity is present when the phonetic form of a sign or word is motivated by a perceptible property of its referent (e.g. what it looks or sounds like) (Taub, 2012). While all human languages include iconic forms, the resources provided by the visual-gestural modality for iconicity are considered to be more abundant than those offered by the aural-oral modality. Another difference is the temporal organisation of linguistic units. Spoken language is primarily organised sequentially, meaning that linguistic units follow each other in a linear order. In contrast, sign languages contain more multilinear structures in which information is organised simultaneously. Finally, sign languages make extensive use of the signing space (i.e. the space surrounding the upper body of the signer) to manage referents and their relationships.

Phylogenetically, LSFB is closely related to Vlaamse Gebarentaal (VGT), the sign language of the Flemish Community of Belgium. Both languages are historically rooted in the Old French Sign Language (VLSF) and were considered under the umbrella of Belgian Sign Language until the 1970s. LSFB received recognition as an official language in the French-speaking Community of Belgium in 2003.

While significant efforts in creating linguistic resources that can be used for LSFB research, education and language technology development are currently ongoing, LSFB for now remains a low-resource language. The most extensive resource today is the LSFB Corpus (Meurant, 2015; Meurant et al., 2016), which consists of 90 hours of unscripted yet task-moderated video-recorded conversations between pairs of signers. French ID-gloss annotations (Johnston, 2008) have been made available for 25 hours of the corpus material. An overview of all ID-glosses that were used to annotate the corpus, along with French translations and isolated video recordings of the corresponding LSFB signs, has been released as the lex-LSFB lexical database (Meurant et al., 2015).

The LSFB online dictionary (https://dico.lsfb.be) provides LSFB video translations for over 5000 French words and a tool for translating LSFB signs into French words was made available by Fink et al. (2022). The LSFB Corpus was made available to the machine learning community through the release of two datasets, one of which wraps the original corpus data (LSFB-CONT) while the other groups occurrences of the most frequently used signs (LSFB-ISOL) (Fink et al., 2021). A spoken Belgian-French counterpart to the LSFB Corpus, where participants were asked to carry out the same conversational tasks, is currently under construction (Lepeut et al., 2024).

[1]The overall number of LSFB users is estimated at 0.44% of the total population following Haeusler et al. (2014). The number of first language users is estimated at 0.1% of the total population following Pasikowska-Schnass (2018).

## 2.2. Sign Language Transcription

On a high level, sign language transcription systems can be grouped into two categories. The first category transcribes each sign within a signed utterance using a gloss, i.e. an identifier for the sign that essentially corresponds to one or more words from the ambient spoken language[2], or to a numerical code. The use of descriptive spoken language words within the label (e.g. THANK-YOU for a sign that expresses gratitude) is generally preferred over opaque numerical labels, as it helps annotators remember the correspondences between signs and their labels. It is important to note, however, that sequences of glosses should not be confused with spoken language translations, since they retain the grammatical structure of the signed language. Of particular interest are ID-glosses (Johnston, 2008, 2010), where contextually different variants of a sign, i.e. variants with the same form but a different meaning, are grouped under the same gloss. (ID-)glosses do not contain any information about the form of a sign itself and are heavily dependent on a spoken glossing language.

The second category of writing systems transcribe signed utterances phonetically or phonologically. Systems in this category do not rely on any external glossing language, but transcribe the form components of signs directly. Phonetic/phonological transcription systems for signed languages were pioneered by Stokoe (1960) for American Sign Language (ASL). The Sutton SignWriting (Sutton, 1995) and HamNoSys (Prillwitz et al., 1987; Hanke, 2004) systems are currently the most widely used language-agnostic transcription systems.

SignWriting represents manual and non-manual components of signs in 2D space using an alphabet of 652 base symbols[3]. HamNoSys provides a linear, phonetic representation of sign components using a more restricted inventory of just over 200 base symbols (Smith, 2013). More recently, the Typannot system for handshape annotation was introduced with the goal of reaching a HamNoSys-like phonetic system while achieving a SignWriting-like readability (Rébulard et al., 2018), but is still to find a more widespread adoption.

Gloss-based transcriptions and phonetic transcriptions are very different in nature, have different properties and serve different purposes. Gloss-based transcriptions are, relatively speaking, less time-consuming to annotate. Their grounding in an ambient spoken language naturally supports their use in domains where the relationship between the signed and the spoken language is of significant interest, such as (machine) translation and different forms of sign language education. Phonetic transcriptions on the other hand are a valuable resource for linguistic research as they allow for a representation of signed forms that do not need to be cast into linguistic theories that were developed for spoken languages. As noted by Hodge and Crasborn (2022), it can be considered good practice in corpus development to incorporate different levels and types of annotation. In this way, the corpus becomes multi-purpose and is thereby likely to better stand the test of time.

## 2.3. Procedural Semantics

The procedural approach to semantics was introduced in the field of artificial intelligence in the 1960s and 1970s by Woods (1967), Winograd (1972) and Johnson-Laird (1977). The innovative idea behind this approach was that the meaning of a natural language expression can take the form of a program that can be evaluated by a machine. The evaluation of such a program involves its grounding in a 'world model', i.e. an external resource such as a database, knowledge graph, sensory observation, or simulation. Let us consider an example drawn from the GeoQuery resource, which will be described in more detail in Section 2.4. The world model consists of a database of geographical facts, including the area and population of cities and states, the height of mountains, the length of rivers, and the states and cities rivers run through. Procedurally, the meaning of a question like "*How big is Texas?*" could be represented through a query that, when evaluated against the database, would return the area of the state of Texas. Importantly, the answer itself is irrelevant when it comes to representing the meaning of the question. If a new treaty is signed and the world model is consequently updated, a query that accurately captured the meaning of the question would still return the correct answer, even if the answer is no longer the same. Likewise, a program that adequately represents the meaning of the statement "*The Mississippi river runs through Iowa*" should evaluate to true in every world where Iowa is among the states the Mississippi river runs through, and evaluate to false in any other world.

The procedural approach to representing meaning subscribes to the Davidsonian view that the meaning of an utterance corresponds to its truth conditions (Davidson, 1967). A complete understanding of an utterance amounts to being able to determine its truth value in every possible world, a task that requires an exact knowledge of the conditions under which the utterance is true. In terms of procedural semantics, the programs that represent

[2]The spoken language predominantly used within the same geographic region as the relevant signed language.

[3]information retrieved from the Sutton SignWriting computer manual: https://www.signwriting.org/archive/docs7/sw0636_SignWriting_Alphabet_Manual_2010.pdf

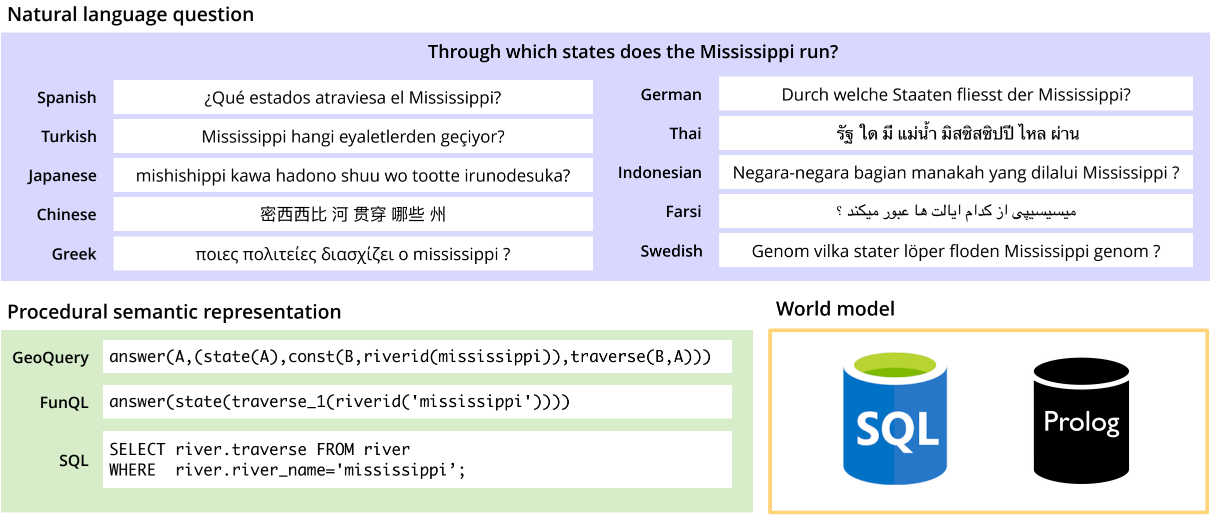


Figure 1: Schematic representation of a single entry in the multilingual GeoQuery corpus, featuring the original English question, its translation into ten other spoken languages, its procedural semantic representation in the GeoQuery, FunQL and SQL languages, and the relational and logic databases that constitute the world model

the meaning of natural language utterances thus correspond to their truth conditions. The process of evaluation then corresponds to the task of determining the truth of an utterance in a particular world[4]. In general, there are many ways in which procedural semantic programs can be formalised and in which their evaluation can be implemented, and the choice will typically depend on the task at hand and the nature of the world model that is provided (see e.g. Verheyen et al., 2023).

Despite the terminology that is used, procedural semantic programs almost never take the form of a sequence of operations to execute. Instead, they state the logic underlying the computation they are supposed to represent, for example through a conjunction of predicates. As long as a procedural interpretation can be linked to their logic interpretation, the programs are considered procedural semantic representations. In this paper and the associated GeoQuery-LSFB corpus, all procedural semantic representations will take the form of logic expressions that correspond to machine-evaluatable database queries.

The ability to evaluate procedural semantic representations against an external source sets them apart from the declarative, non-executable representations used in other sign language applications, such as the Discourse Representation Structures (DRS) used by Marshall and Sáfár (2004), Elliott et al. (2008) and Sáfár and Glauert (2010) or the logical structures from Role and Reference grammar (RRG), used by Murtagh (2011) and Murtagh et al. (2022).

[4]In Fregean terminology, the programs would correspond to the *Sinn* of an utterance, while the evaluation process would reveal its *Bedeuting* (Frege, 1892).

## 2.4. The GeoQuery Corpus

The GeoQuery corpus was originally introduced by Zelle (1995) and Zelle and Mooney (1996) as a benchmark dataset for the supervised learning of semantic parsers. In its original form, it consisted of 250 English questions[5] about US geography that were collected from undergraduate students at the Department of Computer Sciences of the University of Texas at Austin. Each question was annotated by a human expert with a logical query that would answer the question upon evaluation against the GeoBase database of US state geography (Borland International, 1988). The corpus was extended to 880 question-query pairs by Tang and Mooney (2001) and Tang (2003) using data collected from users of the GeoQuery web interface.

The logical queries annotated in the original dataset straightforwardly integrate with logic programming languages such as Prolog. In order to better accommodate semantic parsers written in other programming languages, Kate et al. (2005) introduced FunQL, a functional, variable-free query language for the GeoQuery domain. A set of scripts allows for automatic conversion between the annotated logical queries and their FunQL counterparts (Kate et al., 2005). Popescu et al. (2003) and Iyer et al. (2017) introduced a relational database

[5]We refer to the GeoQuery utterances as *questions* as they constitute requests for information to be answered. Grammatically, these requests can be expressed through interrogative sentences (e.g. '*What is the city with the smallest population?*') but also through imperative sentences (e.g. '*Give me the biggest city in Wisconsin*').

schema for the Geobase database and manually annotated all GeoQuery questions with SQL queries.

The first multilingual version of the GeoQuery corpus was introduced by Wong and Mooney (2006), who provided Spanish, Japanese and Turkish translations of the original GeoQuery questions. Later, translations of all questions in the extended GeoQuery dataset were added for many other languages, in particular for Chinese by Lu and Ng (2011), for German, Greek and Thai by Jones et al. (2012) and for Indonesian, Farsi and Swedish by Susanto and Lu (2017).

Figure 1 shows a single entry in the multilingual GeoQuery corpus. The original English question "*Through which states does the Mississippi run?*" is shown on top along with its translation into the 10 other languages of the dataset. The procedural semantic representations that correspond to the questions are shown in the bottom-left part of the figure. Note that the logical query and the FunQL query are equivalent, but that the SQL query was constructed independently. A schematic representation of the 'world models' is shown in the bottom-right part of the figure. The SQL meaning representation can be evaluated against the relational database, while the GeoQuery and FunQL meaning representations can be evaluated against the Prolog factbase.

## 3. Data Collection and Transcription

The overall corpus creation task consisted of annotating the 250 utterances of the original GeoQuery corpus with corresponding LSFB expressions. Given that LSFB is not a written language, one cannot simply ask a native signer to write down LSFB translations of the English sentences. Our native LSFB informant was however experienced in signing LSFB expressions that semantically correspond to utterances written down in the ambient French language. The first stage in the corpus creation process therefore consisted of making video recordings of LSFB expressions. These recordings could then be transcribed both phonetically and in terms of ID-glosses by an expert transcriber in a second stage.

In order to optimise the workload of the LSFB informant and transcriber, the first challenge was to select the minimal number of expressions that needed to be signed, video-recorded and transcribed. We started from the 221 unique meaning representations in the corpus and reduced them to 95 schemata by replacing named entities by their higher-level entity type (`<state>`, `<river>`, `<city>` or `<capital>`). For each of these schemata, the first corresponding utterance from the GeoQuery dataset was translated into French and provided to the signer. The signer was instructed to interpret the meaning of the French utterance and sign a corresponding LSFB expression. For meaning representations that occurred multiple times in the corpus with the exact same named entities, the signer was asked to sign the same number of variants. The signed utterances were video-recorded, yielding a total of 124 recordings.

The expert transcriber annotated the recordings first with French ID-glosses and then with HamNoSys-based phonetic transcriptions. ID-gloss annotation layers were created for the left and right hands using glosses from the lex-LSFB lexical database (Meurant, 2015). During the annotation process, five new ID-glosses were created for namesigns that do not occur in the database (NS:CALIFORNIE, NS:FLORIDE, NS:TEXAS, NS:DES-MOINES, and NS:MEXIQUE), as well as four new regular ID-glosses (KILOMETRE-CARRE, TAILLE, ME-DIRE and SERRER). Depicting and pointing signs were annotated following the guidelines established by Johnston (2010, 2016). Fingerspelled named entities were transcribed as such and variants of existing ID-glosses were tagged following Johnston (2010)'s handshape categorisation. Phonetic transcriptions were made using a time-aligned, multilinear extension to the HamNoSys convention. Concretely, a separate HamNoSys annotation layer was added for each hand. The segmentation of the layers was copied from the corresponding ID-gloss layers, temporally aligning the HamNoSys and ID-gloss transcriptions. Two-handed signs, which HamNoSys transcribes using a single linear expression, are annotated in the layer for the dominant hand, in which case the corresponding segment in the layer for the other hand is left empty.

An example transcription is shown in Figure 2 for the original utterance "*What are the high points of the states surrounding Mississippi?*". The annotation layers associated to the signer's dominant right hand are divided into eight time-aligned segments, one for each ID-gloss that was identified. The signer opens with the PALM-UP sign, a discourse marker that indicates that the speaker is thinking (Gabarró-López, 2017). This sign is followed by the DANS sign (English: IN) and the fingerspelled proper name Mississippi (FS:MISSISSIPPI, where FS stands for 'fingerspelling'). Then, the IL-Y-A sign is annotated (English: THERE-IS), followed by a depicting sign DS(BENT5):ETAT+, where DS stands for 'depicting sign', BENT5 for the handshape categorised according to Johnston (2016) and + signals that the sign is repeated multiple times. Finally, the pointing sign PT:DET/LOC(1)+ is annotated, where PT stands for 'pointing sign', DET/LOC for 'determiner or location', 1 for the used handshape and + for repetition, followed by the ID-glosses HAUT (English: HIGH) and QUOI (English: WHAT). Note that

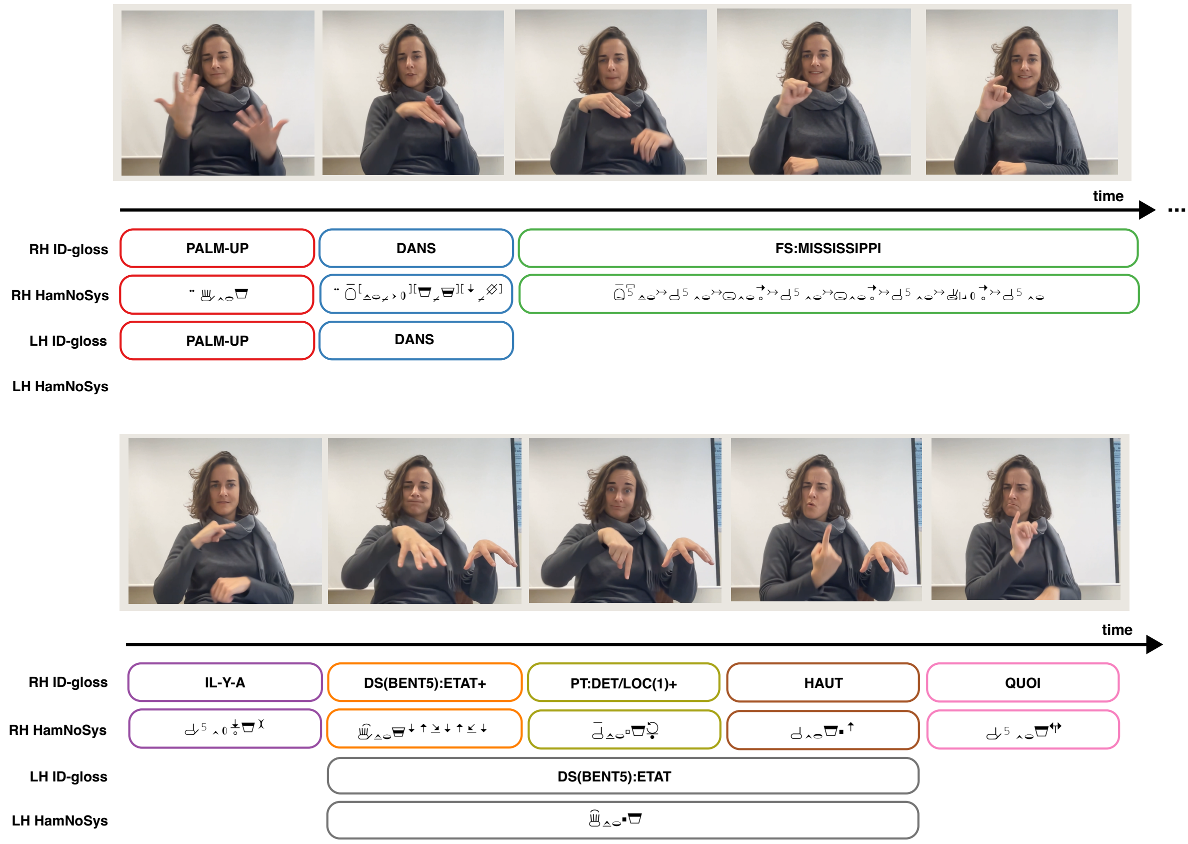


Figure 2: An example transcription of a video fragment in which the LSFB expression that corresponds to the English utterance *What are the high points of the states surrounding Mississippi?* is signed. Tiers for right-hand ID-gloss, right-hand HamNoSys transcription, left-hand ID-gloss and left-hand HamNoSys transcription are shown

the circular motion of the pointing sign is included in its HamNoSys transcription, but is reduced to 'repetition' in the ID-gloss annotation.

The annotation layers for the left hand hold three segments, two of which are ID-gloss annotations for two-handed signs (PALM-UP and DANS). For these segments, the HamNoSys transcription layer for the left hand remains empty, as the symmetry operator ¨ is used in the corresponding right-hand HamNoSys segment. The third left-hand segment concerns a depicting sign referring to a state-like entity that is again signed using Johnston (2016)'s handshape BENT5 and which coincides with three right-hand segments.

After the transcription of the 124 video recordings was completed, the resulting annotation layers were extrapolated to the 250 utterances of the original corpus. This could be done straightforwardly by substituting the named entity segments that were annotated in the recordings with segments containing the namesigns or finger-spelled transliterations of the named entities that occur at the same location in the remaining utterances. A separate, larger corpus of 4519 annotated utterances was also created by extrapolating the transcriptions of the video recordings to all named entities in the GeoBase database, enforcing the constraint that states, rivers, cities and capital cities can only be substituted with an entity of the same type.

As indicated above, the French ID-gloss annotations and HamNoSys transcriptions were contributed by a single expert transcriber. As a result, no extensive inter-annotator agreement study could be performed. While we acknowledge this as a limitation of the resource that we created, we would very much like to emphasize the difficulty of recruiting and training annotators for a low-resource sign language like LSFB. The annotator that was recruited is to the best of our knowledge the only person sufficiently proficient in both LSFB and HamNoSys to perform this task reliably. When it comes to the French ID-gloss annotations, a small-scale inter-annotator agreement study was performed in the pilot phase of the corpus creation process in order to validate its feasibility. This pilot study yielded a Cohen's kappa of 0.83 on 10 corpus entries selected to be maximally different. This result led to the green light needed to initiate the corpus creation process, but cannot be reported as such as a measure of the quality of the final resource.

All annotations were created using the ELAN software toolkit (Crasborn and Sloetjes, 2008)[6], the HamNoSys input palette[7] and the CWA Signing Avatars SiGML Player[8].

## 4. The GeoQuery-LSFB Corpus Release

The GeoQuery-LSFB corpus was released under the GNU General Public License 2.0 and can be downloaded from https://doi.org/10.5281/zenodo.19220833. The release consists of the following parts:

**Documentation** A user guide that documents the contents of the release, especially with respect to using the corpus for research purposes. The documentation provides a detailed description of all aspects of the corpus, including the required installation procedures, the labels used during the transcription process and the data structures that were used.

**Video recordings** All 124 video recordings, encoded using MPEG-4 compression. The videos amount to a total length of 21 minutes and 44 seconds, which corresponds to an average of 10.35 seconds per video.

**HamNoSys font** A version of the Menlo font that was extended with the HamNoSys character set[9].

**Annotation files** The 124 raw ELAN annotation files in `.eaf` format, as well as their corresponding configuration files in `.psfx` format.

**Corpus files** The actual corpus files, which hold for each entry:

- a unique identifier,
- the original English utterance,
- a reference to the higher-level semantic schema of which the meaning of the utterance is an instantiation,
- a reference to the corresponding ELAN file and video recording (where applicable),
- translations of the utterance into Spanish, Turkish, Japanese (Romaji), German, Greek, Thai, Indonesian, Swedish, Chinese and Farsi,
- the French translation of the utterance that was provided to the LSFB signer,
- procedural semantic representations of the utterance in the GeoQuery, FunQL and SQL formats,
- our integrated representation of the ID-gloss and HamNoSys annotations of the corresponding LSFB expression.

Apart from the unique identifiers and references to the higher-level semantic schemata, ELAN files and video recordings, we contributed the French and LSFB annotations and adopted all other annotations from prior versions of the corpus as documented in Section 2.4 above. The corpus files are provided in both the JSON and XML formats, with both versions containing the exact same information.

**Database files** Scripts for executing and evaluating the procedural semantic representations of the utterances in GeoQuery, FunQL and SQL formats, including Iyer et al. (2017)'s script for creating and populating a relational database. Instructions on how to use the execution and evaluation scripts are included in the user guide.

Overall, the transcription layers of the GeoQuery-LSFB corpus for the dominant right hand contain 2681 segments, while those for the left hand hold 1258 segments. 2794 segments (70.93%) were annotated with lexicalised ID-glosses. 143 distinct lexicalised ID-glosses occur, with the most frequent ones being either generally frequent in LSFB, such as UN (English: A/ONE) and IL-Y-A (English: THERE-IS), or being related to the domain of the dataset (e.g.VILLE (English: CITY) and RIVIERE (English: RIVER)). 595 segments (15.11%) were annotated with pointing sign glosses and 380 segments (9.65%) were annotated with depicting sign glosses. The remaining 170 segments (4.32%) were annotated as finger-spelled proper names.

## 5. Potential Uses of the Corpus

The GeoQuery-LSFB corpus can serve various use cases within the fields of NLP and linguistics.

In NLP, the corpus facilitates a range of language processing tasks, such as semantic parsing, database information retrieval, sign language recognition and machine translation. Currently, the performance of multilingual semantic parsers has exclusively been tested on spoken language data.

[6] Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands: https://archive.mpi.nl/tla/elan

[7] https://www.sign-lang.uni-hamburg.de/hamnosys/input/

[8] https://vhg.cmp.uea.ac.uk/tech/jas/std/

[9] Original HamNoSys character set was retrieved from: https://doi.org/10.25592/uhhfdm.9725

The GeoQuery-LSFB corpus provides the first semantic parsing benchmark which combines a variety of spoken languages with a signed language, allowing multilingual parsers to be trained and tested on a set of languages with heterogeneously diverse forms. The developed semantic parsers can be used within natural language database interfaces, which allow users to retrieve information from a database using natural language, increasing accessibility to these databases (Copestake and Jones, 1990). Given the ID-gloss and HamNoSys annotations provided, the corpus can also be used to train sign language recognition models. Such models typically map a sign language video to a single ID-gloss or series of ID-glosses (Camgoz et al., 2020). The inclusion of HamNoSys also allows the development of models that map video to HamNoSys representations. Finally, the parallel nature of the corpus is ideal for machine translation between a signed language and multiple spoken languages.

Within the field of linguistics, the corpus's main use is the onomasiological study of a signed language along with a diverse set of spoken languages. Onomasiology is a branch from linguistics which analyses how different languages express the same concept (Fernández-Domínguez, 2019). The semantic representations and multiple languages in the corpus are ideal for this, as they allow researchers to study how one component of the semantic representation is expressed in each of the corpus' languages. This can give insight into the influence of the visual-gestural versus aural-oral modality on language variation. Additionally, the combination of signed forms and semantic annotations in GeoQuery-LSFB increases the limited amount of semantically annotated data that is available for computational construction grammar development. Computational construction grammar operationalises the core ideas of construction grammar and analyses language through its form-meaning mappings (e.g. Steels 2011; Beuls and Van Eecke 2023, Van Eecke and Beuls 2025). While several computational construction grammars have been engineered (e.g. Nevens et al., 2019; Verheyen et al., 2023) and learned (Nevens et al., 2022; Doumen et al., 2023, 2024) for spoken languages, similar resources for signed languages are currently limited (e.g. van Trijp 2015; De Vos et al. 2025). GeoQuery-LSFB can help fill this gap by providing the first semantically annotated corpus for a signed language.

## 6. Discussion and Conclusions

We have presented the GeoQuery-LSFB corpus as a linguistic resource for the low-resource French Belgian Sign Language (LSFB). Concretely, we have extended the multilingual and semantically annotated GeoQuery corpus with transcriptions of corresponding LSFB expressions. Based on video recordings featuring a native LSFB signer, an expert transcriber has annotated each expression from the original corpus with French ID-glosses (following Meurant et al., 2015; Johnston, 2016) and phonetic HamNoSys transcriptions (following Hanke, 2004). The corpus was released under the GNU General Public License 2.0 and includes (i) an extensive user guide, (ii) the video recordings that were made, (iii) the original English corpus data and procedural semantic annotations, (iv) the multilingual versions of the corpus and FunQL and SQL annotations that were previously contributed by different researchers, (v) our LSFB annotations in terms of French ID-glosses and HamNoSys transcriptions, and (vi) scripts for executing and evaluating the procedural semantic representations. A synthetically augmented version of the corpus was also included in order to meet user demand.

The GeoQuery-LSFB corpus constitutes a contribution to the linguistic resource landscape in three main respects. First of all, the corpus contributes valuable data to the limited pool of resources that are currently available for the linguistic study of LSFB, as well as for the development of LSFB language technologies. Such data is challenging to create, not only for financial reasons, but also given the limited number of people that both sufficiently master LSFB and have experience or interest in the creation and management of linguistic resources. Second, the GeoQuery-LSFB corpus is, to the best of our knowledge, the first corpus that aligns signed utterances with procedural semantic annotations. The corpus thereby facilitates research into language processing technologies, such as data-efficient semantic parsers or natural language interfaces, that could so far only be developed for spoken languages. Finally, the parallel and semantically annotated nature of the corpus offers for the first time the possibility to study, from an onomasiological perspective, a signed language along a diverse variety of spoken languages.

## 7. Limitations

**Single annotator** The LSFB transcriptions and annotations were performed by a single expert transcriber. We acknowledge this as a limitation of the resource that we created. Yet, we would like to emphasize the difficulty of recruiting and training transcribers for a low-resource sign language like LSFB. We were not able to identify a second potential transcriber sufficiently proficient in both LSFB and HamNoSys to perform the transcription task. Regarding the French ID-gloss annotations, a small-scale inter-annotator agreement study was performed in the pilot phase of the corpus creation process in order to validate its feasibility. This pilot study yielded a Cohen's kappa of 0.83 on 10 corpus entries selected to be maximally different. This result led to the green light needed to initiate the corpus creation process, but the final corpus was annotated by only one of the transcribers that was part of the pilot due to a lack of available human resources.

**Limited size and domain of the corpus** With its 250 and 4519 utterances, the resource that we introduce is limited in volume and restricted to a single domain. As such, it might not be suitable for many data-intensive processing methods and should not be taken as a representative sample of language use. At the same time, it is the first resource of its kind for LSFB, and for any signed language for that matter. Given the extremely limited resources available for the low-resource LSFB language, the corpus will find immediate use in LSFB-related research and to the best of our expectations in sign language research more generally.

## 8. Statements and Declarations

**Funding** This research was partially supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche).

**Ethics statement** The French Belgian Sign Language (LSFB) informant consented to the publication and distribution of the video recordings that are part of the GeoQuery-LSFB resource. Both the LSFB informant and the expert transcriber were remunerated for their work according to the official pay scales for scientific personnel employed by universities of the French community of Belgium.

**Competing interests** The authors have no competing interests to declare that are relevant to the content of this article.

**Author Contributions** All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Liesbet De Vos and Laurance Meurant. The original manuscript was written by Katrien Beuls and Paul Van Eecke, with contributions by Liesbet De Vos. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

**Use of Generative AI** The authors declare that this article was conceived and written without the use of generative AI tools.

## 9. Bibliographical References


Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Katrien Beuls and Paul Van Eecke. 2023. Fluid Construction Grammar: State of the art and future outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 41–50. Association for Computational Linguistics.

Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2013. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, pages 481–495. Springer.

Borland International. 1988. *Turbo Prolog 2.0 Reference Guide*. Borland International, Scotts Valley, CA, USA.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*, pages 128–135.

Jianpeng Cheng, Siva Reddy, Vijay Saraswat, and Mirella Lapata. 2019. Learning an executable neural semantic parser. *Computational Linguistics*, 45(1):59–94.

Ann Copestake and Karen Sparck Jones. 1990. Natural language interfaces to databases. *The Knowledge Engineering Review*, 5(4):225–249.

Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43. European Language Resources Association (ELRA).

Donald Davidson. 1967. Truth and meaning. In Jack Kulas, James H. Fetzer, and Terry L. Rankin, editors, *Philosophy, Language, and Artificial Intelligence: Resources for Processing Natural Language*, pages 93–111. Springer.

Liesbet De Vos, Paul Van Eecke, and Katrien Beuls. 2025. A computational construction grammar framework for modelling signed languages. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 1–12, Düsseldorf, Germany. Association for Computational Linguistics.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33–43. Association for Computational Linguistics.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2023. Modelling language acquisition through syntactico-semantic pattern finding. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1317–1327.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2024. Modelling constructivist language acquisition through syntactico-semantic pattern finding. *Royal Society Open Science*, 11(7):231998.

Ralph Elliott, John RW Glauert, JR Kennaway, Ian Marshall, and Eva Safar. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6:375–391.

Jesús Fernández-Domínguez. 2019. The onomasiological approach. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, UK.

Jérôme Fink, Pierre Poitier, Maxime André, Laurence Meurant, Benoît Frénay, and Anthony Cleve. 2022. Dictionnaire contextuel langue des signes belge francophone vers français.

Jérôme Fink, Benoît Frénay, Laurence Meurant, and Anthony Cleve. 2021. LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Gottlob Frege. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.

Sílvia Gabarró-López. 2017. *Discourse Markers in French Belgian Sign Language (LSFB) and Catalan Sign Language (LSC): BUOYS, PALM-UP and SAME*. Ph.D. thesis, Faculty of Arts, Université de Namur, Namur, Belgium.

Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth, and Michael Zillich. 2018. Action Verb Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Laurence Haeusler, Thibaud de Laval, and Charlotte Millot. 2014. Etude quantitative sur le handicap auditif à partir de l'enquête "handicap-santé". Technical Report 131, Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), Paris, France.

Thomas Hanke. 2004. HamNoSys: representing sign language data in language resources and language processing contexts. In *Proceedings of the fourth international Conference on Language Resources and Evaluation (LREC)*, volume 4, pages 1–6. European Language Resources Association (ELRA).

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop held at Hidden Valley*. Morgan Kaufmann.

Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In Trevor Johnston, Julie A. Hochgesang, and Jordan Fenlon, editors, *Signed Language Corpora*, pages 46–89. Gallaudet University Press, Washington, D.C., USA.

Drew A Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017b. Inferring and executing programs for visual reasoning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2989–2998, Washington, D.C., USA. IEEE Computer Society.

Philip N. Johnson-Laird. 1977. Procedural semantics. *Cognition*, 5(3):189–214.

Trevor Johnston. 2008. Corpus linguistics and signed languages: no lemmata, no corpus. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 82–87. European Language Resources Association (ELRA).

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2016. *Auslan Corpus Annotation Guidelines*.

Bevan Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with Bayesian tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496. Association for Computational Linguistics.

Rohit J. Kate, Yuk Wah Wong, and Raymond J. Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings of the 20th National Conference on Artificial Intelligence-Volume 3*, pages 1062–1068.

Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer. In *The AAAI-2004 Workshop on Supervisory Control of Learning and Adaptive Systems*. San Jose, CA.

Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233. Association for Computational Linguistics.

Alysson Lepeut, Clara Lombart, Sébastien Vandenitte, and Laurence Meurant. 2024. Spoken and signed languages hand in hand: parallel and directly comparable corpora of French Belgian Sign Language (LSFB) and French. *Corpora*, 19(2):241–253.

Percy Liang. 2016. Learning executable semantic parsers for natural language understanding. *Communications of the ACM*, 59(9):68–76.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622. Association for Computational Linguistics.

Ian Marshall and Éva Sáfár. 2004. Sign language generation in an ALE HPSG. In *11th International Conference on Head-Driven Phrase Structure Grammar*, pages 189–201, Stanford, CA. CSLI Publications.

Richard P. Meier. 2002. Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In Richard P. Meier, Cormier Kearsy, and Quinto-Pozos David, editors, *Modality and structure in signed and spoken languages*, pages 1–26. Cambridge University Press, Cambridge, United Kingdom.

Richard P. Meier. 2012. Language and modality. In Roland Pfau, Marcus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, volume 37 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 574–601. De Gruyter Mouton, Berlin, Germany.

Laurence Meurant. 2015. *Corpus LSFB. First digital open access corpus of movies and annotations of French Belgian Sign Language (LSFB)*. LSFB-Lab, University of Namur.

Laurence Meurant, Christophe De Clerck, Sibylle Fonze, Gauthier Raes, Susana Sanchez, and Éric Bernagou. 2015. Lex-LSFB: online lexical database from and linked to the 'Corpus LSFB'.

Laurence Meurant, Aurelie Sinte, and Eric Bernagou. 2016. The French Belgian Sign Language corpus a user-friendly searchable online corpus. In *7th workshop on the Representation and Processing of Sign Languages: Corpus Mining (LREC 2016)*, pages 167–174.

Dipendra K. Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. 2016. Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. *The International Journal of Robotics Research*, 35(1-3):281–300.

Irene Murtagh. 2011. Developing a linguistically motivated avatar for Irish Sign Language visualisation. In *2nd International Workshop on Sign Language Translation and Avatar Technology*.

Irene Murtagh, Víctor Ubieto Nogales, and Josep Blat. 2022. Sign language machine translation and the sign language lexicon: A linguistically informed approach. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 240–251.

Jens Nevens, Robin De Haes, Rachel Ringe, Mihai Pomarlan, Robert Porzel, Katrien Beuls, and Paul Van Eecke. 2024. A benchmark for recipe understanding in artificial agents. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 22–42.

Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2019. Computational construction grammar for visual question answering. *Linguistics Vanguard*, 5(1):20180070.

Magdalena Pasikowska-Schnass. 2018. Sign languages in the EU. Technical Report PE 625.196, European Parliamentary Research Service (EPRS), Brussels, Belgium.

Ana-Maria Popescu, Oren Etzioni, and Henry Kautz. 2003. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 149–157.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, Jan Henning, et al. 1987. HamNoSys. Hamburg Notation System for sign language. An introduction. *Zentrum für Deutsche Gebärdensprache*.

Morgane Rébulard, Claire Danet, Adrien Contesse, Claudia S. Bianchini, Jean-François Dauphin, Léa Chèvrefils, Patrick Doan, and Dominique Boutet. 2018. TYPANNOT, a new glyphic system to annotate handshapes in any sign languages. In *8th Conference of the International Society of Gesture Studies "Gesture and diversity" (ISGS8)*, pages 139–140, Cape Town, South Africa. Heather Brooks.

Éva Sáfár and John Glauert. 2010. Sign language HPSG. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 204–207. European Language Resources Association (ELRA).

Robert Smith. 2013. HamNoSys 4.0 user guide. Technical report, Institute of Technology Blanchardstown, Blanchardstown, Ireland.

Luc Steels. 2011. Introducing Fluid Construction Grammar. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*, pages 3–30. John Benjamins, Amsterdam, Netherlands.

William C. Stokoe. 1960. Sign language structure: An outline of the visual communication system of the American deaf. *Studies in Linguistics: Occasional Papers*.

Raymond Hendy Susanto and Wei Lu. 2017. Semantic parsing with neural hybrid trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3309–3315, Washington, DC, USA. AAAI Press.

Valerie Sutton. 1995. *Lessons in SignWriting: textbook & workbook.* Deaf Action Committee for Sign Writing, La Jolla, CA.

Lappoon R. Tang. 2003. *Integrating Top-down and Bottom-up Approaches in Inductive Logic Programming: Applications in Natural Language Processing and Relational Data Mining*. Ph.D. thesis, Department of Computer Sciences, University of Texas, Austin, TX.

Lappoon R. Tang and Raymond J. Mooney. 2001. Using multiple clause constructors in inductive logic programming for semantic parsing. In *European Conference on Machine Learning*, pages 466–477. Springer.

Dan Tasse and Noah A. Smith. 2008. SOUR CREAM: Toward semantic processing of recipes. Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA, USA.

Sarah F. Taub. 2012. Iconicity and metaphor. In Roland Pfau, Marcus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, volume 37 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 574–601. De Gruyter Mouton, Berlin, Germany.

Paul Van Eecke and Katrien Beuls. 2025. PyFCG: Fluid Construction Grammar in Python. *arXiv preprint arXiv:2505.12920*.

Remi van Trijp. 2015. Towards bidirectional processing models of sign language: A constructional approach in fluid construction grammar. In *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science*, pages 668–673. University of Torino.

Remi van Trijp, Katrien Beuls, and Paul Van Eecke. 2024. The proof is in the almond cookies. In Luc Steels and Robert Porzel, editors, *Narrative-based Understanding of Everyday Activities: A Cookbook*, pages 59–77. Venice International University, Venice, Italy.

Lara Verheyen, Jérôme Botoko Ekila, Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2023. Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, pages 2419–2426.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Yuk Wah Wong and Raymond J. Mooney. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of the 2006 main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics.

William A. Woods. 1967. *Semantics for a question-answering system*. Ph.D. thesis, Harvard University, Cambridge, MA, USA.

William A. Woods. 1981. Procedural semantics as a theory of meaning. In Aravind K. Joshi, Bruce L. Webber, and Ivan A. Sag, editors, *Elements of Discourse Understanding*, pages 300–334. Cambridge University Press.

William A. Woods. 2010. ACL lifetime achievement award: The right tools: Reflections on computation and language. *Computational Linguistics*, 36(4):601–630.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, pages 1050–1055, Washington, D.C., USA. AAAI Press.

John Marvin Zelle. 1995. *Using inductive logic programming to automate the construction of natural language parsers*. Ph.D. thesis, The University of Texas at Austin.

Luke Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 658–666. AUAI Press.