Towards a Corpus-based Approach to Sign Language Dictionaries

Universität Hamburg

Thomas.Hanke@sign-lang.uni-hamburg.de

Abstract

This paper discusses those aspects of iLex, a sign language transcription tool, that are relevant to lexical work and the production of e-learning materials. iLex is built upon a relational database, and uses this strength to support the user in type-token matching by giving immediate access to all other tokens already related to a certain type. iLex features a number of classification schemes, both built-in and data-driven, to allow for the incremental process of identifying and describing the lexicon of a sign language. Data cannot only be exported to other transcription tools, but also into authoring systems for teaching materials. Finally, we speculate about the applicability of Zipf's Law for sign language corpora extrapolating from the current contents of the iLex database.

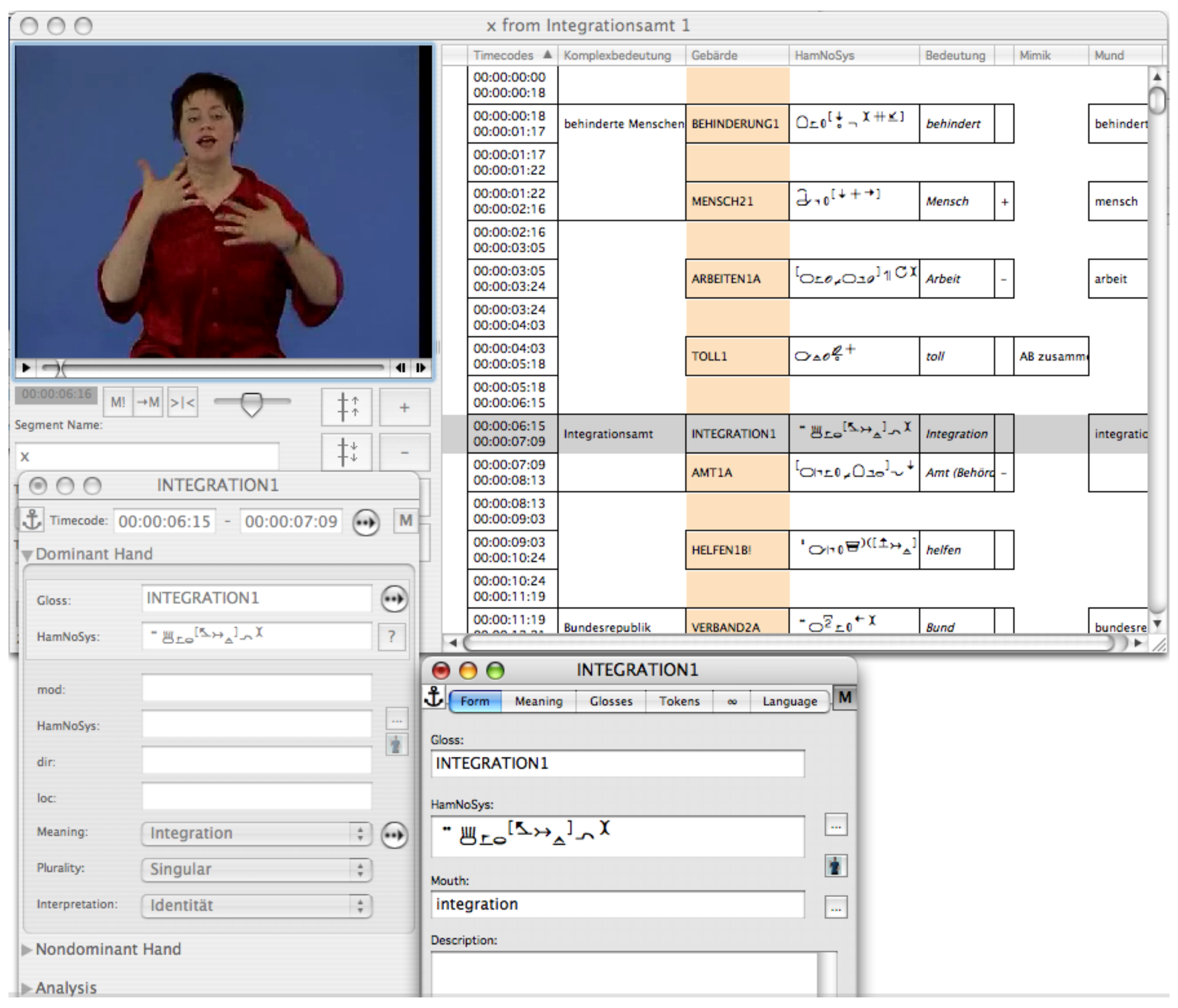
Introduction

Over the past fifteen years, our institute has produced a number of special-terminology dictionaries for German sign language. Having started with introspective data from small focus groups, we quickly moved to an empirical approach where informants were invited to report on their professional experience and to answer to a variety of elicitation settings. The signers' productions were recorded on video and later transcribed. These data provided the almost exclusive source for the dictionaries. A common starting point for all these dictionary projects was to use a list of concepts to be covered in each respective dictionary, usually defined from an educational point of view. The amount of video data collected and thereby the transcription effort needed were mainly determined by the size of this list and the number of informants available.

For these projects, we developed methods and tools to support the transcription and further analysis processes, especially type-token matching – a task much harder than for most spoken languages (Hanke et al., 2001). Moving away from written-language phrases or pictures as elicitation prompts towards semi-structured interviews and discussions, the transcription tools needed to become flexible enough to transcribe any signed discourse, not just short mainly sequential phrases. This now allows us to use the same tool named iLex both for lexicographic work and discourse analysis (Hanke, 2002b, for tools in our earlier work in discourse transcription with syncWRITER cf. Hanke, 2001).

Type-Token Matching

As sign languages have no written form, language resources for sign language often use "phonetic" notations, such as HamNoSys (Prillwitz et al., 1989 and Schmaling/Hanke, 2001, Hanke 2004). However, the current state-of-the-art for sign language notation is far away from being a full compensation for an orthography (Miller, 2001), which in general is the main access key to language data for written language as well as annotated speech. We therefore consider it essential for sign language corpus annotation to explicitly link tokens to lexical entities. The distinctive feature of our transcription tool is that it is built on top of a relational database modelling tokens and types as different entities related to each other. I.e. stretches of signed discourse cannot only be tagged with text, e.g. glosses, but also as tokens related to one specific type.



x from Integrationsamt 1

Timecodes	Komplexbedeutung	Gebärde	HamNoSys	Bedeutung		Mimik	Mund
00:00:00:00 00:00:00:18							
00:00:00:18 00:00:01:17	behinderte Menschen	BEHINDERUNG1		behindert			behindert
00:00:01:17 00:00:01:22							
00:00:01:22 00:00:02:16		MENSCH21		Mensch	+		mensch
00:00:02:16 00:00:03:05							
00:00:03:05 00:00:03:24		ARBEITEN1A		Arbeit	-		arbeit
00:00:03:24 00:00:04:03							
00:00:04:03 00:00:05:18		TOLL1		toll		AB zusamm	
00:00:05:18 00:00:06:15							
00:00:06:15 00:00:07:09	Integrationsamt	INTEGRATION1		Integration			integratio
00:00:07:09 00:00:08:13		AMT1A		Amt (Behörd	-		
00:00:08:13 00:00:09:03							
00:00:09:03 00:00:10:24		HELFEN1B!		helfen			
00:00:10:24 00:00:11:19							
00:00:11:19	Bundesrepublik	VERBAND2A		Bund			bundesre

00:00:06:16

Segment Name: x

INTEGRATION1

Timecode: 00:00:06:15 - 00:00:07:09

Dominant Hand

Gloss: INTEGRATION1

HamNoSys:

mod:

HamNoSys:

dir:

loc:

Meaning: Integration

Plurality: Singular

Interpretation: Identität

Nondominant Hand

Analysis

INTEGRATION1

Gloss: INTEGRATION1

HamNoSys:

Mouth: integration

Description:

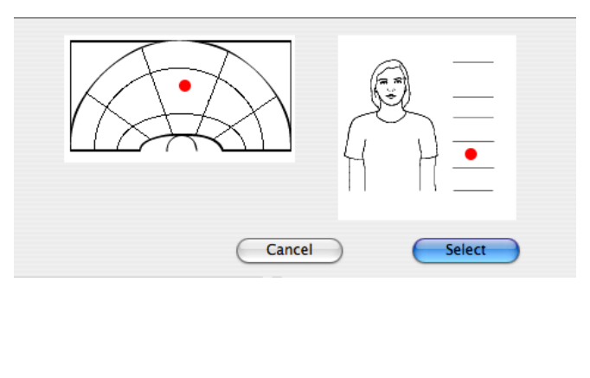
The major advantage of this approach is that in the course of type-token matching, one can always review the video clips showing other tokens related to a candidate type. In addition, the relational model allows a multitude of search approaches to identify candidate types, e.g. by meaning, by gloss, by form, or by grammatical class.

(We share the view of many researchers that glosses are convenient labels for types. It is of course always necessary to keep in mind the danger of using spoken language words for sign language types (cf. Pizzuto/Pietandrea, 2001), and even native signer team members report about various occasions where spoken language labels mislead them. The database approach however implies that token data are constantly reviewed from a number of perspectives, and in many cases glosses play no role so that it is our hope that such cases will be identified even in projects where, for budget reasons, not all transcription work can be independently reviewed.)

Once the type for a token is identified, deviation in form from the type needs to be considered. In the case of grammatical modifications, such as inflection, the system suggests possible categorisations of the modification based on the assumed grammatical class of the type.

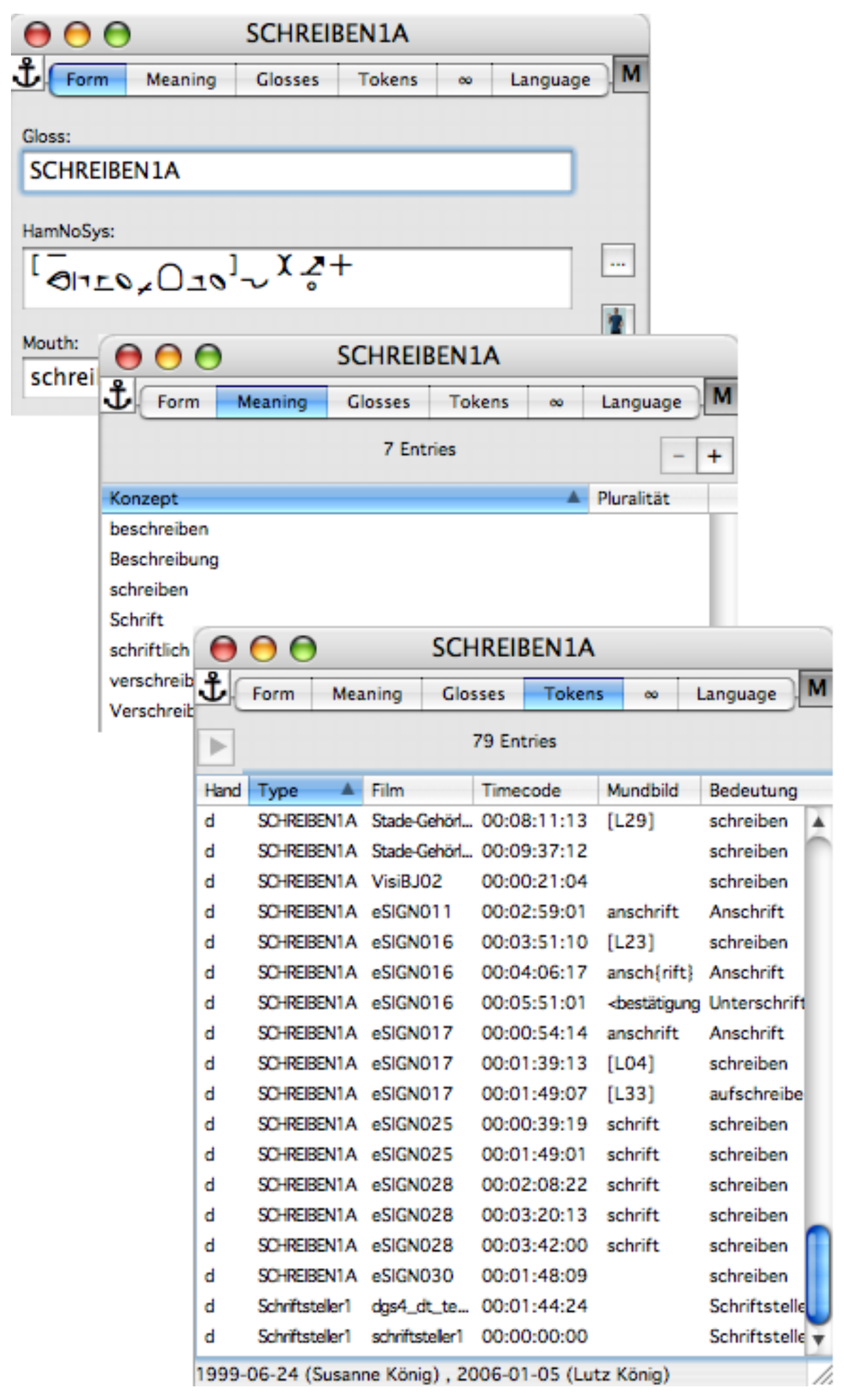
Should a revision of the grammatical classification of a type render modification classifications of some related tokens invalid, the tool suggests these tokens to be reviewed.

As the tool is used by different projects within our institute and as conceptualisations also change over time within one research group, grammatical classes and modifications they allow are modelled by the database as well, so that the grammatical model applied is not determined by the system, but the data. For special graphical editors needed to make data input more efficient, a plug-in model has been implemented.



Type Hierarchy

iLex allows the transcriber to arrange types in a tree hierarchy. We currently use a four-level schema: On the top level, we describe the abstract images or ideas underlying many signs. The manual realisations of these images, i.e. certain forms, are found on the level below. This level is what most researchers would consider the sign inventory. Proper homonyms (non-polysemic) exist on this level: They share their surface form, but are derived from different images.



SCHREIBEN1A

Gloss: SCHREIBEN1A

HamNoSys:

Mouth: schrei

SCHREIBEN1A

7 Entries

Konzept	Pluralität
beschreiben	
Beschreibung	
schreiben	
Schrift	
schriftlich	
verschreib	
Verschreib	

SCHREIBEN1A

79 Entries

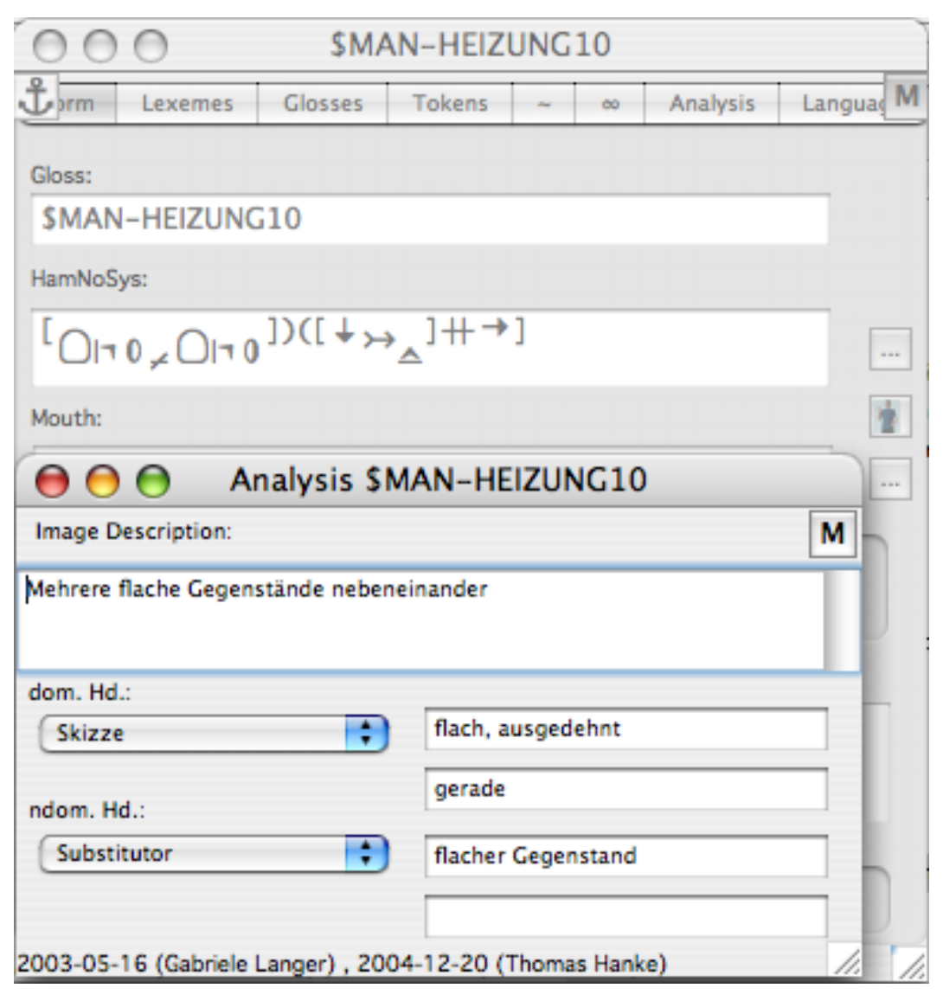
Hand	Type	Film	Timecode	Mundbild	Bedeutung
d	SCHREIBEN1A	Stade-Gehörl...	00:08:11:13	[L29]	schreiben
d	SCHREIBEN1A	Stade-Gehörl...	00:09:37:12		schreiben
d	SCHREIBEN1A	VisiBJ02	00:00:21:04		schreiben
d	SCHREIBEN1A	eSIGN011	00:02:59:01	anschrift	Anschrift
d	SCHREIBEN1A	eSIGN016	00:03:51:10	[L23]	schreiben
d	SCHREIBEN1A	eSIGN016	00:04:06:17	ansch{rift}	Anschrift
d	SCHREIBEN1A	eSIGN016	00:05:51:01	<bestätigung	Unterschrift
d	SCHREIBEN1A	eSIGN017	00:00:54:14	anschrift	Anschrift
d	SCHREIBEN1A	eSIGN017	00:01:39:13	[L04]	schreiben
d	SCHREIBEN1A	eSIGN017	00:01:49:07	[L33]	aufschreibe
d	SCHREIBEN1A	eSIGN025	00:00:39:19	schrift	schreiben
d	SCHREIBEN1A	eSIGN025	00:01:49:01	schrift	schreiben
d	SCHREIBEN1A	eSIGN028	00:02:08:22	schrift	schreiben
d	SCHREIBEN1A	eSIGN028	00:03:20:13	schrift	schreiben
d	SCHREIBEN1A	eSIGN028	00:03:42:00	schrift	schreiben
d	SCHREIBEN1A	eSIGN030	00:01:48:09		schreiben
d	Schriftsteller1	dgs4_dt_te...	00:01:44:24		Schriftstelle
d	Schriftsteller1	schriftsteller1	00:00:00:00		Schriftstelle

1999-06-24 (Susanne König) , 2006-01-05 (Lutz König)

Forms can be assigned certain meanings, and there are numerous examples of signs with many different meanings. Conventionalised form-from-image/meaning pairs are notated on the third level, allowing e.g. their own glosses without obscuring their relation to other meanings that share form and image. In our model, this level corresponds to the lexicalisations level in a dictionary. For DGS, we can often notate a default mouth picture on this level.

On the lowest level, both forms and form/meaning pairs can be split up according to project-specific needs. E.g. projects using the transcription tools in order to produce e-learning materials currently use this level to assign alternative glosses they consider more appropriate in a certain didactic context. It is also possible to consider some modifications of types as separate entities and to use this level to introduce the dependencies, e.g. to introduce separate I and YOU if on the higher level only one (person) reference entity exists.

First experiments indicate that this hierarchical approach also has the potential to model the overlap of the type inventories of different sign languages. For this purpose, types can be attributed with languages that they appear attested for. Filters then allow the user to concentrate on the type inventory and associated tokens for one specific language only or to view data from a multitude of languages at the same time.



$MAN-HEIZUNG10

Gloss: $MAN-HEIZUNG10

HamNoSys:

Mouth:

Analysis $MAN-HEIZUNG10

Image Description: Mehrere flache Gegenstände nebeneinander

dom. Hd.: Skizze — flach, ausgedehnt

gerade

ndom. Hd.: Substitutor — flacher Gegenstand

2003-05-16 (Gabriele Langer) , 2004-12-20 (Thomas Hanke)

iLex allows the user to define irreflexive relations between types of a certain level. We currently use this to further analyse homonymic as well as close-neighbour form relations. In addition, types can be analysed with respect to the image production techniques used (Konrad et al. 2004).

Representation of Form

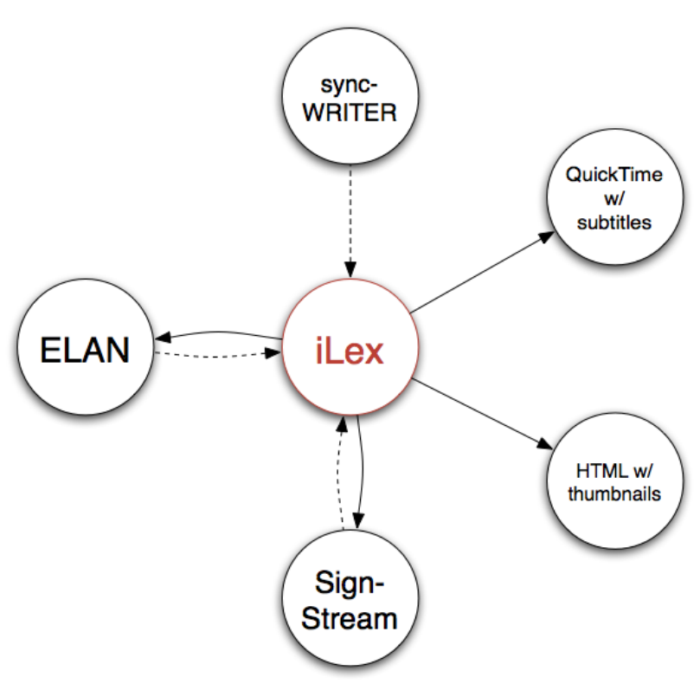
Obviously, since the advent of digital video and especially in a database context such as iLex, where videoclips associated with tokens are immediately available, phonetic transcriptions no longer have an exclusive role to describe the form. But as long as is not possible to automatically search a video for certain features of a sign, phonetic data allows access to the data not otherwise possible. For tokens, HamNoSys as we use it in Hamburg certainly is an adequate form description. HamNoSys-compatible avatar software (Elliott et al. 2004) allows the transcriber to immediately verify HamNoSys notations written in iLex. Types, on the other hand, may require a different system.

While we use HamNoSys here as well with recent additions to allow underspecification or ranges of permissible handshapes, for example (Hanke 2002a), an easier way to abstract away from individual variations is highly desirable. Phonological models, however, still await the availability of large lexical databases in order to be verified.

Data Exchange

For a larger lab, where not only researchers work on transcriptions, but also students, a central database has the major advantage that transcription cannot go lost as people leave or students finish their exams. This cannot, however, mean that sharing data with the research community should become more cumbersome. ilex therefore provides export modules to transfer selected transcriptions to formats used by other researchers, such as ELAN (Crasborn et al. 2004) or SignStream (Neidle 2001). For read-only purposes, transcriptions can also be exported as QuickTime movies with subtitles or as scores in HTML format to be viewed with any browser. In cases where the original video cannot be made available to the public, data can be exported to eSIGN documents that can then be played back by an avatar (Hanke 2004).

Importing data made available by other researchers as ELAN or SignStream documents requires a two-step approach. In a first step, data are imported into transcripts with only text tiers. In a second (optional) step, glosses as text should be replaced by database references. This step can only partially be automated, but finally results in transcripts that make full use of the iLex database structure.

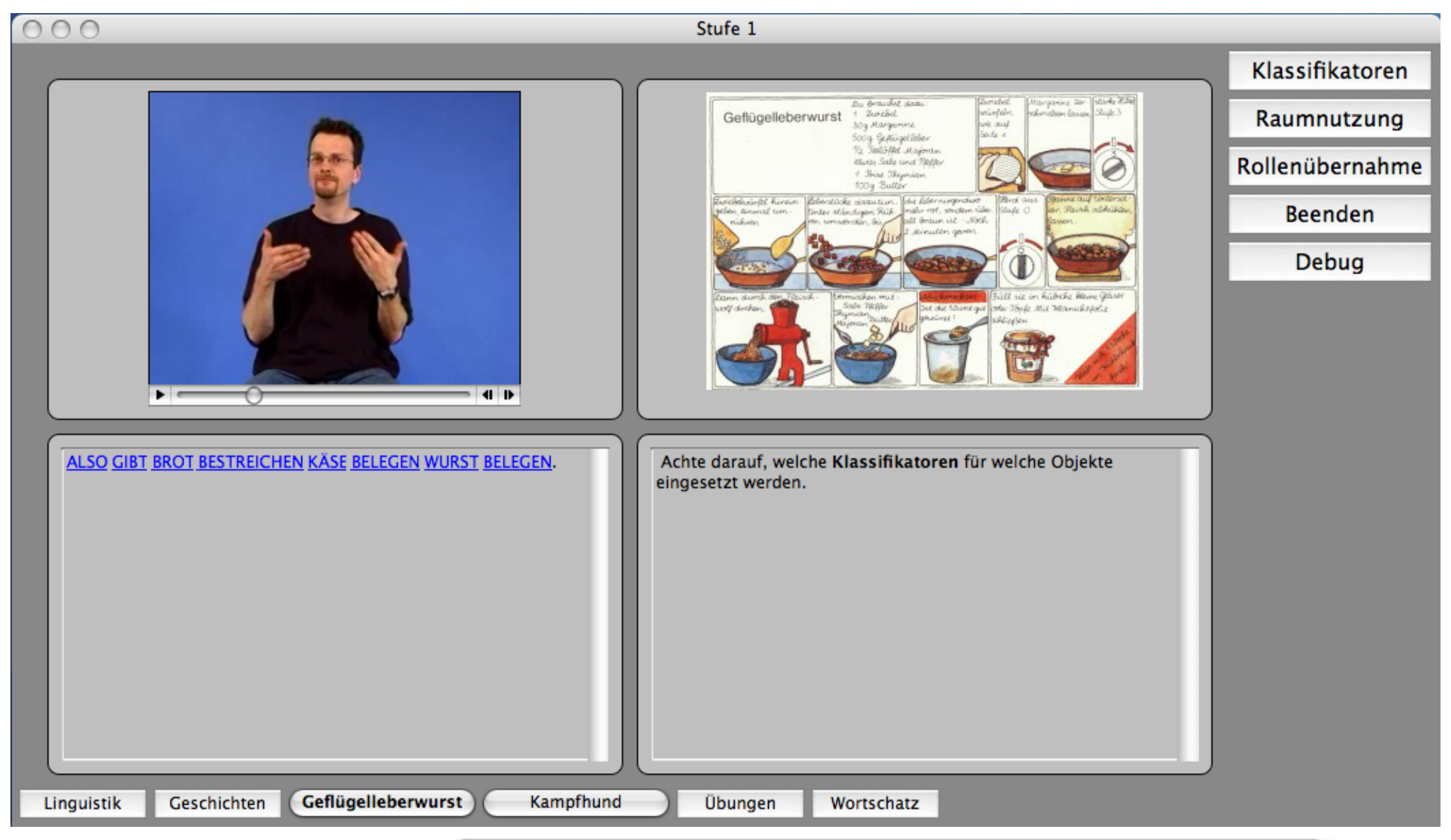


Whereas ELAN, SignStream, and iLex share the idea that tags label intervals of time and therefore can be thought of as variations of the concepts formalised by Bird and Liberman (2001), import from syncWRITER documents requires a number of assumptions as syncWRITER primarily tagged points in time. It can therefore become necessary to "repair" syncWRITER documents before or after the import process.

iLex supports the user in building metadata on all aspects of a signed discourse. For this, it supports all features required by Crasborn and Hanke (2003).

Applications in Teaching Materials

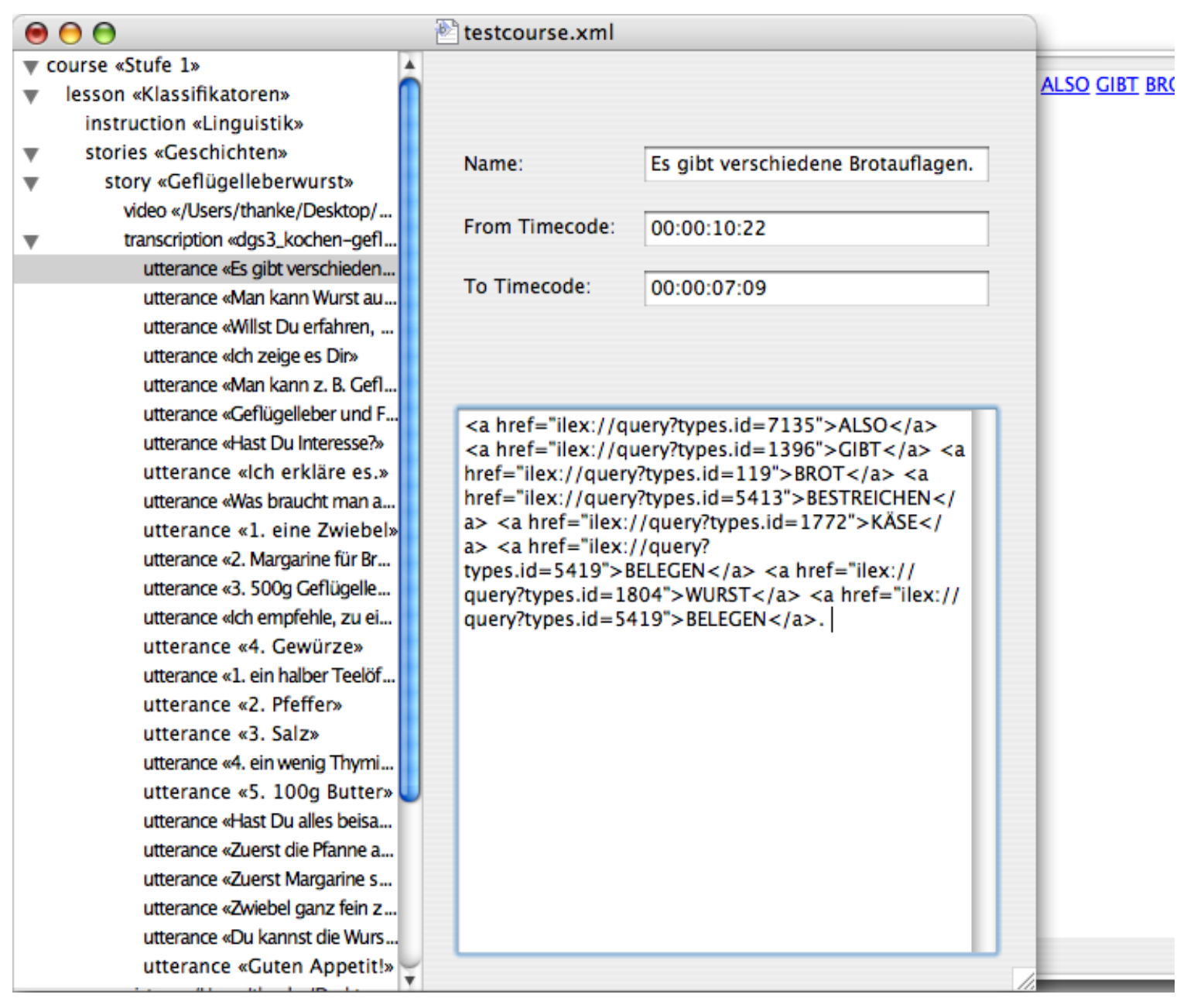
While we have produced high-quality sign language teaching CD-ROMs in the past (Metzger 2005), that have been individually programmed, we also see the need for less sophisticated, but easy and quickly to produce materials for our everyday teaching. Ideally, the lecturers should be able to do the complete production process themselves. Often the most complicated assets in e-learning materials for sign language is videos with time-aligned explanations and links, e.g. into a lexicon. The idea is to produce these assets as



Stufe 1

ALSO GIBT BROT BESTREICHEN KÄSE BELEGEN WURST BELEGEN.

Achte darauf, welche **Klassifikatoren** für welche Objekte eingesetzt werden.



testcourse.xml

```
▼ course «Stufe 1»
▼   lesson «Klassifikatoren»
      instruction «Linguistik»
▼     stories «Geschichten»
▼       story «Geflügelleberwurst»
          video «/Users/thanke/Desktop/...
▼         transcription «dgs3_kochen-gefl...
            utterance «Es gibt verschieden...
            utterance «Man kann Wurst au...
            utterance «Willst Du erfahren, ...
            utterance «Ich zeige es Dir»
            utterance «Man kann z. B. Gefl...
            utterance «Geflügelleber und F...
            utterance «Hast Du Interesse?»
            utterance «Ich erkläre es.»
            utterance «Was braucht man a...
            utterance «1. eine Zwiebel»
            utterance «2. Margarine für Br...
            utterance «3. 500g Geflügelle...
            utterance «Ich empfehle, zu ei...
            utterance «4. Gewürze»
            utterance «1. ein halber Teelöf...
            utterance «2. Pfeffer»
            utterance «3. Salz»
            utterance «4. ein wenig Thymi...
            utterance «5. 100g Butter»
            utterance «Hast Du alles beisa...
            utterance «Zuerst die Pfanne a...
            utterance «Zuerst Margarine s...
            utterance «Zwiebel ganz fein z...
            utterance «Du kannst die Wurs...
            utterance «Guten Appetit!»
```

Name: Es gibt verschiedene Brotauflagen.

From Timecode: 00:00:10:22

To Timecode: 00:00:07:09

```
<a href="ilex://query?types.id=7135">ALSO</a> <a href="ilex://query?types.id=1396">GIBT</a> <a href="ilex://query?types.id=119">BROT</a> <a href="ilex://query?types.id=5413">BESTREICHEN</a> <a href="ilex://query?types.id=1772">KÄSE</a> <a href="ilex://query?types.id=5419">BELEGEN</a> <a href="ilex://query?types.id=1804">WURST</a> <a href="ilex://query?types.id=5419">BELEGEN</a>.
```

ALSO GIBT BRO

transcriptions in iLex, and then to import them into the authoring environment as complex content objects. We have therefore developed an authoring tool closely integrated with iLex. Through the interaction, links into a dictionary and the dictionary itself can be produced almost without any manual intervention. The player module, of course, works standalone and does not require a connection to the iLex database.

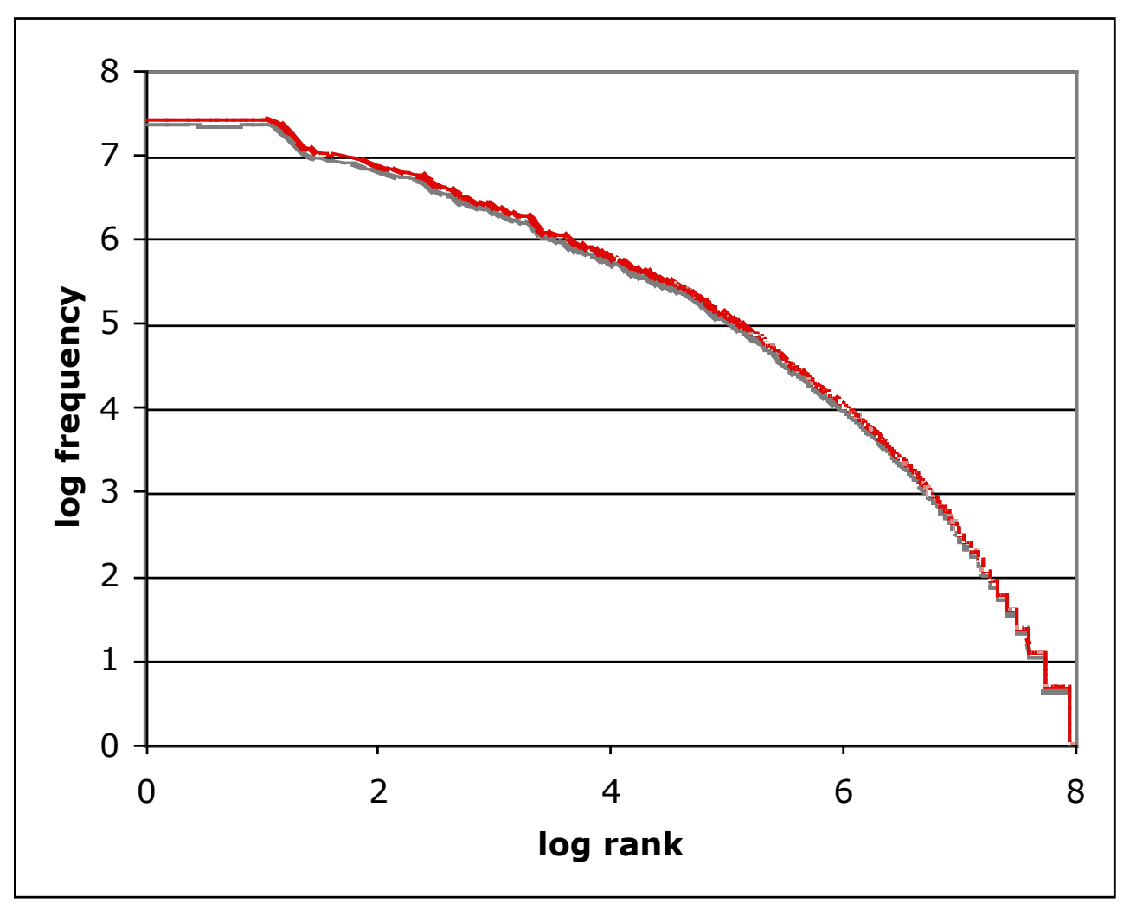
Zipf's Law for Sign Languages

When planning a general dictionary of DGS, there is no word list to start with. For a basic vocabulary, methods developed for spoken languages have been successfully adapted to result in a seed for a basic vocabulary of a signed language (Efthimiou/Katsoyannou 2001). For larger dictionaries, however, we see no alternative to a completely corpus-driven approach. The question then of course is how large a corpus needs to be in order to cover a sufficiently large portion of the lexicon.

For spoken languages, these predictions are often based on the rules of thumb referred to as Zipf's Law. The basic idea is that the product of the frequency of a word in a corpus and its rank is more or less constant over all words in the corpus.

Can we expect such a rule to also apply to sign languages? Function "words" play a significantly smaller role than e.g. in English, and it is not clear how productive signs fit into the game.

Certainly, we do not have a balanced corpus of reasonable size available to "verify" Zipf's law. Nevertheless, we did some math experiment with the current contents of the iLex database, only counting those tokens that refer to types undoubtedly qualifying as "lexical". These accounted for 108000 out of 125000 tokens. Surprisingly, the graph does look relatively smooth. While the graph is not exactly what you would expect for English, the low slope in the first ranks comes close to what Ha and Smith (2004) reported for Irish, a highly-inflected Indo-European language.



So we are tempted to "trust" Zipf predictions and use future work on the production of a general dictionary of German Sign Language to verify this.

References

Bird, S and M. Liberman (2001). A formal framework for linguistic annotation. Speech Communication 33(1,2): 131-162.

Crasborn, O. and T. Hanke (2003). Metadata for sign language corpora. Available online at: http://www.let.ru.nl/sign-lang/echo/docs/ ECHO_Metadata_SL.pdf.

Crasborn, O., J. Mesch and E. v.d. Kooij (2004). European cultural heritage online (ECHO): publishing sign language data on the internet. Poster presented at TISLR 8 Barcelona, Spain.

Efthimiou, E. and M. Katsoyannou (2001). Research issues on GSL: a study of vocabulary and lexicon creation. Studies in Greek Linguistics 2: 42-50 (in Greek).

Elliott, R. et al. (2004). An overview of the SiGML notation and SiGMLsigning software system. In: O. Streiter and C. Vettori (eds.): Proceedings of the Workshop on Representing and Processing of Sign Languages, LREC 2004, Lisbon, Portugal, pp. 98–104.

Ha, L. Q., and F. J. Smith (2004). Zipf and type-token rules for the English and Irish languages. In: Proceedings of the Modelling for the Identification of Languages Workshop, Paris, France, pp. 65-70.

Hanke, T. (2001). Sign language transcription with syncWRITER. Sign Language and Linguistics. 4(1/2): 275-283.

Hanke, T. (2002a). HamNoSys in a sign language generation context. In: R. Schulmeister and H. Reinitzer (eds): Progress in sign language research. In honor of Siegmund Prillwitz / Fortschritte in der Gebär¬den¬sprach¬forschung. Festschrift für Siegmund Prillwitz (pp. 249-264). Hamburg: Signum.

Hanke, T., (2002b). iLex - A tool for sign language lexicography and corpus analysis. In: M. González Rodriguez, Manuel and C. Paz Suarez Araujo (eds.): Proceedings of the third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, Spain. (pp. 923-926). Paris: ELRA.

Hanke, T. (2004). HamNoSys - Representing sign language data in language resources and language processing contexts. In: O. Streiter and C. Vettori (eds.): Proceedings of the Workshop on Representing and Processing of Sign Languages, LREC 2004, Lisbon, Portugal, pp. 1-6.

Hanke, T., R. Konrad and A. Schwarz (2001). GlossLexer – A multimedia lexical database for sign language dictionary compilation. Sign Language and Linguistics 4(1/2): 161–179.

Konrad, R. et al. (2004). What's in a sign? Theoretical lessons from practical sign language lexicography. Poster presented at TISLR 8 Barcelona, Spain.

Metzger, C. (2005). Die Firma. In: H. Leuninger and D. Happ (eds): Gebärdensprachen: Struktur, Erwerb, Verwendung, pp. 309-324. Hamburg: Buske.

Miller, C. (2001). Some reflections on the need for a common sign notation. Sign Language and Linguistics 4(1/2): 11-28.

Neidle, C. (2001). SignStream™: A database tool for research on visual-gestural language. Sign Language and Linguistics. 4(1/2): 203-214.

Pizzuto, E., and P. Pietrandrea (2001). The notation of signed texts: Open questions and indications for further research. Sign Language and Linguistics 4 (1/2): 29-45.

Prillwitz, S. et al. (1989). HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide. Hamburg: Signum.

Schmaling, C. and T. Hanke (2001). HamNoSys 4.0. In: T. Hanke (ed.), Interface definitions. ViSiCAST Deliverable D5-1. [This chapter is available at http:// www.sign-lang.uni-hamburg.de/projekte/ HamNoSys/HNS4.0/englisch/HNS4.pdf]

http://www.sign-lang.uni-hamburg.de/iLex/