

발 표 자 료 집

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전

Records of Visible Language:
Sign Language Dictionary

2018. 11. 8. (목)

이화여자대학교 이화삼성교육문화관
1층 강당

주최 · 주관



국립국어원



국립국어원장
소 강 춘

안녕하십니까? 국립국어원장 소강춘입니다.

귀한 시간을 내어 참석해 주신 모든 분들께 감사의 말씀을 올립니다. 풍요로운 계절에 이 자리에 참석하신 분들을 뵙게 되어 반갑습니다. 국립국어원은 2010년부터 2년에 한 번씩 세계의 언어 정책을 국내에 소개하고 관련 기관들과 교류와 협력 관계를 다지는 국제 학술 대회를 마련해 왔습니다. 지난 2016년의 ‘수어언어와 사회적 의사소통’ 주제에 이어, 올해에는 ‘보이는 언어의 기록, 수어사전’이라는 주제로 국제 학술 대회를 개최하게 되었습니다.

이러한 뜻깊은 자리에 소중한 경험을 나누어 주시기 위해 멀리서 오신 예테 크리스토퍼센 선생님, 요한나 메시 선생님, 트레버 존스틴 선생님, 토마스 항케 선생님께 감사와 환영의 말씀을 올립니다. 애초 기초연설을 해 주시길 하였던 글래디스 탕 교수님은 안타깝게도 갑작스러운 개인 사정으로 참석하지 못하셨습니다.

2016년 ‘한국수어언어법’의 제정과 시행으로 농인의 고유한 언어인 한국수어는 우리 사회에서 언어로서의 자격을 법적으로 보장받게 되었으며, 27만 농인과 한국수어 사용자의 언어권과 삶의 질을 실질적으로 보장하기 위한 발판이 마련되었습니다. 국립국어원은 기존에 구축된 수어사전을 정비하여 더욱 체계적인 수어사전 구축을 계획하고 있는데 이번 학술 대회를 계기로 한 걸음 더 나아간 한국어 수어사전의 모습을 전망하고 농사회의 발전과 공감을 이끌어내려고 합니다. 이번 학술 대회에서 많은 논의가 이루어짐으로써 우리 사회가 소통의 장벽 없는 사회로 한발 더 다가설 수 있는 계기가 마련될 것이라고 생각합니다. 오늘 이 자리에서 말뭉치 중심의 수어사전, 수어사전의 구조와 표제어 표기, 정보 기술을 활용한 수어사전의 미래 등에 대하여 논의하면서 앞으로의 한국수어사전의 발전 방향을 모색하는 귀중한 자리가 되기를 바랍니다.

오늘 이 자리를 빛내 주신 귀빈 여러분, 바쁜 일정에도 우리 학술 대회를 찾아주신 여러 청중들, 그리고 발표를 맡아 주신 세계 각국의 수어언어 전문가 여러분께 다시 한번 깊은 감사의 말씀을 올립니다. 모쪼록 이번 국제 학술 대회를 통해 수어사전의 발전을 이끌어내는 풍성한 수확을 거두기를 기원합니다.

2018. 11. 8.

Opening Remarks



Director General of NIKL
So Kang-Chun

Greetings. I am So Kang-Chun, Director General of the National Institute of Korean Language (NIKL).

I would like to start by expressing my gratitude to all participants for taking the time to attend this conference. It is a pleasure to see everyone here in this season of harvest. The National Institute of Korean Language has hosted the international academic conference biannually since 2010 to introduce language policies from around the world and to strengthen ties for cooperation and exchange with other organizations. Following the theme of “Sign Language and Social Communication” in 2016, this year’s conference is being held under the theme of “Records of Visible Language: Sign Language Dictionary”.

I want to extend my heartfelt gratitude and welcome to our presenters Dr. Jette Kristoffersen, Dr. Johanna Mesch, Dr. Trevor Johnston, and Dr. Thomas Hanke, who travelled far to share their valuable insights with us. Dr. Gladys Tang was also scheduled

to join us to deliver the keynote address, but unfortunately, she could not be here with us due to personal circumstances.

With the enactment of the Korean Sign Language Act in 2016, Korean sign language gained legal recognition as a language and laid the foundation for 270,000 deafs and sign language users to ensure their right to language and quality of life. The National Institute of Korean Language is currently working on a more systematic revision of the Korean Sign Language Dictionary. We believe this conference will bring us one step closer to the new and improved dictionary and help us develop and gain a better understanding of the deaf society. Through active discussions, we can lay the groundwork toward a society free of communication barriers. As we touch on topics such as corpus-based dictionaries, dictionary structure and entries, and the future of sign language dictionaries with IT, I hope there will be meaningful discussions about the future direction of the Korean Sign Language Dictionary.

I would like to once again express my deepest appreciation to our distinguished guests and members of the audience for attending the conference, as well as sign language experts from around the world who will be presenting today. I hope we can reap a fruitful and rich harvest from this conference for the future growth of sign language dictionaries.

November 8, 2018



축사



문화체육관광부
문화예술정책실장
이우성

안녕하십니까? 문화체육관광부 문화예술정책실장 이우성입니다.

‘보이는 언어의 기록, 수어사전’ 국립국어원 2018 국제 학술 대회 개최를 축하드리며, 이 행사에 참석하여 자리를 빛내 주신 모든 분들께 감사와 환영의 마음을 전합니다. 특히, 이번 학술 대회를 위해 홍콩, 덴마크, 스웨덴, 호주, 독일 등에서 찾아와 주신 발표자 여러분께 특별한 감사의 인사를 전합니다.

올해는 한글을 창제하신 세종대왕이 즉위하신 지 600돌이 되는 해입니다. 누구나 자유롭게 소통할 수 있는 세상. 이것이 세종대왕이 꿈꾸던 세상이고, 한글은 바로 그 꿈의 산물이 아닌가 합니다. 600년이 흐른 지금 한글은 한국어 사용자들을 잇는 다리로서 그 소임을 다하고 있습니다. 2016년에 제정된 「한국수어언어법」에는 한글 창제에 담긴 정신을 한국수어 사용자와도 함께 나누기 위한 뜻이 담겨 있습니다. 「한국수어언어법」은 한국수어가 한국어와 동등한 자격을 가진 농인의 고유한 언어임을 천명하고, 한국수어 사용자가 자유롭게 소통할 수 있는 환경을 만들어 나가는 것을 목적으로 하고 있기 때문입니다. 이번 학술 대회는 이러한 목표를 이루기 위한 의미 있는 발걸음이라고 생각합니다.

지난해 문체부는 ‘제1차 한국수어발전기본계획(2018~2022)’을 발표한 바 있습니다. 이 계획에 따라 한국수어 교육자료 개발, 한국수어 교육기관 지원, 한국수어교원 자격제도 운영, 한국수어 문화정보 구축, 한국수어사전 편찬 등 한국수어 발전을 위해 다양한 정책을 추진하고 있습니다. 특히, 새롭게 편찬하고 있는 ‘한국수어사전’은 표제어, 뜻풀이, 용례 모두 한국수어가 중심이 되는 사전을 목표로 하고 있습니다. 농인과 농인, 농인과 청인을 이어주는 ‘한국수어사전 편찬’은 매우 중요한 작업입니다. 이번 학술 대회가 우리 수어사전의 편찬 수준을 한층 높일 수 있는 계기가 되길 바랍니다.

오늘 학술 대회를 위해 멀리서 오신 발표자 선생님을 비롯한 청중 여러분들께 다시 한번 감사드리며, 행사 준비를 위해 애써 주신 소강춘 국립국어원장님을 비롯한 관계자 여러분께도 감사의 마음을 전합니다. 유난했던 여름에 여봐란듯이 쌀쌀한 날씨가 이어지고 있습니다. 모두 건강에 유의하시면서 즐겁게 한 해를 마무리하시기 바랍니다. 고맙습니다.

2018. 11. 8.

Congratulatory Remarks



Head of Culture
and Arts Policy Office
Ministry of Culture,
Sports and Tourism
Lee Woo-sung

Greetings. I am Lee Woo-sung, Head of the Culture and Arts Policy Office at Korea's Ministry of Culture, Sports and Tourism.

First of all, I would like to congratulate the opening of the National Institute of Korean Language's 2018 International Academic Conference, held under the theme of "Records of Visible Language: Sign Language Dictionary". I would like to extend my warmest appreciation and welcome to everyone who came out to the conference today. In particular, I especially want to thank the presenters who travelled all the way from Hong Kong, Denmark, Sweden, Australia, and Germany.

This year marks the 600th anniversary of the enthronement of King Sejong, the inventor of the Korean alphabet Hangul. King Sejong dreamt of a world where all people can communicate freely, and Hangul is the product of that dream. Now, 600 years later, Hangul is fulfilling its role as a bridge between Korean language users. The Korean Sign Language Act, enacted in 2016, captures the determination to share the spirit of Hangul with sign language users. This Act declares Korean sign language as the native language of the deaf people with an equal status as Hangul, and aims to create an environment where Korean sign language users can communicate freely. I believe this conference is a meaningful step toward achieving this goal.

Last year, Korea's Ministry of Culture, Sports and Tourism announced its first Basic Plan on Korean Sign Language Development (2018-2022). Following this plan, many policies have been carried out for the advancement of Korean sign language, including development of educational resources, support for educational institutions, operation of an instructor qualification system, building culture information, and compilation of a dictionary. In particular, the Korean Sign Language Dictionary will consist of entries, definitions, and examples that are specific for Korean sign language. Compilation of a Korean Sign Language Dictionary is an important task, bringing together deaf people with other deaf people, as well as non-deaf people. I hope this conference will be a valuable opportunity to enhance the quality of the new dictionary.

Once again, I would like to thank the presenters and the members of the audience for attending today's conference. I also want to extend my gratitude to Director General So Kang-Chun of the National Institute of Korean Language and his staff for all of their hard work in preparation of this event. With the scorching summer behind us, the chilly weather is finally making an ostentatious appearance. I hope you take care of your health, and I wish you all a good rest of the year. Thank you.

November 8, 2018



Program

시간	내용
09:30~10:00 ('30)	등록 / Registration
10:00~10:30 ('30)	개회식(개회사, 축사) / Opening Ceremony

1부 / Session 1

10:30~11:10 ('40)	주제 1: 수어 말뭉치 중심의 수어사전 • 발표자: 트레버 존스톤(Trevor Johnston, 호주, Macquarie University)
11:10~11:50 ('40)	주제 2: 수어사전의 구조 • 발표자: 예테 크리스토퍼센(Jette H. Kristoffersen, 덴마크, University College Capital)
11:50~13:50 ('120)	점심 식사 / Luncheon

2부 / Session 2

13:50~14:30 ('40)	주제 3: 수어사전 편찬의 과제 • 발표자: 요한나 메시(Johanna Mesch, 스웨덴, Stockholm University)
14:30~15:10 ('40)	주제 4: 대한민국 수어사전의 현황과 발전 방향 • 발표자: 최혜원, 이현화(한국, 국립국어원)
15:10~15:30 ('20)	휴식 / Coffee Break

3부 / Session 3

15:30~16:10 ('40)	주제 5: 미래의 수어사전 • 발표자: 토마스 항케(Thomas Hanke, 독일, Hamburg University)
16:10~16:30 ('20)	폐회 / Closing

목차 / Contents

1부 발표 | Session 1

주제 1

수어 말뭉치 중심의 수어사전 11

트레버 존스턴_ 호주, 맥쿼리 대학교

A Corpus-oriented Sign Language Dictionary 28

Trevor Johnston_ Macquarie University, Australia

주제 2

수어사전의 구조 43

예테 크리스토퍼센_ 덴마크, 유시시(UCC) 대학교

Structure of a Sign Language Dictionary 52

Jette H. Kristoffersen_ University College Capital, Denmark

2부 발표 | Session 2

주제 3

수어사전 편찬의 과제 61

요한나 메시_ 스웨덴, 스톡홀름 대학교

Challenges of creating a Sign Dictionary 70

Johanna Mesch_ Stockholm University, Sweden

주제 4

대한민국 수어사전의 현황과 발전 방향 79

최혜원, 이현화_ 한국, 국립국어원

Current Status and Development Direction of Korean Sign Language Dictionary 95

Hyewon Choi, Hyunhwa Lee_ National Institute of Korean Language, Korea

3부 발표 | Session 3

주제 5

미래의 수어사전 113

토마스 한케_ 독일, 함부르크 대학교

Sign Language Dictionaries of Tomorrow 125

Thomas Hanke_ Hamburg University, Germany

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전



Records of Visible Language:
Sign Language Dictionary

주제 1 | 수어 말뭉치 중심의 수어사전

트레버 존스톤
호주, 맥쿼리 대학교

Trevor Johnston
Macquarie University, Australia

1부 발표

수어 말뭉치 중심의 수어사전

트레버 존스톤
호주, 맥쿼리 대학교

개요 [슬라이드 2]

본 강의에서는 말뭉치 중심 수어사전, 특히 ‘오슬란 사인뱅크(Auslan Signbank)’로 알려진 호주 수어(이하 ‘오슬란’) 온라인 사전을 다룬다.

본 강의는 두 부분으로 구성된다. 먼저 파트 1에서는 1985년 이후로 출판되어 온 다양한 오슬란 사전의 발전과 역사를 살펴본다. 특히 1984년에 만들어졌으며 1985년 이후 책, 시디롬, 온라인 형식으로 출판되어 온 모든 오슬란 사전의 기초가 되는 수어 어휘 데이터베이스를 논의한다. 이들 모두 오슬란 말뭉치(Auslan Corpus)가 만들어지기 전에 출판되었다. 해당 데이터베이스는 ‘오슬란 어휘 데이터베이스’라 불린다. 또 다양한 사전 형식의 항목 구조에 대해서도 설명한다.

파트 2에서는 오슬란 사인뱅크를 사례로 들어 말뭉치 중심 수어사전을 설명한다. 오슬란 사인뱅크는 2004~2007년 사이에 오슬란 기록 보관소가 만들어지고(Johnston, 2008) 해당 기록 보관소가 오슬란 말뭉치라는 이름의 언어 말뭉치에 주석이 달리기 시작하면서 말뭉치 중심 수어사전이 되었다. 본 사전에 기록된 수어는 현재 크게 두 가지 측면에서 말뭉치에서 확인된 수어와 연관이 있다. 첫째는 식별용 주석을 사용한 것이고, 둘째는 수어 사용자와 그들이 사용하는 수어 관련 말뭉치로부터 정보를 얻은 사전의 항목으로 수어를 통합한 것이다.

파트 1: 말뭉치 이전 수어사전 [슬라이드 3-39]

오슬란 어휘 데이터베이스 [슬라이드 4-5]

오슬란 어휘 데이터베이스는 엠에스 워드(MS word) 문서 형태로 된 일련의 표 항목으로 1984년에 시작되었다. 1984년부터 2018년까지 34년 동안 해당 데이터베이스는 지속적인 조사에 따른 수정 및 추가를 통

해 끊임없이 개선되어 왔다.

1984년 이후로 반복을 거듭해 온 오슬란 어휘 데이터베이스는 표제어로 입력된 기존 오슬란 어휘의 저장소로 간주되어 왔다. 수어 사용자들이 서로 의사소통할 때 만들어내곤 하는 다른 유형의 수어 항목은 체계적으로 본 데이터베이스에 저장되지 않는다. 비어휘적인 이들 수어는 묘사(또는 ‘식별자’ 기호), 지시 기호, 행위 및 몸짓으로 불러왔다. 굴절형태소에 해당하는 일부 표제어가 변형되거나 바뀐 경우가 있다. 동작주·피동작주 역할이나 상 정보의 동사를 표시하는 변화 또는 명사의 복수형을 표시하는 변화가 그 예이다. 것처럼 굴절된 형태는 사전에 별도로 들어가지 않는다. 대부분의 언어 사전들과 마찬가지로 변형된 형태는 어휘보다는 해당 언어의 문법에서 논의되도록 한다. 하지만 경우에 따라서는 기호의 이전 굴절 변경은 사실상, 그리고 불규칙적으로 새롭고 예측 불가능한 의미, 다시 말해 새로운 어휘와 관련이 있다. 그와 같은 경우는 사전에 추가된다.

이 조사는 다음의 몇 가지 기법을 이용하여 수행되었다. (1) 원어민 수어 사용자의 직관과 자기 성찰, (2) 농인 공동체에서의 참여자 관찰 세션 동안 녹화한 청각장애 원주민 수어 사용자 간의 대화 영상 시청 및 분석, (3) 원주민 수어 사용자를 대상으로 실시한 그룹·개인 면담상에서 개별 기호의 용법과 의미를 논의, 검사 또는 유도

2004년 오슬란 어휘 데이터베이스는 오슬란 사인뱅크 사전의 온라인 출판을 목적으로 대대적인 정비를 거쳤다. 이 사전에는 피드백 메커니즘이 적용되어 있어서 등록 사용자들이 누락된 수어나 부정확한 정의를 지적하거나 보고할 수 있었다. 머지않아 오슬란 말뭉치가 탄생했다. 그 이후로 오슬란 사인뱅크상에 이미 기록되어 있는 수어와 관련하여 식별용 주석의 형태로 말뭉치의 주석을 표시했다.

시간의 경과에 따른 오슬란 어휘 데이터베이스 [슬라이드 6]

오슬란 어휘 데이터베이스상의 데이터는 1984년 이후로 다양한 유형의 플랫폼 사이를 이동했다. 다양한 버전의 오슬란 사전이 기초를 두고 있는 데이터셋인 것이다.

오슬란 어휘 데이터베이스는 1985-1989년부터 엠에스 워드(MS Word)의 초창기 버전에 저장된 기록 약 2,000건의 목록으로 존재했다. 1990년대 초 엠에스 워드(MS Word) 표는 애플의 하이퍼카드(HyperCard) 스택으로 이동했다(1990-1992). 하이퍼카드 스택 덕분에 컴퓨터상의 각 기호 기록 내에 시각 정보를 쉽게 포함시킬 수 있게 되었다. 각 기호 항목/기록은 라인드로잉(라인드로잉이 오슬란 사전에 등장한 것은 1989년이다)을 이용하여 표시하였다.

1990년대 초 데이터는 또 다시 상업적으로 이용 가능한 데이터베이스로 이동했는데 처음은 폭스프로(FoxPro)였고 그 다음이 파일메이커 프로(FileMaker Pro)였다. 이들 새로운 데이터베이스상의 각 기록은 기호별로 음운론, 의미론, 방언, 도상성 등 다양한 정보가 기호화된 100여 개의 필드를 지니고 있었다(본 강의 파트 2에서 이를 다시 다룰 예정이다). 이들 프로그램은 일부 필드에 영상 및 그래픽 정보를 입력할 수 있었으며 이를 통해 스프레드시트 내에서 각각의 기호 표제어를 시각적으로 보여주는 것이 가능했다.

이와 비슷한 신규 오슬란 어휘 데이터베이스가 1990년대 말에 생성되었다. 파일메이커 프로 데이터베이스

1부 발표

스의 데이터를 내보내 시디롬을 지원하는 전용 컴퓨터 프로그램으로 통합시키는 방식이었다. 해당 프로그램은 멀티미디어 시디롬 오슬란 사전을 만들기 위해 사용되었다. 시디롬을 지원하는 프로그래밍 언어는 각 기호와 관련 있는 복수의 필드, 그리고 데이터베이스상의 기호 사이에, 다시 말해 동의성 관계, 방언 변이, 제한적 용법 등 다양한 기록 및 필드 사이에 수립된 수많은 교차 참조를 이용할 수 있었다.

2000년대 초 비디오 필드를 포함한 파일메이커 프로의 전체 데이터셋은 시디롬과 기본적으로 기능이 동일한 온라인 웹사이트로 이동했다. 이 웹사이트에서는 파이썬(Python) 프로그래밍 언어로 구현된 장고(Django) 웹 애플리케이션 프레임워크를 사용했다. 2004년에 출범한 이 웹사이트는 오슬란 사인뱅크(Auslan Signbank)라는 이름으로 불렸다(2004).

오슬란 사인뱅크를 지원하는 스크립팅 및 프로그래밍은 2014년부터 오픈소스가 되었고 깃허브(Github) 웹사이트(<https://github.com/signbank>)상에서 제공된다. 2018년 현재 오슬란의 항목 수는 7,298개에 이른다(이 가운데 대략 절반은 stand-alone 핵심 표제어이고 나머지 절반은 핵심 표제어가 변형된 형태이다).

사전 형식: 인쇄, 시디롬, 온라인 [슬라이드 7-14]

오슬란 어휘 데이터베이스를 이용하여 출판된 사전은 세 가지 형태로 나뉜다.

첫째, 인쇄 형식: 1987 《호주 수어(오슬란) 예비 수어사전》, 1989 《오슬란 사전: 호주 농인 공동체 수어사전》, 1997 《호주 수어: 새로운 오슬란 사전》, 2003 《오슬란 생존 가이드: 초보자를 위한 호주 수어 포켓 사전》.

둘째, 시디롬 멀티미디어 형식: 1997 《호주 수어: 새로운 오슬란 사전》, 2002 《오슬란의 발견: 초보자를 위한 호주 수어사전》.

셋째, 웹기반 온라인 형식: 2004, 2008, 2018 오슬란 사인뱅크(<http://www.auslan.org.au>).

사전의 다양한 버전과 서로 다른 형식에도 불구하고 위에서 언급한 모든 사전의 구성 원칙이 바뀌지 않았다는 사실은 주목할 필요가 있다. 지난 수 년 동안 있었던 변화는 (1) 형식상의 특징에 따른 기호의 순차적 처리와 관련한 사소한 변경, (2) 용어 검색 및 분류 항목의 데이터베이스 기능 개선, (3) 기존 매체에서 새로운 매체로의 전환, 즉 인쇄 형식에서 시디롬 형식, 나아가 온라인 형식으로의 전환에 따른 적응 및 발전 정도이다.

어떤 형식으로도든 오슬란 어휘 데이터베이스는 호주 농인 공동체의 수어 어휘를 정확하게 문서화하고자 한다.

사전 구조 [슬라이드 15-17]

오슬란 사전은 모두 제한적인 양방향, 2개 국어 사전이었다. 단, 한 가지 예외가 있었는데 그것은 바로 《오슬란 생존가이드: 초보자를 위한 호주 수어 포켓 사전》(Johnston & Schembri, 2003) (아래 내용 참

조)이다.

진정한 사전이라면 한 언어의 단어나 기호는 단순히 나열되는 것이 아니라 정의가 내려진다. 2개 국어 사전에서는 A 언어의 단어나 기호가 B 언어로 정의 내려진다. 양방향 사전의 경우에는 B 언어로 정의 내려진 A 언어 단어 또는 기호의 목록과 A 언어로 정의 내려진 B 언어 단어 또는 기호의 목록으로 구성되어 있다. 그러한 이유에서 수어의 양방향, 2개 국어 사전은 어휘 정보에 접근하는 두 가지 방법이 존재한다. 하나는 수어를 통한 방법, 다른 하나는 음성언어를 통한 방법이다. 한 언어의 각 단어 또는 기호는 상대 언어의 단어 또는 기호로 정의 내려진다.

오슬란 사전의 본문은 영어(B 언어)로 정의 내려진 오슬란 기호(A 언어)를 열거한다. 오슬란의 표제어는 손 모양과 위치 등 언어 내적인 구성 원칙에 따라 순서대로 나열된다. 인쇄 사전의 표제어는 라인 일러스트로 나타났다. 1990년대 말까지는 데이터베이스의 표제어 또한 함노시스(HamNoSys)를 이용한 전용 수어 전사로 나타냈으나 일반인이 볼 수 있는 것은 아니었고 인쇄 사전에 실리지도 않았다.

함노시스를 이용한 전사는 기호의 형태를 알고자 할 때 더 이상 필수적인 것으로 간주되지 않는다. 영상 클립(또는 이전 버전의 경우에는 라인드로잉)이 포함된 필드를 확인하는 것만으로도 기호의 형태를 쉽게 알 수 있기 때문이다. 또 음운론적 특성과 관련하여 기호 항목을 분류 또는 설명하려고 할 때 전사가 필요하지 않다(과거에는 손 모양, 위치 등에 사용된 유니코드 심볼에 순서를 지정하는 방식으로 함노시스 전사를 분류하는 것이 가능했다.). 게다가 새로운 데이터베이스는 특별한 전사 스크립트에 의존하지 않고도 기호의 음운론적 특징 또는 그 밖의 특징을 태그로 달 수 있는 상당수의 필드를 수월하게 수용할 수 있다. 함노시스 전사는 1990년대 말까지 생성되었고 이후에는 디지털 영상 기술과 데스크톱 성능이 전용 영상 클립으로 각각의 표제어를 나타내는 것이 가능할 만큼 충분히 강력해졌고 비용도 저렴해졌다. 결국 함노시스 전사는 중단되었다. 그에 따라 오슬란 어휘 데이터베이스의 모든 항목은 영상 클립으로 만족해야 하며 음운론적 형태를 기호화하는 관련 필드를 지니고 있다.

인쇄, 시디롬, 웹사이트 형식의 사전에는 오슬란-영어란 또는 데이터를 볼 수 있는 뷰가 있다. 하지만 세 가지 형식 모두 영어-오슬란 뷰(B 언어-A 언어)는 제한적이다. 인쇄 형식의 경우 그 이유는 단순하다. 라인드로잉 또는 함노시스를 이용하여 오슬란으로 영어 표제어를 정의하기란 사실상 불가능했기 때문이다. 따라서 비교 방향은 단순히 각 단어에 해당하는 오슬란 번역어의 항목 번호를 사용자에게 알려주는 영어 단어 색인을 통해 드러냈다. 시디롬과 사인뱅크 웹사이트 형식의 경우에는 음운론적 특성(손 모양, 위치, 움직임 등) 및 관련 영어 키워드에 대한 각 기호 항목(기호 형태)을 태그할 수 있다. 사용자는 시디롬 또는 사인뱅크 웹사이트의 검색창에 영어 단어를 입력하여 해당 키워드와 관련 있는 기호를 찾는다. 따라서 여전히 양방향 사전의 특징을 지니고 있지만 제한적인 의미에서만 그러하다.

오슬란 사전은 형식상의 특징(아래 참조)에 기반한 원칙에 따라 표제어를 분류하고 배열했다. 본 사전은 수어 초반 우세 손의 모양을 바탕으로 하여 수어를 손 모양 섹션으로 분류한다. 각 섹션 내에는 그 외에도 손의 위치와 방향, 움직임 및 손의 개수, 그리고 경우에 따라서는 얼굴 표정 등 형식상의 특징이 담겨 있다. 비(非)우세 손의 모양과 기호의 대칭은 우세 손 모양이 동일한 기호의 분류 및 배열을 결정한다.

1부 발표

인쇄 사전의 정렬 원칙(알파벳 순서) [슬라이드 18-25]

여기서는 기호 항목의 ‘알파벳 순서’ 근간을 이루는 주요 원칙만을 다룬다. 더 자세한 내용은 존스턴(Johnston, 2003년)을 참조하기 바란다.

손 모양 기준 [슬라이드 18-20] 손 모양은 하나의 기호를 다른 기호와 구분하고 여러 기호를 섹션으로 분류하기 위한 원칙으로 삼기에 이상적이다. 오슬란의 손 모양은 어느 정도 자의적인 순서에 따라 정렬된다. 초창기 오슬란 사전(1987년, 1989년)의 경우에는 어휘에서 차지하는 손 모양의 빈도를 바탕으로 손 모양의 순서를 정하되 가장 빈도가 높은 손 모양(평평한 손 모양)이 가장 먼저 나온다. 《호주의 수어》(1997)에서는 수를 셀 때 사용하는 오슬란 손 모양을 바탕으로 더 자연스럽고 직관적인 순서를 사용했다. 순서는 0부터 시작해서 수를 셀 때 사용하는 손 모양을 따른다. 이러한 손 모양의 미세한 변형은 수를 세는 주요 손 모양들 가운데에서 순서가 정해진다. 예를 들어 둥글게 오므린 손(Round Hand)은 오슬란 손 모양으로 숫자 0을 의미하기 때문에 사전상에서 가장 먼저 등장한다. 오케이 모양의 손(Okay Hand)은 숫자 0을 의미하는 대체 가능한 손 모양이기 때문에 두 번째로 온다. 그 다음은 숫자 1을 의미하는 검지를 세운 손(Point Hand)이다. 검지를 세운 손과 비슷한 갈고리 모양의 손(Hook Hand)이 그 뒤에 나오며 이와 같은 식으로 숫자 2부터 12까지의 손 모양을 통해 순서가 정해진다(오슬란에서는 12까지의 숫자 가운데 그 어느 것도 낮은 숫자에서 이미 나온 손 모양을 사용하지 않는다). 오슬란에서 사용하는 손 모양 중 몇 가지는 숫자와 아무런 관련이 없으며 임의적으로 목록과 사전의 마지막에 온다. 예를 들어 ‘SUN’은 사전 속 둥글게 오므린 손(Round Hand) 섹션, ‘IGNORE’는 검지를 세운 손(Point Hand) 섹션, ‘FRONT’는 평평한 손(Flat Hand) 섹션에 나온다[슬라이드 20].

기본 위치 기준 [슬라이드 21-22] 그 다음으로 오슬란 사전은 위치에 따라 기호의 순서를 나열한다. 수어 공간 또는 신체상의 높은 위치에서 표시한 기호(예: 머리 또는 얼굴 위에서의 기호)는 낮은 위치에서 표시한 기호(예: 허리 또는 팔 위에서의 기호)보다 먼저 온다. 그러한 이유에서 ‘THINK’는 ‘STOMACH’보다 앞서는 ‘WHO’의 앞에 온다.

손의 개수 기준 [슬라이드 23] 각각의 손 모양 섹션 내에서 한 손만을 사용하는 기호가 두 손 모두를 사용하는 기호보다 먼저 나온다. 예를 들어 한 손만 사용하는 기호 ‘CONSIDER’는 두 손을 모두 사용하는 변형 기호보다 앞선다.

동일한 또는 다른 손 모양 기준 [슬라이드 24-25] 각각의 손 모양 섹션 내에서 두 손을 사용해야 하는 기호는 한 손만 사용해야 하는 기호보다 뒤에 나온다. 두 손을 사용하는 기호 내에서는 양손의 모양이 같은 기호가 비(非)우세 손의 모양이 다른 기호보다 먼저 나온다. 그러한 이유에서 ‘CONSIDER’는 ‘CROSS-OUT’보다 앞선다. 두 손을 사용하지만 비(非)우세 손의 모양이 다른 기호의 순서는 단순히 사전 전반의 손 모양 순서를 따른다. 결국 우세 손의 모양이 해당 기호의 손 모양 섹션을 결정하고, 비(非)우세 손의 모양이 우세 손 모양이 동일한 두 손 기호들 사이의 순서를 결정한다. 그러한 이유에서 ‘QUESTION’, ‘POTATO’, ‘DEBT’는 상대적으로 이 같은 순서를 따른다. 오케이 모양의 손은 검지를 세운 손보다 앞서고,

검지를 세운 손은 평평한 모양의 손보다 앞서기 때문이다.

생존 가이드 구조 [슬라이드 26]

《오슬란 생존 가이드: 초보자를 위한 호주 수어 포켓 사전》(Johnston & Schembri, 2003)은 정확한 의미에서 사전이라기보다는 단어 목록이라 할 수 있다. 그러한 이유에서 과거에 비언어학자들이 만든 대부분의 이른바 수어사전들과 다를 바 없었다. 이들은 본질적으로 단일어 단어 목록이었으며 표제어는 사실 해당 언어의 알파벳 문법에 따라 배열된 공동체 음성언어의 단어였다. 표제어와 동일한 기호의 일러스트는 바로 옆에 위치해 있었다(《생존가이드》는 한정된 특정 독자를 염두에 두고 출판되었다. 농인과 기본적인 의사소통을 할 수 있을 정도의 매우 기본적인 어휘력을 쌓고 싶지만 시간적 여유가 많지 않은 성인이 그 대상이다). 이 같은 유형의 단순한 어휘 목록에 대한 특정한 틈새 분야가 있지만 오늘날 언어학자들이 만들고 있고 본 강의에서 다루고 있는 진정한 수어사전과 혼동해서는 안 된다.

시디롬 구조 [슬라이드 27-33]

시디롬 버전의 〈호주 수어〉(1997)는 오슬란 최초의 멀티미디어 사전이다. 오슬란 어휘 데이터베이스는 시디롬 제작에 필요한 데이터를 제공했다.

시디롬의 가장 중요한 혁신을 꼽자면 시디롬 형식상의 특징을 기반으로 기호를 검색할 수 있다는 점이다. 친숙한 ‘영어 단어로 검색’하는 방법(인쇄 형식 사전의 영어 색인을 참고하는 방법) 대신 ‘오슬란으로 검색하기’ 옵션 버튼을 누르는 방법으로 기호 검색을 요청할 수 있다. 사용자는 시디롬상에서 ‘오슬란으로 검색하기’ 옵션을 선택한 후 다음의 구체적인 내용을 묻는 일련의 추가 질문에 답해야 한다. (1) 한 손을 사용하는 기호인지, 두 손을 사용하는 기호인지의 여부, (2) 수어 시작 시 우세 손의 모양, (3) 수어의 위치, (4) 두 손을 사용하는 기호의 경우 비(非)우세 손의 모양, (5) 2차적인 위치(즉, 두 손이 만나거나 서로 상호작용하는 경우의 비(非)우세 손의 위치)

그 다음으로 중요한 시디롬의 혁신은 시멘틱 필드(음식, 신체 부위, 감정, 건강 등) 또는 방언(호주의 주요 방언 지역 또는 주 기준)과 같이 다양한 기준에 따라 즉시 기호를 선택하고 분류할 수 있다는 점이다.

세 번째로 중요한 시디롬의 혁신은 미리 설정된 여러 링크(예: 동의어, 반의어, 동형이의어)를 기반으로 사전 내의 기호 사이를 즉시 이동할 수 있다는 점이다.

온라인 사인뱅크 구조 [슬라이드 34-39]

오슬란 어휘 데이터베이스는 또한 일반적으로는 사인뱅크(Signbank), 구체적으로는 오슬란 사인뱅크(Auslan Signbank)라고 하는 온라인 버전 오슬란 사전의 근간을 이룬다. 사실 데이터베이스 자체는 현재 웹사이트 내에 있다. 다시 말해 오슬란 사인뱅크는 오슬란 어휘 데이터베이스의 웹사이트명에 불과하다.

1부 발표

오슬란 사인뱅크는 단순히 데이터베이스상의 웹 포털인 것이다. 하지만 연구원들은 .csv 파일 형태로 사인뱅크 웹사이트상의 내용을 다운로드 받아 추가 데이터 처리나 조사를 목적으로 다양한 데이터베이스로 통합할 수 있다(따라서 데이터베이스 버전은 여전히 파일메이커 프로 및 기타 스프레드시트 형식으로 남아있다.).

모든 이전 버전의 오슬란 사전과 마찬가지로 오슬란 사인뱅크는 영어 정의와 함께 표제어로 표시된 언어의 전통적인 어휘적 기호를 포함하고 있다. 오슬란 사전에는 묘사하는 기호 또는 식별자 기반의 기호가 다수 포함되어 있는데 이들 기호의 어휘 상태에는 의심의 여지가 있다.

오슬란 사인뱅크상에서 사전을 보여주는 방식에 몇 가지 혁신이 적용되어 있다. 첫째, 인터넷 접속이 가능한 사람이라면 누구든 퍼블릭 뷰를 통해 이용 가능하다. 둘째, 등록 사용자만 이용 가능한 제한적 뷰 또는 프라이빗 뷰가 있다. 등록 뷰의 경우 데이터베이스에 대한 접근이 여러 단계로 나뉘어져 있다. 일반 사용자는 웹사이트에 대한 피드백(사전상의 기호 오류 또는 누락 보고 등)을 작성하기 위해 등록을 해야 한다. 그 외에도 애플리케이션, 통역사, 연구자, 편집자, 출판사를 위한 네 단계의 조사용 접근이 가능하다. 데이터베이스 내에는 이와 관련하여 다양한 접근 및 편집 권한이 존재한다. 데이터베이스 열람 및 변경 접근은 다양한 등급의 사용자들에게 웹사이트에 대한 서로 다른 단계의 권한을 부여하는 방식으로 통제한다. 추가적인 페이지 뷰에서는 이미 시디롬에서 이용 중인 유형의 정보를 보여준다. 하지만 시디롬과는 달리 오슬란 사인뱅크는 어느 때라도 업데이트가 가능하며(단, 접근 권한을 갖고 있다는 전제 하에) 이때의 변경 사항은 즉시 효력을 발휘하여 뷰에 반영된다. 또한 새로운 정보가 이용 가능해지면 기록은 삭제되거나 추가될 수 있다.

이는 오슬란 사전이 공동체의 피드백 또는 오슬란 말뭉치의 점검을 통해 확보한 새로운 정보에 신속하게 대응할 수 있음을 의미한다(다음 섹션 참조).

오슬란 사인뱅크의 2004년 초기 웹버전은 당시의 스크립트 언어에 따른 제약 때문에 시디롬의 기능 몇 가지만을 지니고 있었다. 2004년과 2018년 사이에 오슬란 사인뱅크에 사용된 스크립트 언어는 전면적으로 개선되었고 이전까지 시디롬에서만 제공하던 기능이 웹사이트에서도 이용할 수 있게 되었다. 하지만 현재 이 기능의 대부분은 특별 접근 권한을 부여받은 등록 사용자만 이용 가능하다.

수어는 인쇄 버전의 사전에서 먼저 부여받은 내재적 순서를 지니고 있지만 웹사이트상에서의 기호 순서는 인쇄 사전에서만 명확하지 않다. 각각의 페이지는 단 하나의 기호 항목을 보여주므로 사용자가 사전 한 페이지에서 다음 페이지로 이동하는 경우에만 순서가 분명히 드러난다.

가장 최근인 2018년에 이행된 오슬란 사인뱅크의 주요 혁신은 ‘오슬란상의 정의’ 필드가 제공되는 것이다. 이 필드에는 영상 클립이 삽입될 수 있으며 이때 해당 표제어에 대한 오슬란상의 정의가 제공된다. 이러한 유형의 항목이 생겨난 것이 불과 최근의 일이지만 향후 단일어 사전에서처럼 오슬란 사인뱅크 항목을 보여주는 것이 가능할 수 있음을 시사한다. 이는 획기적인 일이다. 어떤 수어라도 단일어 사전은 아직까지 출판된 적이 없다. 수어용 문자체계가 존재하거나 개발되지 않고서는 인쇄 형태의 수어 단일어 사전이란 사실상 불가능했다. 비록 영상의 도래에 힘입어 단일어 수어사전이 이론상 가능해졌지만 실행은 여전히 불가능했다. 그래도 디지털 영상기술이 나오면서 단일어 사전 제작이 기술적으로 가능해졌지만 아직까지 제작되지

나 출판된 사전은 하나도 없다. 최신 2018년 버전 오슬란 사인뱅크는 현재 단일어 항목 생성을 가능하게 해 준다. 오슬란을 이용하여 오슬란 표제어가 정의 내려지는 항목 말이다. 이 기술을 이용한 최초의 항목이 생성되기 시작했다.

파트 2: 말뭉치 중심 수어사전 [슬라이드 40-70]

말뭉치 중심 수어사전에 대한 근거 [슬라이드 40-44]

앞에서 설명했듯이 오슬란 사전은 참여자 관찰, 면담, 도출 등을 수반한 '전통적' 언어 기록과 서술 연구 프로그램의 산물이다. 이와 같은 모델은 아직 문서화되지 않고 서술되지 않은 언어에 대한 연구를 목적으로 언어학과 인류학에서 폭넓게 사용되고 있다. 일반적으로 이러한 유형의 연구를 진행하는 언어 공동체는 대체로 서술의 역사가 없고 지리적으로 동떨어져 있으며 소규모인 전통사회에서 찾아볼 수 있다.

하지만 수어 공동체는 이와는 다른 경향을 보인다. 거의 모든 곳에서 공동체 수어는 음성 언어(및 때로는 문자 언어)를 사용하고 구성원 대부분이 청인인 공동체 내의 소수 농인들이 사용하고 있다. 일부 국가 또는 지역의 몇몇 농인들은 수 세대에 걸쳐 청각장애를 앓아온 가정에서 태어나지만 이는 비교적 드문 경우이다. 농인 대부분의 부모와 형제자매는 청각장애를 갖고 있지 않다. 그러므로 대부분의 수어를 습득하여 사용하는 것은 음성 언어를 습득하여 사용하는 것에 비해 매우 이례적인 것으로 보인다. 농인 공동체의 잘 알려진 사회언어학적 특성에 따르면 상당수 연구자들은 수어 공동체가 다른 예상보다 어휘와 문법 양 측면에서 훨씬 높은 수준의 변화를 보이는 듯하다고 말한다.

1990년대 말과 2000년대 초 오슬란 연구자들을 포함한 상당수 수어 연구자들에게 분명해진 사실이 있다. 그것은 바로 수어의 어휘와 문법에 대하여 경험적으로 건전한 서술을 하려면 원어민 또는 그에 가까운 수어 사용자의 자연스러운 수어를 바탕으로 언어학적 말뭉치를 만들어야 한다는 점이다(Johnston, 2010, 2014).

언어학적 말뭉치를 만드는 일은 기계가 읽을 수 있도록 전사, 주석, 번역을 통해 언어 사용 기록을 변환하는 작업이 수반된다. 수어의 경우, 물론 이는 디지털 영상 기록물을 제작하고 멀티미디어 주석 소프트웨어를 사용하는 것을 의미한다.

일단 수어 말뭉치를 만드는 준비 과정에서 기록 보관소가 마련되면 해당 수어의 기존 수어사전은 여기에 적응해야 할 필요가 있다. 예를 들어 식별되어 말뭉치에 주석이 달린 기존의 모든 어휘 기호들 또한 사전에 실려야 한다. 그렇지 않을 경우 해당 사전은 불완전한 것으로 여겨질 것이다. 사실상 사전과 말뭉치 간의 관계는 두 가지 방향으로 존재한다. 한편으로는 사전이 기호의 식별과 주석을 알려주고, 다른 한편으로는 말뭉치가 지금까지 문서화되지 않은 기존 기호와 관련하여 잠재적으로 새로운 데이터를 제공하거나 이미 문서화된 형태의 가능한 변이를 식별해준다.

1부 발표

말뭉치 주석과 사전 내 기호 항목간의 결부 [슬라이드 45-49]

식별용 주석

사전과 말뭉치가 조화를 이루기 위해서는 사전에 입력된 기호 유형의 형태가 말뭉치 내의 토큰(token)과 부합할 수 있도록 해야 한다. 이 말뭉치 중심 데이터베이스의 기호 항목은 영어로 된 키워드와 고유의 식별용 주석(Identifying-gloss, ID-gloss)과 관련이 있다. 고유의 식별용 주석은 사전 내 각각의 표제어마다 배정되어 있다. 사전 내 기호 유형과 관련한 식별용 주석은 말뭉치 내 기호 유형의 모든 토큰에 주석을 다는 목적으로 사용된다. 하지만 식별용 주석을 효과적, 효율적으로 사용하려면 어휘 유형이 문서화되어 있는 참조 데이터베이스가 필요하다. 이것이 바로 말뭉치 중심 수어사전의 정수이다. 해당 언어의 모든 기존 어휘 기호는 사전상에 등재되며 말뭉치 주석으로 분할되어 표시를 할 때 고유의 식별자를 갖게 된다.

따라서 말뭉치에 주석을 다는 동안 수어사전에 끊임없이 참조 표시를 해야 한다. 이를 위해 오슬란 주석자는 오슬란 사인뱅크를 참조한다. 주석자는 특정 기호의 식별용 주석에 확인이 필요할 때면 언제든지 해당 기호와 관련 있는 영어 키워드를 오슬란 사인뱅크 검색창에 입력하고 사인뱅크에서 제시하는 일치 여부를 확인한다. 그 후 주석자는 말뭉치에서 주시하고 있는 기호의 형태와 가장 유사한 기호의 식별용 주석을 선택한다.

현재 오슬란 사인뱅크는 말뭉치 구축에 사용되는 ELAN 주석 프로그램의 외부에 있다. 주석자는 ELAN 상에서 주석을 만들 때 데스크톱상에 온라인 사전이 열려 있도록 할 뿐이다. 하지만 사인뱅크 NGT(오슬란 사인뱅크 기반의 네덜란드 수어용 사인뱅크. 뒤에서 이를 다룰 계획이다)에서 둘은 서로 연결되어 있다(Crasborn & Meijer, 2012; Crasborn 등, 2012). 예를 들어 사인뱅크 NGT는 NGT 말뭉치의 주석을 다는 데 사용되는 ELAN 소프트웨어 템플릿과 연결되어 있다. 주석자는 ELAN상에 기호의 식별자 주석을 입력할 수 있으며 사인뱅크 NGT상의 모든 관련 식별자 주석을 포함하는 통제 어휘가 생겨나게 되면 최적의 조합을 선택할 수 있을 것이다. 만약 의구심이 들 경우 주석자는 통제 어휘 내의 어떠한 주석이라도 클릭할 수 있고 사인뱅크로부터의 식별용 주석을 위한 섬네일 영상 클립이 팝업창으로 나타나게 된다.

음운 변이 [슬라이드 50-51]

가장 흔한 형태의 기호(말뭉치의 토큰 카운트에서 가장 빈번하게 발생하는 변이의 횟수로 평가)는 어휘 기호의 어간 또는 기본 형태로 여겨진다. 바로 중심 표제어인 것이다. 표제어의 변이는 해당 항목에서 드러난다. 근본적으로 변이는 기본 기호와 다른 형식상의 특징이 하나 또는 많아야 둘인 기본 기호의 한 형태이다(변이는 보통 다른 손 모양, 위치 또는 수어에서 사용되는 손의 개수를 수반한다.). 또한 변이는 형태 기반 사전의 적절한 위치에 고유의 항목 기록을 갖고 있다. 만약 말뭉치에 기록된 것처럼 사용 데이터에 따라 기본 및 변이 형태의 지위가 뒤바뀌어야 된다고 한다면 그대로 될 것이다. 예를 들어 사용 데이터는 두 손을 모두 사용한 형태의 기호가 한 손만을 사용한 형태보다 훨씬 더 빈번하다는 점을 보여줄 수 있다. 만약

한 손만을 사용한 형태가 사전상에 중심 표제어로 올라와 있을 경우 관계는 뒤바뀌어야 한다.

입증된 사용 [슬라이드 52]

시간이 지나면서, 그리고 말뭉치가 모든 사용자와 모든 텍스트 유형을 대표하고 또 충분히 크다면 수어사전 항목의 대부분이 말뭉치를 통해 입증된다는 점을 확인할 수 있다. 일반적인 말뭉치에서 찾아보기 힘든 기호는 너무 오래 전에 사용되어 쓸모없어졌거나 중요도가 떨어지는 기호 정도일 것이다(예: 특정 분야의 소규모 사람들만이 사용하는 어휘). 말뭉치에서 입증된 사용은 사전이 경험적으로 볼 때 견고하다는 점을 보여준다. 2012년, 4년 간의 주석 사용 이후(Johnston 2012 참조) 오슬란 사인뱅크의 3,500개 핵심 항목 가운데 약 75%가 오슬란 말뭉치에서 증명되었다. 당시 오슬란 말뭉치의 규모(약 70,000 토큰)가 음성 언어의 말뭉치에 비해 작다는 점을 고려하면 이는 매우 높은 비율이다(오슬란 사인뱅크의 항목은 7,000여개로 두 배가 넘지만 이는 모두 3,500개 핵심 어휘 기호가 변이된 형태이다).

누락 데이터 [슬라이드 53]

물론 어떠한 사전도 완전하지는 않기 때문에 지금까지 기록되지 않은 기호가 말뭉치에서 식별될 가능성도 높다. 만약 해당 언어의 진정한 기존 어휘 기호로 확인된다면 일정 기간 동안 공동체 구성원들과의 협의 및 말뭉치에서 발견한 추가 토큰의 확인을 통해 사전에 추가될 수 있다. 이런 식으로 말뭉치는 지금까지 기록되지 않은 기호로 어휘 데이터베이스를 확대할 수 있다. 그것이 기존 어휘 항목의 변이형태이든, 언어에 새롭게 추가된 기호이거나(신조어 또는 차용어) 단순히 이전까지 간과되었던 기호이든 말이다. 약 50개의 새롭게 추가되거나 이전에 기록되지 않았던 기존 어휘 기호가 오슬란 말뭉치에서 식별되었고 4년 간의 주석 사용 이후 오슬란 사인뱅크에 추가되었다.

사용 데이터에 근거한 정의 개선 [슬라이드 54]

애초에 말뭉치에서의 입증된 사용을 고려했을 때 지나치게 좁은 의미가 기호의 형태에 부여되었을 수 있다. 그러한 경우 해당 기호의 실제 자연스러운 사용을 반영하여 사인뱅크 내의 정의를 개선하게 된다. 이는 정의에 대한 변경 및 개선, 또는 관련 키워드의 추가, 제거 또는 변경이 될 수 있다.

일부 기호의 어휘 지위 조정 [슬라이드 55]

초기 현장 조사에 근거하여 일부 매우 상징적이거나 묘사적인 기호가 기존 어휘 기호로 사전에 포함되었을 수 있다(직관, 자기 성찰, 도출 또는 공동체 피드백 결과 관습적이고 어휘화된 것으로 보고된 기호이다). 하지만 말뭉치에서 입증된 것처럼 기호를 사용한다는 것은 기호가 애초에 생각했던 것만큼 구체적이거나

1부 발표

관습적(또는 ‘어휘적’)인 의미를 지니고 있지 않음을 시사할 수 있다. 그와 같은 기호는 어휘화 측면에서 경계선상의 기호이다. 즉 의미 측면에서 해당 언어의 특징을 지니지 못하고 그러한 이유로 오슬란 사전에 적절하게 포함되지 못했을 수 있다(물론 해당 언어의 문법 측면에서 논의가 필요할 것이다). 따라서 그러한 항목은 그 지위를 제대로 반영할 수 있도록 업데이트되거나 아예 제거되어야 할 수 있다.

말뭉치에서 사전까지? [슬라이드 56-57]

이상적인 말뭉치 구축 상황에서는 그 누구도 먼저 언어에 대하여 기본적인 어휘론적 연구를 수행한 다음 그 결과로 나온 초기 어휘를 데이터베이스나 사전에 기록 및 반영하는 일 없이 수어의 기록을 분할하고 주석을 달려고 하지 않을 것이다. 하지만 최근 몇 년 동안 대규모 수어 말뭉치 프로젝트의 일환으로 몇 가지 수어사전 프로젝트가 시작되었다. 그 목적은 말뭉치를 사용하여 사전을 만드는 것에 있다.

거의 아무 것도 없는 상황에서 주석 자체를 바탕으로 참조 어휘를 구축하는 것은 가능한 일이다. 즉, 어휘의 기초로 말뭉치 자체의 입증된 사례만을 사용하는 방법이다. 이와 비슷한 접근방식이 BSL 말뭉치와 관련하여 설명되었다(Cormier 등, 2012). 비록 이런 식으로 진행할 수도 있겠지만 이 접근방식은 어렵고 많은 시간이 소요되며 일관성이 떨어지기 쉽다. 특히 주석 작업 초기에 주석자가 이용할 공통의 참조점이 없기 때문이다. 연구자들은 초기의 데이터베이스와 사전을 끊임없이 보완하고 수정해야 할 상당수의 다양한 사용 및 변이 형태와 맞닥뜨릴 가능성이 높다. 이는 커져가는 어휘 데이터베이스에도 연쇄효과를 일으킬 수 있다(Crasborn, Meijer, 2012).

실제로 일부 연구자들은 실증적인 수어사전이 참조 말뭉치상에서 입증된 기호에만 의존해야 한다고 생각한다. 어쩌면 혹자는 이를 두고 본 강의에서 언급한 말뭉치 중심 수어사전이라기보다는 말뭉치 기반 또는 말뭉치 주도 수어사전이라 할 것이다. 하지만 수어사전에 기호를 포함시키기 위해 것처럼 엄격한 기준을 적용하기에는 위험이 따른다. 한 언어의 어떠한 말뭉치라도(무수히 많은 단어나 기호와 접하지 않는 한) 해당 언어의 알려진 모든 단어나 기호의 최소 하나의 토큰을 포함할 가능성은 매우 낮다. 만약 매우 기본적인 일상 어휘를 말뭉치에서 찾아볼 수 없게 된다면 특히 문제가 되며, 규모가 그리 크지 않다는 점에서 이미 구축되었거나 구축 중인 수어 말뭉치의 경우에는 그렇게 될 가능성이 매우 높다. 수어사전의 사용자라면 누구라도 기본 개념을 위한 기호가 기록되어 있지 않다는 사실을 깨닫게 될 것이다. 말뭉치상에서 과거에 단 한 번도 없었던 일이기 때문이다. 이는 만족스럽지 못한 일이며 사전의 사용자는 사전이 불완전하다는 타당한 결론을 내릴 것이다. 만약 입증된 데이터를 제한하고 싶다면 이 문제를 피할 수 있는 방법 한 가지는 관습적인 또는 서술적인 텍스트뿐만 아니라 도출된 텍스트와 어휘까지 아우르는 말뭉치를 자체적으로 구축하는 것이다. 그렇게 되면 기본적인 어휘를 도출하는 기록 과정에서 작업을 통합할 수 있다.

말뭉치 결과를 이용한 사전 항목 강화 [슬라이드 58-68]

오슬란 어휘 데이터베이스와 더불어 오슬란 사인뱅크는 각 항목의 각 기호에 대한 매우 다양한 정보를 수

용할 수 있도록 만들어졌다. 해당 정보는 수많은 전용 필드나 셀에서 찾을 수 있다. 여기에는 다음에 관한 정보가 담겨 있다. (1) 표제어의 형태, (2) 표제어의 의미, (3) 표제어의 사용자, (4) 표제어의 검색이나 분류 또는 온라인상 표시를 위한 기준. 이들 정보 상당 부분은 말뭉치 자체를 조사해 보면 찾을 수 있다(예: 방언 선호도, 연령 집단, 학교 기반 기호, 소속 종교, 빈도 등). 이 정보는 끊임없이 검토되며 사인뱅크 연구자용 인터페이스상에서 쉽게 편집 가능하다. 일부 정보는 일반인이 즉시 열람할 수 있고 또 다른 일부 정보는 데이터베이스의 프라이빗 뷰에서만 연구자가 참고 목적으로 열람 가능하다.

동형이의어 처리 [슬라이드 69-70]

오슬란 사인뱅크는 형태 기반의 기호 데이터베이스이기 때문에 만약 두 기호가 동형이의어인 경우, 다시 말해 오슬란상에서는 서로 다른 의미를 갖고 있지만 기호 형태가 동일한 경우(예: ‘금속 용기’와 ‘할 수 있는’을 의미하는 영어 단어 ‘can’), 두 기호 모두 동일한 형태를 참조하게 된다. 오슬란 사인뱅크에서의 해결책은 영어 사전에서와 마찬가지로이다. 즉 표제어가 되풀이되고 생성된 두 개(또는 그 이상)의 항목에서 동일한 음운론적 형태에 대하여 두 가지(또는 그 이상)의 의미로 각각 규정한다. 오슬란 사인뱅크에서는 의미 1, 의미 2 등과 같은 식으로 번호를 부여한다.

다수의 수어 [슬라이드 71-78]

오슬란 사인뱅크 & 기타 사인뱅크 [슬라이드 72]

오슬란 어휘 데이터베이스 구조의 기본 원칙은 처음 만들어져 데이터베이스에 적용된 이후로 어떠한 수어에도 적용 가능했다(Johnston, 2001). 오슬란 사인뱅크(및 인쇄, 시디롬 초기 형식)의 표제어는 형태에 기반한 것이었다. 이들 표제어는 그 형식상의 특징이 전용 전사 시스템(함노시스 등) 또는 각 기호가 지닌 관련 형식상의 특징(손 모양, 위치 등)의 존재 여부를 명시한 일련의 필드에 기호화되어 있다. 그 결과 어떠한 수어의 기호라도 데이터베이스 내에 수용될 수 있다. 단순히 형태별로 서술된 항목이 생성되는 것이다(만약 데이터베이스가 형태별로 정렬된다면 새로운 기호는 알파벳 순서에 따라 적절한 자리에 들어가게 된다.). 심지어 기호들 사이의 상관관계 유형(가령 동의어, 반의어, 동형이의어, 의미 필드 항목, 방언 분포 등)이 프로그램 속으로 구축되어 있다면 이 또한 어떠한 수어의 경우에도 적절하게 이용 가능하다.

이러한 잠재력이 실현되자 다른 수어를 연구하는 타 언어학자들도 오슬란 사인뱅크의 기본 아키텍처를 본인의 수어 연구에 도입하게 되었다. 그 결과 네덜란드 수어(NGT) 사인뱅크, 영국 수어(BSL) 사인뱅크, 핀란드 수어(FinSL) 사인뱅크의 온라인 사전이 제작되었고 미국 수어(ASL) 사인뱅크도 작업에 착수했다.

1부 발표

동음어에서 다수의 수어까지 [슬라이드 73-77]

오슬란 사인뱅크 구조의 한 가지 중요한 점에 힘입어 단일 사인뱅크, 다시 말해 범용 어휘 데이터베이스(Universal Lexical Database)에 다수의 수어를 수용하는 어휘 데이터베이스 구축의 가능성이 열렸다. 그러한 데이터베이스는 기존 오슬란 사인뱅크를 타 사인뱅크의 템플릿으로 사용하는 데 그치지 않고 범용 사인뱅크(Universal Signbank)를 구축해준다.

앞에서 언급한 것처럼 오슬란 사인뱅크는 형성 과정의 특징에 따라 수어 형태를 기록해 낸다. 그 결과 인쇄 형식으로 제공된 경우에는 라인드로잉 또는 사진이 시간적인 순서로 정렬되고, 스프레드시트의 기록과 셀의 형태로 제공된 경우에는 유사한 영상 클립들이 늘어서게 된다. 비록 사인뱅크 같은 온라인 사전에서는 페이지당 하나의 항목만 눈에 보이지만 동일한 원칙에 따라 각 항목에 기호화된 관련 형식상의 특징별로 데이터베이스 내에서 기록을 즉시 분류할 수 있다. 그러한 이유로 기호는 손 모양(인쇄 형식에서처럼)이나 위치, 또는 요청받은 다른 특징별로 정렬할 수 있다.

더욱 중요한 점은 이러한 형태 기반 항목들이 쉽게 기호화될 수 있고 음운론적 변이, 방언 변이, 의미 관계(반의어, 동의어, 동형이의어) 및 타 수어(영국 수어, 미국 수어 등)와의 관계를 대변한다는 사실이다. 오슬란 사인뱅크의 기호 형태는 이제 방언뿐만 아니라 수어에도 태그로 달릴 수 있다. 태그는 기호 형태가 영국 수어나 미국 수어에도 존재한다는 것을 의미한다.

언어에 태그를 달 수 있고 더불어 다른 기호와 동형이의어 관계에 있는 기호에 태그를 달 수 있게 되면서 확장 가능한 데이터베이스의 가능성이 열렸다. 어떠한 수어로부터의 기호라도 동일한 데이터베이스에 들어갈 수 있다. 오슬란 어휘 데이터베이스가 처음 구상되었을 당시에만 해도 컴퓨터의 데이터 처리 속도(및 비용), 디지털 영상기록물의 품질, 서버의 저장공간(및 비용)에 실질적인 제약이 있었다. 오늘날에는 사인뱅크의 항목 수, 각 항목과 관련한 코드의 수와 유형, 표제어나 정의에 사용되는 영상 클립의 길이 등은 더 이상 실행 불가능하지도, 많은 비용이 들지도 않는다.

그 결과 이러한 유형의 범용 사인뱅크의 가능성과 사용을 시험해보려는 노력이 시작되었다. 이는 글로벌 사인뱅크(Global Signbank)라 한다.

글로벌 사인뱅크? [슬라이드 78]

글로벌 사인뱅크는 실현 가능한가? 지금까지 심층적으로 연구되어 온 수어 어휘의 두 가지 관찰에 따르면 그 가능성은 충분하다.

첫째, 규모이다. 지난 50여 년 동안 출판되어 온 수어사전 대부분은 수어 표제어의 수가 수천 개를 넘지 못한다(3,000~8,000개 사이). 이는 대다수 음성 언어의 사전, 심지어는 문자가 존재하지 않는 음성 언어와 비교해 보더라도 낮은 수치이다.

둘째, 상당수 연구자들은 수어의 도상성이 높다 보니 아주 똑같거나 매우 유사한 형태의 기호가 서로 다른 수어에서 거듭 나타나는 상황이 발생하곤 한다고 밝혔다. 비록 도상적으로 자극을 받은 이들 기호 형태가

종종 서로 관계없는 수어 사이에서 똑같거나 유사한 의미를 갖기도 하지만 그 의미는 전혀 다를 수도 있다. 그럼에도 불구하고 이들 수어는 동일한 기호 형태를 공유하고 있다. 한 가지 수어를 사용하는 사람들이 외국 수어를 본 후에 형태 측면에서 본인이 사용하는 수어의 기호화 매우 유사해 보인다고 말하는 경우가 종종 있다. 때로 그들은 그 의미를 정확하게 추측할 수 있을 것이고 그러지 못할 때도 있을 것이다. 반면 서로 관계없는 음성 언어 사이에서 동일한 단어 형태가 거듭 나타나는 경우는 매우 드물다. 그 이유는 다음과 같다. 대부분의 음성 언어에서 음운론적으로 형태가 올바른 단어의 잠재적 수는 음소와 가능한 음절 수의 조합을 고려했을 경우 사실상 무제한인 것으로 보이기 때문이다. 하지만 수어에서의 상황은 일부분 도상성의 영향 때문에 다를 수 있다.

가령 글로벌 사인뱅크가 실현 가능하다면 사인뱅크의 다국어 버전이라 할 수 있는 글로벌 사인뱅크는 현재 재구성되고 있는 구성상의 순서로부터 가장 큰 수혜를 입을 가능성이 높다. 글로벌 사인뱅크는 심지어 손 모양이 앞으로도 가장 ‘중요한’ 수어의 특징으로 남을 가능성이 높다 하더라도 대다수 수어 연구자들이 받아들이는 ‘중립적인’(다시 말해, 특정 언어의 특징을 지니지 않는) 구성상의 순서를 갖게 될 것이다. 예를 들어 특정 언어와 관계없는 순서는 수를 세는 별도의 체계보다는 각 손 모양에서 펼치고 있는 특정 손가락의 수에 근거할 수 있다. 두 가지 이상의 수어가 동일하거나 거의 동일한 기호 형태를 지니고 있다면 이들은 마치 오슬란 사인뱅크상의 동형이의어와 마찬가지로 글로벌 사인뱅크상의 동일한 영상에 태그가 달릴 것이다. 글로벌 사인뱅크는 수어 음운 목록, 도상성의 정도와 유형, 어휘화·문법화 패턴 등에 대한 교차언어·비교연구 목적으로 이용 가능하다.

참고문헌

- Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosh, M., Even, S., Komen, E. and Johnston, T. 2018. Signbank: Software to Support Web Based Dictionaries of Sign Language. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 7–12, 2018, Miyazaki, Japan.
- Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R. and Woll, B.. 2012. From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen and J. Mesch (eds.), Proceedings of

1부 발표

- the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Workshop as part of 8th International Conference on Language Resources and Evaluation, LREC] (pp. 7–12). Istanbul, Turkey., Istanbul, Turkey 21–27 May 2012.
- Crasborn, O & Meijer, A de. 2012. From corpus to lexicon: the creation of ID-glosses for the Corpus NGT. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (pp. 13–18).
- Crasborn, O., Hulsbosch, M. and Sloetjes, H. 2012. Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (pp. 19–22). Istanbul, Turkey., Istanbul, Turkey 21–27 May 2012.
- Johnston, T. 1987. A preliminary signing dictionary of Australian Sign Language (Auslan). Adelaide: TAFE National Centre for Research and Development.
- Johnston, T. 1989. Auslan Dictionary: A dictionary of the sign language of the Australian deaf community. Sydney: Deafness Resources Australia.
- Johnston, T (ed.) 1997. Signs of Australia: A new dictionary of Auslan. Sydney: North Rocks Press.
- Johnston, T. 2001. The lexical database of Auslan (Australian Sign Language). Sign Language & Linguistics 4(1/2). 145–69. DOI: 10.1075/sll.4.12.11joh.
- Johnston, T., Thornton, D. and Napier, J. 2002. Discovering Auslan: A beginner's dictionary of Australian Sign Language. Sydney: North Rocks Press.
- Johnston, T & Schembri, A (eds.) 2003. The Survival Guide to Auslan: a beginner's pocket dictionary of Australian Sign Language. Sydney: North Rocks Press.
- Johnston, T. 2003. Language standardization and signed language dictionaries. Sign Language Studies 3(4). 431–68. 10.1353/sls.2003.0012.
- Johnston, T. 2004. Auslan Signbank (<http://www.auslan.org.au>). Sydney: Royal Institute for Deaf and Blind Children & Catalyst Training Systems.
- Johnston, T. 2008. The Auslan Archive and Corpus. In D. Nathan (ed.). The Endangered Languages Archive—<http://elar.soas.ac.uk/languages>, London: Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London.

- Johnston, T. 2010. From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1). 104–29. DOI: 10.1075/ijcl.15.1.05joh.
- Johnston, T. 2012. Lexical Frequency in Sign Languages. *Journal of Deaf Studies and Deaf Education* 17(2). 163–93.
- Johnston, T. 2014. The reluctant oracle: using strategic annotations to add value to, and extract value from, a signed language corpus. *Corpora* 9(2). 155–89.

Session 1

A Corpus -oriented Sign Language Dictionary

Trevor Johnston,
Macquarie University, Australia

Outline [Slide 2]

In this lecture I describe a corpus-oriented sign language dictionary, specifically the on-line internet dictionary of Australian Sign Language (henceforth Auslan) known as “Auslan Signbank”.

The lecture consists of two parts. In Part 1, I give the history and evolution of the various dictionaries of Auslan that have been published since 1985. In particular I discuss the lexical database of sign forms that was first created in 1984 and which underlies all the dictionaries of Auslan published since 1985 in book, CD-ROM, and internet formats. These were all published before the Auslan Corpus was created. The database is referred to as the Auslan Lexical Database. I also explain the structure of entries in the various dictionary formats.

In Part 2, I describe a corpus-oriented sign language dictionary, using Auslan Signbank as an example. Auslan Signbank became a corpus-oriented dictionary after the creation of the Auslan Archive between 2004-2007 (Johnston, 2008), and the beginning of annotation of that archive into a linguistic corpus, called the Auslan Corpus. The signs recorded in the dictionary are now related to the signs identified in the corpus in two main ways: by the use of ID-glosses and by incorporating into the entries for these signs in the dictionary of information gleaned from the corpus about signers and the signs they use.

Part 1: The pre-corpus sign language dictionary [Slides 3-39]

The Auslan Lexical Database [Slide 4-5]

The lexical database of Auslan (or Auslan Lexical Database) began in 1984 as a series of tabular entries in a Microsoft Word document. Over the 34 years from 1984 to 2018 the database has been constantly improved with

corrections and additions as a result of on-going research.

In all its iterations since 1984, the Auslan Lexical Database has been conceived as a repository of conventional lexical signs of Auslan, which are entered as headsigns. The database does not systematically include entries for the other possible types of signs that signers often produce when communicating with each other. These non-lexical signs have been called depicting (or ‘classifier’ signs), pointing signs, en-actments and gestures. There are variations or modifications of some headsigns that represent inflectional morphology, e.g., changes that mark verbs for agent and patient roles or for aspectual information, or that mark nouns for plurality. Such inflected forms are not entered separately in the dictionary. As with most language dictionaries, these modified forms are left to be discussed in the grammar of the language, rather than in a lexicon. However, in some cases an erstwhile inflectional modification of a sign is actually and irregularly associated with a new and unpredictable meaning, i.e., a new lexical item. Such cases are added to the dictionary.

The research was conducted using several techniques: (i) native signer intuitions and introspections; (ii) videotape recordings of natural signed exchanges between deaf native signers recorded during participant observation sessions in the deaf community were repeatedly viewed and analysed; and (iii) group and individual interviews of native signer informants in which the usage and meaning of individual signs were discussed, tested or elicited.

The Auslan Lexical Database underwent a major overhaul for the on-line publication of Auslan Signbank dictionary in 2004. This dictionary included a feedback mechanism that allowed registered users to and comment and report on missing signs or inaccurate definitions. Soon after, the Auslan Corpus was created. From that point on, the annotation of the corpus was made with reference to the signs already recorded in Auslan Signbank, by way of ID-glosses.

The Auslan Lexical Database over time [Slide 6]

The data in the Auslan Lexical Database has migrated across several different types of platforms since 1984. It has been the dataset upon which the various editions of the Auslan dictionaries were based.

From 1985-1989 the Auslan Lexical Database existed as a list of approximately 2,000 records stored in a Microsoft Word processing program table (the earliest versions of Microsoft Word). In the early 1990s the Microsoft Word tables was migrated to an Apple HyperCard stack (1990-1992). The HyperCard stack made it possible to include visual information easily within each sign record on a computer. Each sign entry/record was illustrated with a line-drawing (the line drawings appeared in the 1989 Auslan Dictionary).

In the early 1990s the data was again migrated to commercially available databases: first to FoxPro, and then to FileMaker Pro. Each record in these new databases had more than 100 fields that encoded a variety of information about each sign, e.g., phonology, semantics, dialect, iconicity, etc. (I will mention some of these again in the second part of this lecture.) These programs allowed for video and graphic information to be entered in some fields, making it possible to have a visual representation of each sign headword within a spreadsheet.

Session 1

A new parallel Auslan Lexical Database was created in the late 1990s by exporting the data from the FileMaker Pro database and incorporating it within into a dedicated computer program that supported the CD-ROM. The program was used to create a multi-media CD-ROM dictionary of Auslan. The programming language that supported the CD-ROM was able to exploit the multiple fields associated with each sign and the numerous cross references that had been established between signs in the database, i.e., between numerous records and fields, such as relationships of synonymy, dialectal variations, restricted usage, and so forth.

In the early 2000s the entire FileMaker Pro dataset, including video fields, was migrated into an internet website which had the same basic functionality as the CD-ROM. The site used Django web application framework written in Python programming language. The site was called Auslan Signbank (2004).

From 2014 the scripting and programming that supports Auslan Signbank and has become open source and house on a Github site (<https://github.com/signbank>). Today (2018), Auslan had at last count 7,298 entries (approximately half of these are core stand-alone headsigns, and half are common variant forms of those core headsigns).

Dictionary formats: print, CD-ROM and Internet [Slides 7-14]

The dictionaries published using the Auslan Lexical Database have appeared in three formats:

One, print format: 1987 A preliminary signing dictionary of Australian Sign Language (Auslan); 1989 Auslan Dictionary: A dictionary of the sign language of the Australian deaf community; 1997 Signs of Australia: A new dictionary of Auslan; and 2003 The Survival Guide to Auslan: a beginner's pocket dictionary of Australian Sign Language.

Two, CD-ROM multi-media format: 1997 Signs of Australia: A new dictionary of Auslan; and 2002 Discovering Auslan: A beginner's dictionary of Australian Sign Language.

Three, in a web-based internet format; 2004, 2008, 2018 Auslan Signbank (<http://www.auslan.org.au>).

It is important to note that despite the various editions and different formats, the organizing principles of all these dictionaries have remained unchanged. The changes that have happened over the years relate to (i) minor alterations in sign sequencing according to formational features, (ii) improvements to the functionality of the database in terms searching and sorting entries, and (iii) adaptations and developments occasioned by the transfer from old to new media, i.e., from printed book format to CD-ROMs through to the web-based application.

In all its manifestations, the Auslan Lexical Database attempts accurately document the lexicon of the sign language of the deaf community in Australia.

Dictionary Structure [Slides 15-17]

The Auslan dictionaries have all been limited bidirectional, bilingual dictionaries. There has been one exception —The Survival Guide to Auslan: a beginner's pocket dictionary of Australian Sign Language (Johnston

& Schembri 2003) (see below).

In a true dictionary the words or signs of a language are defined, not just listed. In a bilingual dictionary, the words or signs of Language A are defined in Language B. In a bidirectional dictionary one has a list of Language A words or signs defined in Language B, and a list of Language B words or signs defined in Language A. Thus in a bidirectional, bilingual dictionary of a sign language there should be two ways to access lexical information: one through the sign language and one through the spoken language. Each word or sign of each language is defined in the words or signs of the other language.

The main part of the Auslan dictionaries lists Auslan signs (Language A) which are defined in English (Language B). The headwords of Auslan, i.e., the headsigns, are ordered according to language-internal formational principles such as handshape and location. In the print dictionaries, the headsign was represented by a line illustration. In the database until the late 1990s, the headsigns were also represented by a dedicated sign language transcription in HamNoSys, but this was not viewable by the public, nor included in the printed dictionary.

Transcriptions in HamNoSys are no longer regarded as essential in order to know what the form of a sign is because one can easily see the form of a sign by looking at the field that contains the video clip (or, in earlier editions, line drawing). Nor are transcriptions required in order to sort or describe sign entries in terms of phonological features. (Previously, it was possible to sort the HamNoSys transcriptions themselves by assigning an order to the Unicode symbols used for handshape, location, and so on.) Moreover, the new databases can easily accommodate large numbers of fields in which phonological and other features of signs can be tagged, without recourse to a special transcription script. The HamNoSys transcriptions were created up to the late 1990s, before digital video technology and desktop computing power became powerful and cheap enough to make it feasible to represent each headsign with its own dedicated videoclip. They have been discontinued. All entries in the Auslan Lexical Database, therefore, make do with a video clip, and have associated fields that code for phonological form.

In the print, CD-ROM and web-site formats of the dictionary there is an Auslan to English section or view of the data. However, in all three the English to Auslan view (Language B to Language A) is limited. In the print format, the reason is simple: a definition of an English headword in Auslan using line drawings or HamNoSys was virtually impossible. Thus, this direction of comparison was represented simply by an English word index which pointed users to the entry numbers of relevant Auslan translation equivalents of each word. In the CD-ROM and Signbank web-site formats, one can tag of each sign entry (sign form) for its phonological characteristics (handshape, location, movement, etc.) and for the English key-words associated with each sign. The user typed an English word in the search field of the CD-ROM or the Signbank web-site to find signs associated with those keywords. Thus it is still bidirectional, but only in this limited sense.

The dictionaries sorted and arranged the headsigns according to principles based on formational features (see below). The dictionary groups the signs into handshape sections on the basis of the handshape of the strong hand at the beginning of the sign. Within each section, other formational features, such as location, orientation,

Session 1

movement, the number of hands, and, in some cases, facial expression. The hand-shape on the weak hand, and the symmetry of the sign determine the grouping and sequencing of signs with the same dominant handshape.

Ordering principles for print dictionary ('alphabetical' order) [Slides 18-25]

Only the main principles that underlie the 'alphabetic order' of sign entries are discussed here. For full details, see Johnston (2003).

By handshape [Slides 18-20] Handshape is ideally suited to act as the principle for distinguishing one sign from another and for grouping signs into sections. The handshapes of Auslan are ordered according to a semi-arbitrary sequence. The earlier editions of the Auslan dictionaries (1987, 1989) based the sequence of handshapes on the frequency of occurrence of each in the lexicon, with the most frequent handshape—the flat handshape—appearing first. Signs of Australia (1997) used a more natural and intuitive sequence based on Auslan handshapes used in counting. The order follows the handshape used in counting starting from zero to then. Slight variations on these handshapes are ordered between the major counting handshapes. For example, the Round hand is first in the dictionary because it is the Auslan handshape for the number 0. The Okay hand is second because it is an alternative handshape for 0. The Point hand follows because it is the handshape for the number 1. It is followed by the Hook hand because the latter is similar to the Point hand (it resembles a hooked 1) and so on through handshapes for the numbers 2 through 12 (In Auslan, none of the numbers above twelve involves the use of a handshape not already found in lower numbers.) A few handshapes found in Auslan cannot be associated with numbers in any way and are arbitrarily placed at the end of the list and at the end of the dictionary. For example, SUN is found in the Round hand section of the dictionary, IGNORE is found in the Point hand section, and FRONT is found in the Flat hand section [Slide 20].

By primary location [Slides 21-22] The dictionary then sequences the signs according to location (tabulation). It lists signs made high in the signing space or high on the body (e.g., on the head or face) before those that are made low in the signing space or low on the body (e.g., on the waist or arm). Thus THINK comes before WHO which comes before STOMACH.

By number of hands [Slide 23] Within each handshape section, the dictionary lists as a group the signs that use only one hand before those that use two hands. For example, the one-handed sign CONSIDER precedes its double-handed variant.

By same or different handshapes [Slides 24-25] Within each handshape section, the dictionary arranges signs that require the use of two hands after those that require only one hand. Within these two-handed signs those that have the same handshape on both hands are listed before those that have a different handshape on the subordinate or weak hand. Thus CONSIDER precedes CROSS-OUT. The order of two-handed signs with a different handshapes on the weak hand simply repeats the order of handshapes in the dictionary overall. Thus the dominant or strong hand determines the handshape section of the dictionary the sign appears in, and the subordinate or weak handshape determines its ordering with respect to other two handed signs with the same

dominant handshape. Thus QUESTION, POTATO and DEBT appear in this order relative to each other because the Okay hand comes before the Point hand and the Point hand comes before the Flat hand.

The structure of the survival guide [Slide 26]

The Survival Guide to Auslan: a beginner's pocket dictionary of Australian Sign Language (Johnston & Schembri 2003) is a wordlist rather than a true dictionary. As such, it was not unlike most so-called sign language dictionaries that have been created in the past by non-linguists.. These were essentially mono-lingual wordlists, in which the headwords were actually the words of the community spoken language arranged according to the alphabetic writing conventions of that language. An illustration of a sign equated with the headword was placed next to it. (The Survival Guide was published with a specific narrow audience in mind: adults with limited time, who only wanted to know very simple vocabulary items to facilitate basic communication with deaf people.) There is a specific niche for these types of simple wordlists but they should not be confused with true dictionaries of sign languages which present day linguists are creating and which are the topic of discussion here.

Structure of CD-ROM [Slides 27-33]

The CD-ROM edition of Signs of Australia (1997) was the first multi-media dictionary of Auslan. The Auslan Lexical database was again the source of data used in creating the CD-ROM.

The most important innovation in the CD-ROM was the ability to search for signs based on their formational features. A sign search request could be launched selecting the 'search by Auslan' button option rather than the familiar 'search by English word' (by consulting the English index in the book format). After selecting the 'search by Auslan' option on the CD-ROM the user was guided through a series of filtering questions which asked them to specify: (i) if the sign was one-, double- or two-handed; (ii) the handshape on the strong hand at the beginning of the sign; (iii) the location of the sign; and, if the sign was used two hands, (iv) the handshape on the weak hand, and (v) the secondary location (i.e., the place on the weak hand where the two hands contacted or interacted with each other).

The second most important innovation was the ability to instantly select and sort signs according to various criteria, such as semantic field (food, body parts, emotions, health, etc.) or dialect (according to major dialect zones or states of Australia).

The third most important innovation was the ability to jump instantly between signs within the dictionary based on a number of pre-established links between them, such as synonyms, antonyms and homonyms.

Structure of Internet Signbank [Slides 34-39]

The Auslan Lexical Database also lies at the basis of the internet version of the Auslan Dictionary, called

Session 1

Signbank generally, or Auslan Signbank specifically. Indeed, the database itself now resides in the web-site. In other words, Auslan Signbank is just the website name of the Auslan Lexical Database in its current form: Auslan Signbank is just web portal on the database. It is however possible for researchers to download the contents of the Signbank site as a .csv file and then import the records into a variety of databases for further processing or investigation. (Versions of the database thus still exist in FileMaker Pro and other spreadsheet formats.)

As with all previous versions of the dictionary Auslan Signbank contains conventional lexical signs of the language displayed as the headsign with a definition in English. The dictionary also contains a number of depicting or classifier-based signs the lexical status of which is open to question.

There are some innovations in the presentation of the dictionary in Auslan Signbank. First, there is now a public view available to anyone with an internet connection. Second, there is a restricted or private view available only to registered users. The registered view has several levels of access to the database. Ordinary users register in order to be able to post feedback on the site (including reporting errors or missing signs from the dictionary). There are four other levels of research access available on application, interpreter, researcher, editor, publisher. These have various access and editing rights within the database. Access to viewing and/or changing the database is controlled allocating different classes of users with different levels of rights and privileges to the site. These additional page views display the type of information already exploited in the CD-ROM. However, unlike the CD-ROM, Auslan Signbank can be up-dated at any time (provided one has the access privileges) and these changes can be effected and displayed immediately. Records may also be deleted or added as new information becomes available.

This means that the dictionary is able to respond quickly to new information that comes available either through community feedback or through examination of the Auslan Corpus (see next section).

The initial web version of Auslan Signbank (2004) had only some of the functionality found in the CD-ROM due to constraints imposed by the scripting language used at that time. Between 2004 and 2018, the scripting language behind Auslan Signbank was completely overhauled and functionality previously found only on the CD-ROM became available on the site. However, to date, most of this functionality is only available to registered users who have special access rights.

The signs still have the inherent order first assigned in the print versions of the dictionary, but the ordering of signs is not transparent on a web-site as it would be in a printed book. The order is only evident if one scrolls through the dictionary from one page to the next, as each page only presents one sign entry.

The latest major innovation in Auslan Signbank, implemented in 2018, is the provision of an “Definition in Auslan” field. In this field, a video clip can be inserted in which a definition in Auslan is given for the headsign. The creation of entries of this type has only recently begun, but it does mean that it will be possible, in the future, to present the Auslan Signbank entries as in a monolingual dictionary. This is ground-breaking. No monolingual dictionary of any signed language has yet been published. Without the existence or development of writing systems for signed languages, monolingual dictionaries of signed languages in printed form were virtually

impossible. Although monolingual sign language dictionaries became possible in principle with the advent of film and video, they were still totally impracticable. However, with the availability of digitized video, it has become technically possible to make a monolingual dictionary, but none have yet to be made or published. This latest version of Auslan Signbank (2018) now makes it possible to create monolingual entries—entries in which an Auslan headsign is defined using Auslan. The first entries using this technique have begun to be created.

Part 2: corpus-oriented sign language dictionary [Slides 40-70]

Rational for a corpus-based sign language dictionary [Slides 40-44]

Recall that the Auslan dictionaries described above were the product of a ‘traditional’ language documentation and description research program which involved participant observation, interviews, and elicitation. Such a model is widely used in linguistics and anthropology for work on undocumented and undescribed languages. Generally speaking, the language communities in which this type of research is undertaken tend to be found in geographically discrete, small-scale traditional societies without a history of writing.

Sign language communities, however, tend not to be like this. Almost everywhere, community sign languages are used by small numbers of deaf individuals who are embedded in much larger hearing communities which use a spoken (and sometimes written) language. In some countries or regions, some deaf people are born into families that have several generations of deafness, but this is relatively uncommon. Most deaf people have hearing parents and siblings. Thus the acquisition and use of most sign languages appears quite atypical when compared to that of most spoken languages. These and other well-known sociolinguistic characteristics of deaf communities mean that many researchers have remarked that sign language communities seem to display a much higher degree of variation in both lexis and grammar than might otherwise be expected.

In the late 1990s and early 2000s, it became obvious to many sign language researchers, including Auslan researchers, that an empirically sound description of the lexicon and grammar of a sign language requires the creation of a linguistic corpus based on naturalistic signing from native or near-native signers (Johnston 2010, 2014).

The creation of a linguistic corpus entails the transformation of recordings of language use through transcription, annotation, and translation so that it becomes a machine readable resource. In the case of sign languages this, of course, means creating digital video recordings, and using multimedia annotation software.

Once an archive of recordings has been created in preparation for creating a corpus of a sign language, any pre-existing sign language dictionary of this sign language needs to orient itself towards it. For example, one needs to ensure that every conventional lexical sign that is identified and thus glossed in the corpus is also listed in the dictionary, otherwise the dictionary would clearly be judged as incomplete. Actually, the relationship between the dictionary and the corpus is two way: on the one hand, the dictionary informs sign identification and annotation;

Session 1

on the other hand, the corpus provides potentially new data on hitherto undocumented conventional signs, or identifies possible variants of already documented forms.

Relating corpus annotations to sign entries in the dictionary [Slides 45-49]

ID-glosses

In order to harmonize the dictionary and the corpus, it is necessary to be able to relate the form of sign types entered in the dictionary with tokens found in the corpus. In this corpus-oriented database, sign entries are associated with keywords and a unique identifying gloss in English. The unique identifying gloss (which I have called the ID-gloss) is assigned to each headsign in the dictionary. The ID-gloss associated with a sign type in the dictionary is used to annotate any token of this sign type in the Corpus. However, in order to use ID-glosses effectively and efficiently, one needs a reference database in which lexical types are documented. This is the essence of a corpus-oriented sign language dictionary: all conventional lexical signs of the language are listed in the sign dictionary and have a unique identifier when segmented and labelled in the corpus annotations.

Thus during the annotation and glossing of the corpus, constant reference needs to be made to the sign language dictionary. To do this the Auslan annotator refers to Auslan Signbank. Whenever the ID-gloss for a particular sign needs to be checked, the annotator types in a English keyword associated with the sign into Auslan Signbank search window and views the matches suggested in Signbank. The annotator then select the ID-gloss for the sign that most closely resembles the form of the sign they observe in the corpus.

Auslan Signbank is currently external to the ELAN annotation software used to create the corpus. The annotator simply has the on-line dictionary open on their desktop as they create annotations in ELAN. However, in Signbank NGT (a Signbank for Sign Language of the Netherlands based on Auslan Signbak – we will discuss this later) the two are linked (Crasborn & Meijer 2012, Crasborn et al 2012). For exam-ple, Signbank NGT is linked to the ELAN software templates that are used to annotate the NGT corpus. Annotators can begin to type the ID-gloss of a sign annotation in ELAN and a controlled vocabulary listing all related ID-glosses in Signbank NGT will appear so they may choose the best fit. If in doubt, the annotator may click on any gloss within the controlled vocabulary and a thumbnail video clip for that ID-gloss from Signbank will appear in a pop-up window.

Phonological variation [Slides 50-51]

The most common form of a sign (judged by the most frequent occurrence of a number of variant in a token count from the corpus) is treated as the stem or base form of a lexical sign. It is the main headsign. Variants of the headsign are shown at that entry. A variant is basically a form of base sign that has one, or at most two, formational features which are different to the base sign (the variant usually involves a different handshape,

location, or the number of hands used in the sign). Variants also have their own entry record at the appropriate place in the form-based dictionary. If usage data, as recorded in the corpus, suggests that the status of base and variant forms need to be reversed, they will be. For example, usage data may show that a double-handed form of a sign is far more frequent than a one-handed form. If the one-handed form has been listed as the main headsign in the dictionary, the relations should be reversed.

Attested usage [Slide 52]

Overtime, and providing that the corpus is representative of all users and all text-types, and also large enough, one should expect to find that the majority of one's sign language dictionary entries are attested in the corpus. The only signs expected not to appear in a general corpus are those already known to be archaic, obsolete or marginal (e.g., vocabulary only used by small groups of people in specialist fields). Attested usage in the corpus indicates that the dictionary is empirically sound. In 2012 after 4 years of annotation (see Johnston 2012) approximately 75% of the 3,500 core entries in Auslan Signbank had been attested in the Auslan Corpus. This is a very high percentage of the known lexicon given the relatively modest size of the Auslan Corpus at that time (approximately 70,000 tokens), when compared to spoken language corpora. (There is double this number of entries in Auslan Signbank, i.e., over 7,000 entries, but these are all variant forms of the core 3,500 lexical signs.)

Missing data [Slide 53]

Of course, no dictionary is ever complete so it is more than likely that hitherto unrecorded signs will be identified in the corpus. If these are confirmed as genuine conventional lexical signs of the language, then, after a period of consultation with community members and/or confirmation by additional tokens being found in the corpus, they should be added to the dictionary. In this way the corpus is able to augment the lexical database with hitherto unrecorded signs, either variants of existing lexical entries, or lexical signs which may be recent additions to the language (neologisms or borrowings) or which were simply previously overlooked signs. Approximately 50 such new or previously unrecorded conventional lexical signs had been identified in the Auslan Corpus and so they were, in turn, added to Auslan Signbank, after 4 years of annotation.

Improved definitions based on usage data [Slide 54]

A sign form may be initially assigned a meaning which is too narrow given its attested usage in the corpus. In such cases, the definition in Signbank is improved to reflect the way the sign is actually used in naturalistic signing. This could mean changing and improving the definitions, or adding, removing or changing associated keywords.

Session 1

Adjust the lexical status of some signs [Slide 55]

Some highly iconic or depicting signs may have been entered in the dictionary as conventional lexical signs on the basis of initial fieldwork (intuitions, introspection, elicitation or community feedback had reported the sign as conventional and lexicalized). However, the use of the sign as attested in the corpus may indicate that the sign does not have a meaning which is as specific or conventional (or 'lexical') as first thought. Such signs are borderline signs in terms of lexicalization: they may not be language-specific in their meanings and thus not really appropriately listed in a dictionary of Auslan. (Of course, they would still need to be discussed in a grammar of the language.) Such entries may therefore need to be updated to better reflect their status, or even removed.

From corpus to dictionary? [Slides 56-57]

In the ideal corpus-building situation, it is not expected that one would begin to segment and gloss a recording in a sign language without first having conducted basic lexicological research into the language, and then having recorded and described the resulting initial lexicon in a database or dictionary. However, in recent years some sign language dictionary projects have begun as part of large sign language corpus projects. The aim is to use the corpus to build the dictionary.

It is possible to construct a reference lexicon more or less from scratch based on the annotations themselves as they are being made, i.e., using only attested instances in the corpus itself as the basis for one's lexicon. Something akin to this approach has been described for the BSL corpus (Cormier et al., 2012). Though this may provide a way to proceed, this approach is difficult, time consuming and prone to inconsistency because annotators have no common reference point, especially in the early stages of annotation and glossing. Researchers are likely to encounter many different uses and variant forms that require constant revision and correction of the incipient database and dictionary. This can also have knock on effects in the growing lexical database (Crasborn and Meijer, 2012).

Indeed, some researchers believe that an empirical dictionary of a sign language should only draw on signs attested in the reference corpus. One could perhaps call these corpus-based or corpus-driven sign language dictionaries rather corpus-oriented sign language dictionaries, like the one described here. However, there are dangers in adopting such a strict criteria for inclusion of signs in a sign language dictionary—it is highly unlikely that any corpus of a language (unless it runs into hundreds of millions of words or signs) will include at least one token of every known word or sign in the language. This is particularly problematic if very basic everyday vocabulary items just happen not to be found in a corpus, something that is highly likely to be the case in the sign language corpora that have been or are in the process of being created because they are really not that large. Any user of such a dictionary may find that signs for basic concepts is not recorded because it simply never occurred in the corpus. This is not satisfactory and users of such a dictionary would rightly conclude that it is incomplete. One way to avoid this problem, if one wants to restrict oneself to attested data, is to construct one's

corpus so that it includes elicited text and vocabulary items, not just conversational or narrative text. One could then incorporate tasks during recording sessions that elicit basic vocabulary items.

Enriching dictionary entries from corpus findings [Slides 58-68]

The Auslan Lexical Database, and hence Auslan Signbank, is designed to accommodate a large variety of information about each sign at each entry. This information is found in scores of dedicated fields or cells. These contain information on: (i) the form of the headsign; (ii) its meanings; (iii) who uses it; and (iv) criteria by which the entry can be searched, sorted or displayed on-line. Much of this information emerges from the study of the corpus itself (e.g., dialect preferences, age cohorts, school-based signs, religious affiliation, frequency, etc.) This information is constantly under review and is easily editable in the Signbank researcher interface. Some information is visible to the public immediately, other information is only visible on the private view of the database, where it is available for researchers to consult.

Dealing with homophones [Slide 69-70]

Because Auslan Signbank is a form-based database of signs, it follows that if two signs are homonyms, i.e., two distinct meanings in Auslan have the same sign form (cf. English ‘can’ meaning ‘metal container’ and ‘able to’), then they would both reference the same form. The solution in Auslan Signbank is the same as in an English dictionary—the headword or headsign is repeated, and at the two (or more) entries created the same phonological form is defined with each of its two (or more) distinct meanings. In Auslan Signbank these are numbered: Sense 1, Sense 2, and so on.

Multiple sign languages [Slides 71-78]

Auslan Signbank & other Signbanks [Slide 72]

From their first conception and application in the Auslan Lexical database, the principles underlying its structure were claimed to be applicable to any sign language (Johnston 2001). Auslan Signbank (and its earlier instantiations in book and CD-ROM formats) has form-based entries. These entries have their formational features encoded either in a dedicated transcription system (such as HamNoSys) or in a series of fields that specify the presence or absence of relevant formational features of each sign (such as handshape, location, etc.). The result is that signs from any sign language can be accommodated within the database—entries are simply created which are described by form. (If the database is ordered by form, then the new sign is simply entered at its appropriate slot according to ‘alphabetical’ order.) Even the types of interrelationships between signs (such as synonymy, antonymy, homonymy, semantic field membership, dialect distribution, etc.) that are built into the

Session 1

program can also be exploited, as appropriate, for any language.

Realization of this potential has lead other linguists working on other sign languages to adapt the basic architecture of Auslan Signbank for their own sign languages. As a result, on-line dictionaries have been created for Sign Language of the Netherlands (NGT Signbank), British Sign Language (BSL Signbank), and Finnish Sign Language (FinSL Signbank), and work on an American Sign Language (ASL Signbank) has commenced.

From homophones to multiple sign languages **[Slide 73-77]**

One important aspect of the structure of Auslan Signbank opens up the further possibility of creating a lexical database that accommodates multiple sign languages in the one Signbank, i.e., in a Universal Lex-ical Database as it was described in Johnston (2001). Such a database does not simply use the existing Auslan Signbank as a template for other Signbanks, it creates a Universal Signbank.

Recall, that Auslan Signbank creates records of sign forms according to formational features. This re-sults in a visual ordered sequence of line drawings or photographs when presented in print format, or in columns of similar video clips when presented as records and cells in a spreadsheet. Though only one entry per page is visible in an on-line dictionary such as Signbank, the same principle allows instant sort-ing of records within the database by any one of the relevant formational features encoded with each entry. Thus signs may be ordered by handshape (as in the print format), or by location, or by any other feature requested.

More importantly, as we have seen, these form-based entries also easily encoded and represent phono-logical variation, dialect variation, semantic relations (antonyms, synonyms, homonyms), and relations to other sign languages, such as BSL and ASL. In Auslan Signbank sign forms can now be tagged not only dialect, but also sign language. The tag means the sign form exists in BSL or ASL.

The ability to tag for language together with the ability to tag a sign as being in a homonym relation with another creates the potential for an open-ended database. Signs from any sign language can be en-tered into the same database. When the Auslan Lexical Database was first conceive there were practical limitations on personal computer processing speeds (and cost), the quality of digital video recordings, and the storage capacity (and cost) of servers. Today, the number of entries in Signbank, the number and type of codes associated with each entry, or the length of the video clips used as the headsign or as definitions are no longer impracticable nor too costly.

As a result, work to test the possibilities and uses of this type of universal Signbank, has begun. It is called Global Signbank.

Global Signbank? [Slides 78]

Is a Global Signbank practical? Two observations about the lexicons of sign languages that have been studied in depth to date suggests it may well be.

First, size. Most dictionaries of sign languages that have been published over the past 50 years do not seem to contain more than several thousand main headsigns (somewhere between 3,000 to perhaps 8,000). This appears relatively modest compared to most spoken language dictionaries, even spoken languages without a written form.

Second, many researchers have remarked that the high degree of iconicity in sign languages has tended to create a situation in which the same or very similar sign forms reappear repeatedly in different sign languages. Though these iconically motivated sign forms often have the same or similar conventional meaning in unrelated sign languages, their meanings may also be completely different. Nonetheless they share the same sign form. It is not uncommon for signers of one sign language reporting that when they view a foreign sign language they often see signs that appear to be very similar in form to ones in their own sign language. Sometimes they may be able to correctly guess the meaning, often they will not be able to. By contrast, it is quite rare to see the same word form appear repeatedly in unrelated spoken languages. The reason for this is that the potential number of phonologically well-formed words in most spoken languages appears to all practical purposes to be without limit when the combination of phonemes and the possible number of syllables are taken into account. However, the situation in signed languages may be different partly due to the impact of iconicity.

Assuming it is practicable, a multiple language version of Signbank—Global Signbank—would most likely benefit from the current formational order being re-designed. Global Signbank would preferably have ‘neutral’ (i.e., non-language specific) formational order accepted by a majority sign language re-researchers, even if the ‘primary’ feature is most likely to remain handshape. A non-language specific order could, for example, be one based on the number of selected fingers extended in each handshape, rather than any given counting system as such. Whenever two or more sign languages have an identical or near identical sign form they would be tagged to the same video in Global Signbank, just as are homophones in Auslan Signbank. Global Signbank could be used for the cross-linguistic and comparative study of the sign language phonological inventories, degrees and types of iconicity, and lexicalization and grammaticalization patterns.

References

- Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosh, M., Even, S., Komen, E. and Johnston, T. 2018. Signbank: Software to Support Web Based Dictionaries of Sign Language. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds.). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 7-12, 2018, Miyazaki, Japan.
- Cormier, K., Fenlon, J., Johnston, T., Rentelis, R., Schembri, A., Rowley, K., Adam, R. and Woll, B.. 2012. From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen and J. Mesch (eds.). Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Workshop as part of 8th International Conference on Language Resources and Evaluation, LREC] (pp. 7-12). Istanbul, Turkey., Istanbul, Turkey 21-27 May 2012.

Session 1

- Crasborn, O & Meijer, A de. 2012. From corpus to lexicon: the creation of ID-glosses for the Corpus NGT. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (pp. 13-18).
- Crasborn, O., Hulsbosch, M. and Sloetjes, H. 2012. Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (pp. 19-22). Istanbul, Turkey., Istanbul, Turkey 21-27 May 2012.
- Johnston, T. 1987. A preliminary signing dictionary of Australian Sign Language (Auslan). Adelaide: TAFE National Centre for Research and Development.
- Johnston, T. 1989. Auslan Dictionary: A dictionary of the sign language of the Australian deaf community. Sydney: Deafness Resources Australia.
- Johnston, T (ed.) 1997. Signs of Australia: A new dictionary of Auslan. Sydney: North Rocks Press.
- Johnston, T. 2001. The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics* 4(1/2). 145-69. DOI: 10.1075/sll.4.12.11joh.
- Johnston, T., Thornton, D. and Napier, J. 2002. Discovering Auslan: A beginner's dictionary of Australian Sign Language. Sydney: North Rocks Press.
- Johnston, T & Schembri, A (eds.) 2003. The Survival Guide to Auslan: a beginner's pocket dictionary of Australian Sign Language. Sydney: North Rocks Press.
- Johnston, T. 2003. Language standardization and signed language dictionaries. *Sign Language Studies* 3(4). 431-68. 10.1353/sls.2003.0012.
- Johnston, T. 2004. Auslan Signbank (<http://www.auslan.org.au>). Sydney: Royal Institute for Deaf and Blind Children & Catalyst Training Systems.
- Johnston, T. 2008. The Auslan Archive and Corpus. In D. Nathan (ed.). *The Endangered Languages Archive—<http://elar.soas.ac.uk/languages>*, London: Hans Rausing Endangered Languages Documentation Project, School of Oriental and African Studies, University of London.
- Johnston, T. 2010. From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15(1). 104-29. DOI: 10.1075/ijcl.15.1.05joh.
- Johnston, T. 2012. Lexical Frequency in Sign Languages. *Journal of Deaf Studies and Deaf Education* 17(2). 163-93.
- Johnston, T. 2014. The reluctant oracle: using strategic annotations to add value to, and extract value from, a signed language corpus. *Corpora* 9(2). 155–89.

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전

Records of Visible Language:
Sign Language Dictionary

주제 2 | 수어사전의 구조

예테 크리스토퍼센
덴마크, 유시시(UCC) 대학교

Jette H. Kristoffersen
University College Capital, Denmark

수어사전의 구조

– www.tegnsprog.dk

예테 크리스토포르센
덴마크, 유시시(UCC) 대학교

초록

수어사전학은 수어연구계에서 신생 학문에 해당된다. 수어사전학자가 직면하는 가장 큰 어려움은 수어가 시각 양식(visual modality)을 사용하는 언어로 수어사용자 간 널리 사용되는 문어 체계가 없다는 점이다. 지난 수십 년간 수어사전학은 전자매체의 혜택을 크게 누렸다. 수어사전은 더 이상 과거와 같이 어휘만을 정리한 목록이 아닌 구어사전과 유사한 정보를 담은 과학에 기반한 사전이다. 하지만 수어의 시각적 요소는 어휘 검색과 수어(sign)에 관한 정보 수록 측면에서 여전히 과제로 남아있다. 본 논문에서는 덴마크 수어사전(Danish Sign Language Dictionary, Kristoffersen & Troelsgård, 2016)의 구조를 소개하고 수어사전의 구조를 제안하고자 한다.

1. 서론

가장 최근까지도 대부분의 수어사전은 구어에서 해당되는 어휘와 수어를 나타내는 그림, 사진 또는 설명을 수록한 어휘 목록에 불과했다. 이때 원천 언어는 주로 구어에 해당되며 수어가 목표 언어이다.

지난 수십 년간 수어사전학은 빠르게 발전하였으며 일련의 과학적 수어사전이 편찬되었다. 그 첫 시도 중 하나로 언어학적 원칙에 기반한 미국수어사전(A Dictionary of American Sign Language on Linguistic Principles, Stokoe et. al. 1965)이 있다. 오늘날 수어사전은 더 이상 어휘의 나열이 아닌 구어사전과 유사한 정보를 담은 진정한 의미의 사전이다. 또한 수어가 원천 언어에 해당된다.

수어사전학자는 기술발전의 혜택을 크게 누렸다. 과거에는 수어를 사전에 표기하는 방식을 결정하는데 어려움을 겪었다면, 오늘날에는 수어를 디지털 동영상 형태로 기록하여 쉽게 전자사전에 통합시킬 수 있다

(Zwitserlood et al, 2013 참조).

하지만 여전히 수어사전학자가 겪는 몇 가지 어려움이 있다. 예를 들어, 수어를 지칭하거나 기호 검색을 수행하고 검색한 결과, 기호의 순서 정렬 등을 위해서는 수어에 대한 어떠한 공식 표기법이 필요하다. 또한 수지적 요소(manual features) 검색이 가능하도록 하기 위해서 수어를 원천 언어로 하는 수어사전은 구어 사전에 비해 고도화된 검색 기능이 필요하다.

덴마크 수어사전'(Kristoffersen & Troelsgård, 2016)은 사실상 덴마크 수어(DTS)-덴마크어 이중언어 양방향 사전이지만, 그 구성은 이중언어 사전과 단일어 사전의 중간 정도에 해당된다(Kristoffersen & Troelsgård, 2012). 표제어는 DTS 기호만을 기준으로 하며 수어의 의미에 따라 정렬되어 있고, 해당되는 덴마크어의 동형 이의어 또는 다의어를 구별하기 위한 설명을 제외한, 용례를 포함한 거의 모든 정보는 DTS를 기준으로 작성되어 있다. 덴마크 어휘에 대한 정보는 기존의 '덴마크 단일어 범용 사전'으로 쉽게 확인할 수 있기에 본 사전에서는 DTS를 중심으로 기술하도록 이 접근법을 채택했다. 따라서 덴마크 수어사전은 표제어의 정의 대신 해당 사전의 또 다른 메타언어인, 검색 가능한 대응 덴마크어를 수록한 단일어 사전이라 할 수 있다.

2. 덴마크 수어사전 표제어 배열

덴마크 수어사전의 표제어는 4개 부분으로 구성되어 있다(그림 1 참조).

섹션 1: 표제어항

이 섹션은 수어 표제어의 사진과 용어 해설(gloss)을 담고 있다. 첫 번째 수위(location)와 제일 두드러지는 수형(handshape)을 아이콘으로 설명한다. 마지막으로 버튼을 클릭하여 해당 표제어 해설을 출력할 수 있도록 별도 창을 열거나 전체 사전에 수록된 기호를 포함한 모든 용례를 담은 용어 색인을 확인할 수 있다.

섹션 2: 동영상 창

수어 표제어의 기본형을 볼 수 있다. 단어의 음운적 변이형(variant)이 있는 경우, 동영상 상단에 각 변이형별 재생 버튼으로 확인할 수 있다. 컨트롤 버튼으로 재생속도를 낮추거나 창을 확대할 수 있다. 용례 등의 동영상은 다른 섹션에 있는 재생 버튼을 클릭하여 확인할 수 있다.

1부 발표

섹션 3: 의미 창

섹션3에서는 수어 단어의 뜻을 설명한다. 대응 덴마크어 혹은 단어의 용법에 관한 설명(기능어의 경우), 입 모양에 대한 정보, 동의어에 관한 상호 참조, 사용 제한 설명 및 예문 등이 기술되어 있다. 의미상 이해하기 어려운 합성어도 일반적인 뜻 아래 기술하고 있다. 대응 덴마크어는 덴마크어 사전(Danish Dictionary)¹⁾에서 검색할 수 있으며 연결 링크가 제공된다. 따라서 사용자들은 해당 덴마크어의 뜻과 용법에 대해 더 자세히 알아볼 수 있다.

섹션 4: 추가 정보

동일한 것으로 보이는 수어(동형이의어)와 분류사의 경우에는 자주 사용되는 고정 형태(frozen form)에 관한 상호 참조를 제시한다. 또한 표제어와 흔히 함께 사용되는 수어도 기록하고 있다.

The screenshot displays a digital dictionary interface for Danish Sign Language. It is organized into four numbered sections. Section 1, titled '1. dele, del, detalje, fordele, fordeling, andel', includes a video of a woman signing, a list of related terms like 'HALV, ADSKILLE, DIVIDERE', and example sentences. Section 2, '2. afdeling, afsnit, filial', also features a signing video, related terms like 'HOSPITAL EKSISTENS MASSER-AF DEL PROFORM', and example sentences. Section 3, '3. gå op i detaljer', shows a signing video, related terms like '[I], <intet>', and example sentences. Section 4, '4. Samme form som: SKARP, SVIGERINDE' and 'Almindelige tegnforbindelser: BY~1 DEL bydel, EN DEL halvanden', provides information on related forms and common sign combinations. Each section includes a video player with a play button and a progress bar.

[그림 1. 덴마크 수어사전 표제어의 4개 섹션 구성 예시]

1) 덴마크어 사전(Danish Dictionary)은 현대 덴마크어를 기록한 단일어, 범용 사전이다. 사전은 www.ordnet.dk/ddo에서 무료로 사용할 수 있다.

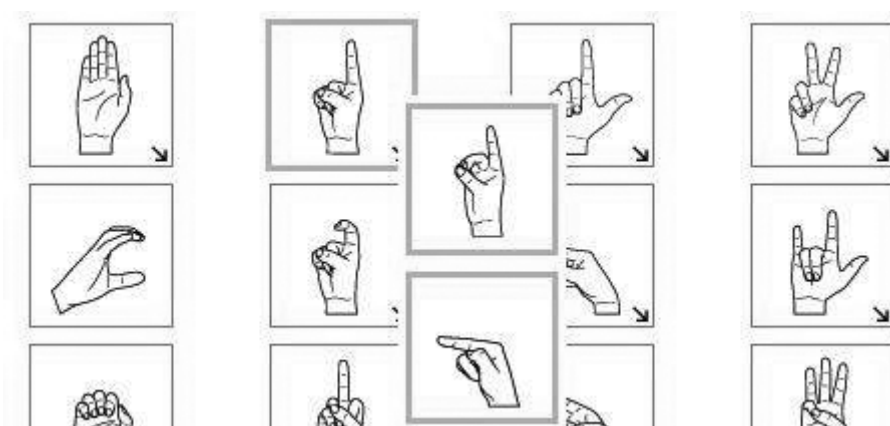
3. 덴마크 수어사전 검색 기능

3.1. 검색 기준 선택

덴마크 수어사전에 등재된 수어는 여러 카테고리에 따라 검색할 수 있다. 수형, 수위, 주제 및 텍스트로 검색이 가능하다.

수형 검색

검색이 가능한 수형은 총 65가지이다. 검색 기준으로 사용될 수형은 연관 수형 그룹별로 선택되지만 원치 않는 수형은 개별적으로 제외할 수 있다. 이 방식으로 사용자는 기본적으로 묶이는 수형 그룹과 별개로 단일 수형 또는 원하는 유형의 수형을 선택할 수 있다. 그림 2는 집게손가락을 편 유형을 검색한 수형 선택 창의 예시를 보여주고 있다.



[그림 2. 수형 선택 창 부분]

움직이는 손(active hand)과 수동적인 손(passive hand)별 특정 수형 검색이 가능하다.

수위 검색

수위는 15개의 신체부위별 혹은 신체와 가까운 위치의 수위를 나타내는 수위 아이콘 페이지에서 선택할 수 있다. 또한 4개의 그룹(선택 창 상단) 중에서 수위를 선택하는 방법도 있다. 그룹을 지정하면 해당 그룹에 속한 모든 수위가 검색 기준에 포함된다.

추가로 (우측 열에서) 두 개의 수위를 선택할 수 있다. 하나는 전체 몸의 앞 부분이며 다른 하나는 얼굴 앞 수위이다. 그림 3은 수위 선택 창을 나타낸다.

1부 발표



[그림 3. 수위 선택 창]

텍스트 검색

대응 텐마크어, 수어 용어 해설(글로스) 및 예문(표기와 번역문)에 대한 텍스트 검색이 가능하다. 이 방법을 통해 기본형은 아니지만 예문에 나타나는 수어 단어를 찾을 수 있다. 구 단위 검색 및 불특정 검색도 가능하다.

주제 검색

70가지 주제 중에서 검색 기준을 선택할 수 있다. 70가지 주제는 9개 그룹과 2개의 하위 그룹으로 구분된다. 상위 그룹에 해당되는 주제를 지정하면 자동적으로 모든 하위 주제가 검색 기준으로 선택된다. 그림 4는 주제 구성 체계를 묘사하고 있다.

복수의 검색 기준을 결합하여 검색도 가능하다. 카테고리별로는 ‘또는(or)’ 조건이 적용되며 검색 카테고리의 조합으로 검색하는 경우 ‘그리고(and)’ 조건을 사용한다.

현재 선택한 기준은 화면의 좌측에 표시된다. 그림 4는 ‘집게손가락을 편’ 수형과 ‘종교’ 주제로 검색한 수어 목록을 보여주고 있다.

General	Time	Day and night
		Weekdays
		Months
		Seasons
	Numbers	
	Measures	
	Colours	
	Shapes	
	Materials	
	People	
Society	Language	Manual Alphabet
		Initialized Signs
		Mouth-Hand-System
		Classifiers
	WH-words	
	Greetings	
	Communication, press	
	Religion	
	Social matters	
	Politics, public administration	
Business, economy	Law	
	Army, warfare	
	Transportation	
	Professions	
	Economy, trade	Money, currencies
Technology, electronics, it	Agriculture, fishing, hunting	
	Industry, handicrafts, tools, machines	
Nature	Building	
	Technology, engineering	
	Electronics, it	
	Biology	Plants
		Animals
	Geography	Geology
		Weather, Climate
		Environment
		Cities, boroughs
		regions, districts
		Countries, nationalities, continents
Education, science	School, teaching	
	Mathematics	
	Astronomy, space research	
	Physics, chemistry	
Home, family	Relatives	
	Home, housekeeping	
	Food and drink	
	Holidays, festive occasions	
	Clothing	
Man	The body	
	Illness, medicine, nursing	
	Personal care	
	Sex, marital relations	
	Moods, feelings, characteristics	
Culture, leisure time	Arts, music, literature	
	Film, theatre, tv, radio	
	Sports, games, playing, hobbies	
	Travel, holiday, outdoor life	

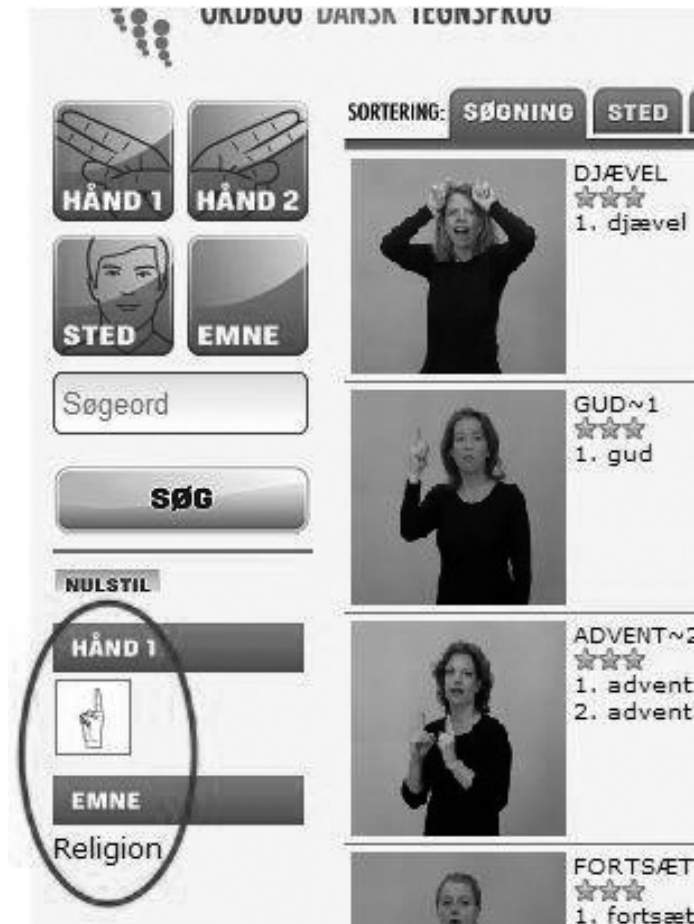
[그림 4. 주제 검색에 사용되는 3 단계 체계]

3.2. 검색 기능 조절

검색을 조율할 수 있도록 검색 기준과 검색 결과는 좌우 나열되어 표시된다(그림 5. 우측에 검색 결과 확

1부 발표

인 가능). 원치 않는 검색 기준을 개별적으로 클릭하여 제외하거나 리셋(reset) 버튼을 클릭하여 모두 삭제할 수 있다.



[그림 5. 수형 및 주제 결합 검색에 따른 기준 목록과 검색 결과 일부]

모든 연관 결과가 검색 결과에 포함되도록 하기 위해 덴마크 수어사전의 검색 기능은 상당히 포괄적이다. 수지적 요소의 경우, 수어의 수위 및 수형은 그 중요도와 무관하게 모두 검색 기능에 고려된다는 것을 의미한다. 같은 이유로 텍스트 기준으로 검색할 때 글로스, 대응어 및 용례 번역 모두가 검색 결과에 반영되어 표시된다.

검색 결과는 연관성 순서대로 나타난다. 수형 및 수위로 검색하는 경우, 수어에 나타나는 순서에 따라 점수가 부여돼 먼저 나타나는 수형 또는 수위가 뒤따르는 수형 또는 수위보다 점수가 높다. 텍스트 검색 결과는 글로스, 대응 덴마크어, 용례 글로스, 용례의 번역문에 있는 단어 순으로 가중치가 부여된다. 즉, 글로스 또는 대응어가 용례에 사용된 단어보다 점수가 높다. 이렇게 계산된 ‘연관성 점수’에 따라 각 검색 결과

는 0-3 ‘연관성 별(relevance star)’로 평가되어 높은 순서대로 표시된다(그림 5 우측 검색 결과 참조). 연관성 점수가 동일한 결과는 수위-수형 순으로 표시된다.

또한 사용자는 연관성 점수와 무관하게 수위 및 수형 기준으로 검색 결과를 정렬하여 확인할 수 있다. 이 방법은 검색 결과 목록 상단에 있는 탭을 클릭하여 정렬할 수 있다. 세번째 탭은 다시 연관성 기준의 정렬 방법으로 검색한 결과를 표시한다.

4. 향후 과제

덴마크 수어사전은 몇 가지 분야에서 개선의 여지가 있다. 단일어 구어사전은 흔히 원천 언어를 메타언어로 사용한다. 이를 수어사전에 적용하기란 거의 불가능에 가깝지만 DTS에 관한 추가 정보를 더함으로써 사전이 더욱 풍부해질 수 있다(예. 의미의 정의). 또한 굴절형에 관한 설명 또는 어원 및 품사와 같이 아직 연구가 많이 진행되지 않은 분야 등 새로운 유형의 정보도 추가할 수 있을 것이다.

참고문헌

- Det Danske Sprog- og Litteraturselskab (2009-2011). Den Danske Ordbog [The Danish Dictionary]. Available at <http://ordnet.dk/ddo> (last accessed 18th September 2018)
- Kristoffersen, J. H. & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language dictionary. In S. Granger & M. Paquot (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-318.
- Kristoffersen, J. H. & Troelsgård, T. (eds) (2016). *Ordbog over Dansk Tegnsprog 3. udgave* [The Danish Sign Language Dictionary 3rd edition]. Center for tegnsprog, Copenhagen. Available at <http://www.tegnsprog.dk> (last accessed 18th September 2018)
- Stokoe, W.C.; Casterline, D.C.; Croneberg, C.G. (1965). *A Dictionary of American Sign Language on Linguistic Principles*, Washington D.C.: Gallaudet College Press.
- Zwitserlood, I, Kristoffersen, J.H. & Thomas Troelsgård (2013). *Issues in Sign Language Lexicography*. In: Jackson, H. (ed.): *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing

Session 1

Structure of a Sign Language Dictionary – www.tegnsprog.dk

Jette H. Kristoffersen,
University College Capital, Denmark

0. Abstract

Sign language lexicography is a young discipline in the world of sign language research. The main problem faced by the sign language lexicographer is that sign languages use the visual modality and have no written representation commonly used among the native signers. Sign language lexicography has over the last decades benefited tremendously from the electronic medium, and sign language dictionaries are no longer mere wordlists as in the past, they are real scientifically based dictionaries, with information types similar to those found in dictionaries of spoken languages. The visual modality is still a challenge on both how to perform searches and how to present the information on the signs. This paper describes the structure of the Danish Sign Language Dictionary (Kristoffersen & Troelsgård, 2016) and demonstrate a possible structure of a Sign Language Dictionary.

1. Introduction

Until recent times most sign language dictionaries were mere wordlists, with words from a spoken language, each accompanied by a picture or a verbal description of a sign. The source language was typically the surrounding spoken language and the sign language was the target language.

Over the last decades, sign language lexicography has developed rapidly and a series of scientifically based sign language dictionaries has emerged, the first being A Dictionary of American Sign Language on Linguistic Principles, Stokoe et. al. (1965). Today Sign language dictionaries are no longer wordlist – they are real dictionaries with the types of information similar to those found in dictionaries of spoken languages. And they typically have sign language as the source language.

The sign language lexicography has benefited tremendously from technological advances. Thus, the decision of

how to represent signs in a dictionary used to be a major problem, but now digital video recordings of signs can easily be integrated into electronic dictionaries (cf. Zwitterlood et al, 2013).

Still, certain obstacles remain for the sign language lexicographer. For example, some form of formal notation of the signs is needed in order to refer uniquely to the signs, to perform sign searches and to order the signs in search results etc. Also, a sign language dictionary with sign language as the source language requires more advanced search functionalities compared to a dictionary of a written language, for example in order to enable searches on manual features.

The Danish Sign Language Dictionary (Kristoffersen & Troelsgård, 2016) is practically a bidirectional bilingual Danish Sign Language (DTS) - Danish dictionary, but its structure places it somewhere between bilingual and monolingual (Kristoffersen & Troelsgård, 2012). There is only one set of entries, with DTS signs as headwords. The entries are structured according to the semantics of the signs, and almost all information, including usage examples, concerns DTS, except for disambiguating explanations of homographic or strongly polysemous Danish equivalents. This approach was chosen in order to keep focus on DTS, as the need for information about Danish words can easily be met by existing Danish all-round monolingual dictionaries. As a result of this decision, the Danish Sign Language Dictionary could be described as monolingual dictionary, which instead of definitions has (searchable) equivalents in another language, Danish, which is also the general metalanguage of the dictionary.

2. Entry layout in the Danish Sign Language Dictionary

The entries of the Danish Sign Language Dictionary have four sections (cf. figure 1):

Section 1: Entry header

In this section the sign headword is shown as a photo and a gloss. The first occurring location and the most prominent handshape of the sign are shown as icons. Lastly, there are buttons for opening separate windows with a printable sign overview and with a concordance view of all usage examples in the entire dictionary including the sign.

Section 2: Video window

By default the base form of the sign headword is shown. If there are phonological variants of the sign, a play button for each variant is shown in top of the video window. There are also control buttons for slow-motion and enlarged view. Other videos, e.g. of usage examples, are activated by clicking play buttons in different sections of the entry.

Section 3: Meanings window In this section the meanings of the sign are described. The meaning description includes: Danish equivalents and/or a description of the sign's usage (for function signs), information about possible mouth movements, cross-references to synonyms etc., information about restricted use, and example


Session 1

sentences. Semantically opaque compounds with the sign are shown below the regular meanings.

Danish equivalents that can be looked up in the Danish Dictionary¹⁾, are supplied with a clickable link to this dictionary. Thus, users can learn more about the meaning and use of the Danish words.

Section 4: Additional information


This section holds cross-references to homonyms and to common frozen forms of the sign (for classifier entries). In addition to this, frequent co-occurrences with the sign are shown in this section.





1

DEL



2

Grundform af: DEL.

4

Samme form som:
SKARP, SVIGERINDE

Almindelige tegnforbindelser:
BY~1 DEL bydel
EN DEL halvanden

1. dele[~], del[~], detalje[~], fordele[~], fordeling[~], andel[~]

/halv/

HALV, ADSKILLE, DIVIDERE

TIME DEL FEM-TIMER FEM-TIMER TRE-TIMER TRE-TIMER

Vi fordelte timerne sådan, at de to hver fik fem timer og han og jeg fik tre timer hver.

KLATRE TO DEL JEG TAGE-AFSTED KØBE HELLERE OMGANG JEG INVESTERE OK

Skal vi ikke deles om en stige? Jeg tager hen og køber den. I første omgang lægger jeg ud, ok?

2. afdeling[~], afsnit[~], filial[~]

HOSPITAL EKSISTENS MASSER-AF DEL PROFORM

Der er mange afdelinger på et hospital.

3. halv[~]

HALV

MADRAS PEG DEL KØBE BANDE BILLIG JEG TAGE

Tæppet var sat ned til halv pris! Det var fandeme billigt, så jeg tog det!

4. speciale[~], specialist[~]

EKSPERT~2, EKSPERT~1

JEG SKRIVE DEL OPGAVE SKAL OVERRÆKKE INDEN JUL TRAVL JEG

Jeg skriver nu speciale, som skal afleveres inden jul. Jeg har travlt.

JEG SLEM ALLERGI EKSISTENS LÆGE HENVENDE JEG HUD DEL

Jeg er slemt ramt af eksem, så lægen har henvist mig til en hudspecialist.

5. gå op i detaljer

[i], <intet>

DYB

PEG DEL JEG DÆMPE

Han var så detaljeret, at jeg bad ham slappe lidt af.

3

[Figure 1. The four sections of an entry in the Danish Sign Language Dictionary.]

1) The Danish Dictionary is an all-round, monolingual dictionary of contemporary Danish. The dictionary is freely accessible at: www.ordnet.dk/ddo

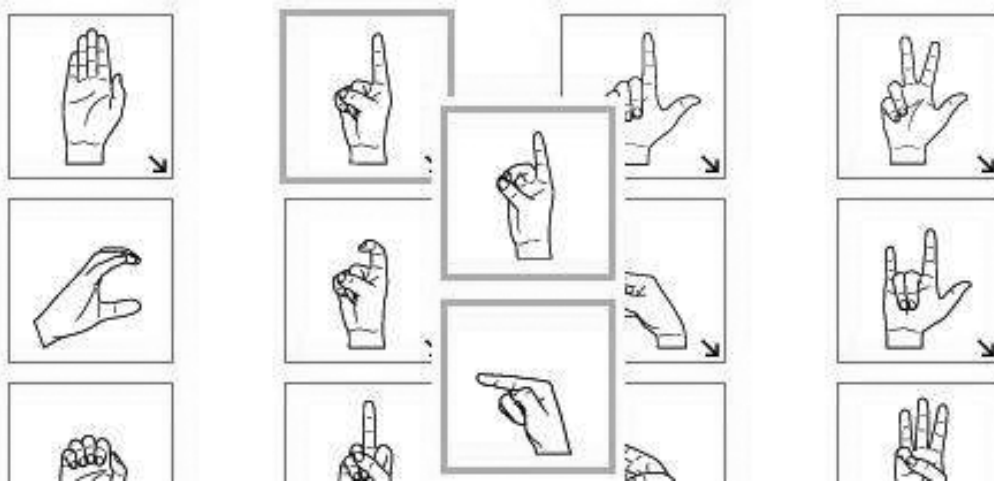
3. Search facilities in the Danish Sign Language Dictionary

3.1. Choosing search criteria

The signs in the Danish Sign Language Dictionary can be looked up through search criteria from different categories: handshape, location, topic and text.

Handshape

There are 65 searchable handshapes. The handshapes to be used in the search criteria are selected in groups of related handshapes, but unwanted handshapes can be deselected individually. This enables the user to select single handshapes or custom groups of handshapes, independently of the default handshape grouping. Figure 2 shows a part of the handshape selection window, with the extended index finger group currently activated.



[Figure 2. Partial view of the handshape selection window.]

Particular handshapes for the active and the passive hand can be specified.

Location

Location is chosen from a page with 15 location icons, representing locations on or near the body. In addition to these it is possible to choose location through four “group headers” (= top row of the selection window). If a group header is chosen, all locations in the group are included in the search criteria. Additionally (in the right-most column) two options are available for choosing locations in the space in front of the body: one for the neutral space in general, and one for locations in front of the face. The location selection window is shown in figure 3.

Session 1



[Figure 3. The location selection window.]

Text

Text searches are performed on Danish equivalents as well as on sign glosses and example sentences (both transcriptions and translations). This enables users to find signs that are not themselves lemmas in the dictionary, but appear in example sentences. Phrase searches and wildcard searches are possible.

Topic

Topics can be chosen as search criteria from a list of 70 topics. The topics are ordered into nine groups, with subgroups in two levels. Topics which are group headers automatically include all subordinate topics in the search criteria. Figure 4 shows an overview of the topic hierarchy.

Several criteria categories can be combined. Inside each category an “or” condition is used, while “and” conditions are used for category combinations.

The currently chosen criteria are shown in the left side of the screen. Figure 4 shows the criteria list for a search for signs that are produced with the handshape extended index finger and concern the topic religion.

Structure of a Sign Language Dictionary | Jette H. Kristoffersen

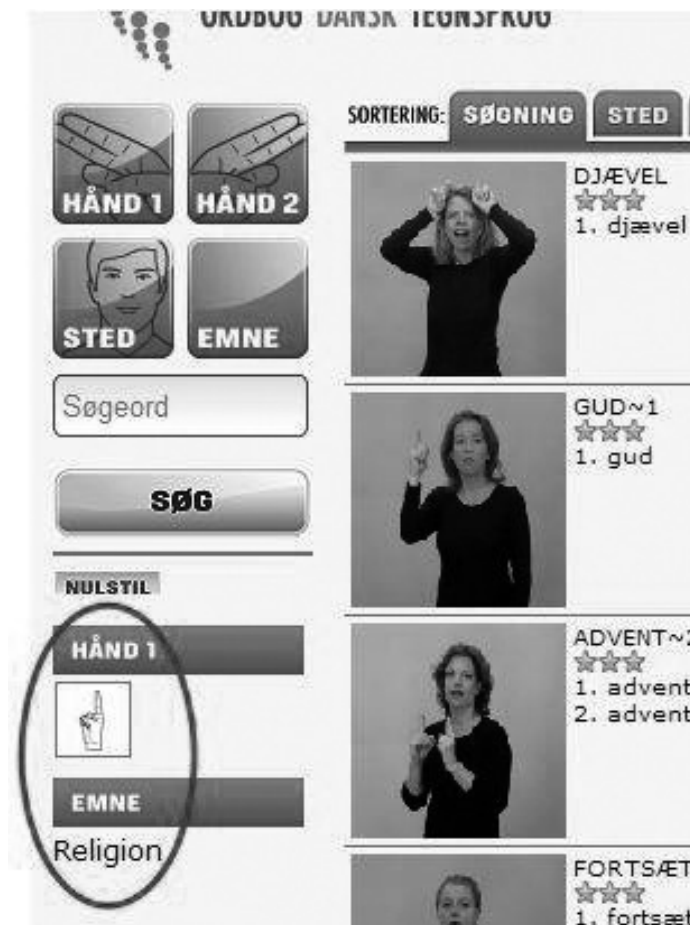
General	Time	Day and night
		Weekdays
		Months
		Seasons
	Numbers	
	Measures	
	Colours	
	Shapes	
	Materials	
	People	
	Language	Manual Alphabet
		Initialized Signs
		Mouth-Hand-System
		Classifiers
Society	WH-words	
	Greetings	
	Communication, press	
	Religion	
	Social matters	
	Politics, public administration	
	Law	
	Army, warfare	
Business, economy	Transportation	
	Professions	
	Economy, trade	Money, currencies
	Agriculture, fishing, hunting	
	Industry, handicrafts, tools, machines	
Technology, electronics, it	Building	
	Technology, engineering	
Nature	Electronics, it	
	Biology	Plants
		Animals
	Geography	Geology
		Weather, Climate
		Environment
		Cities, boroughs
		regions, districts
		Countries, nationalities, continents
Education, science	School, teaching	
	Mathematics	
	Astronomy, space research	
	Physics, chemistry	
Home, family	Relatives	
	Home, housekeeping	
	Food and drink	
	Holidays, festive occasions	
	Clothing	
Man	The body	
	Illness, medicine, nursing	
	Personal care	
	Sex, marital relations	
	Moods, feelings, characteristics	
Culture, leisure time	Arts, music, literature	
	Film, theatre, tv, radio	
	Sports, games, playing, hobbies	
	Travel, holiday, outdoor life	

[Figure 4. The three-level hierarchy used for topic search.]

Session 1

3.2. Adjusting the search

The search criteria and the result list are shown side by side in order to facilitate adjustments of the search (cf. figure 5, where the result list can be seen to the right). For the same purpose the search criteria can be deselected individually, by clicking them, or cleared altogether, by clicking a reset button.



[Figure 5. Criteria list (and part of the result list) for a combined search on handshape and topic.]

In order to secure that all possibly relevant matches are included in a search result, the search function in the Danish Sign Language Dictionary is quite broad. For manual features, that means that all occurring locations and handshapes in a sign, both prominent and non-prominent, are taken into account by the search function. Similarly, all text string matches, both among glosses, equivalents, and translations of usage examples are included in the search result.

The matches of a search are ordered by relevance. For handshape and location matches this means that they are weighted according to their appearance in the sign, so that the first occurring handshape or location scores higher than the following. Text search matches are weighted in the following order: glosses, Danish equivalents, glosses

in usage examples, words in the translations of usage examples. Thus, a matching gloss or equivalent scores higher, than a match in a usage example. Based on the calculated relevance scores each match receives from none to three “relevance stars”, and the matches are ordered accordingly (see the result list shown to the right in figure 5). Matches with equal relevance score are ordered firstly by location, secondly by handshape.

In addition to this, the user can choose between two sort orders: firstly by location, secondly by handshape, with no regard to the relevance score. These alternative sort orders are activated by clicking their according tabs in top of the result list. A third tab re-establishes the default, relevance-based sort order.

4. Future challenges

The Danish Sign Language Dictionary could be developed in several areas. Spoken monolingual dictionaries often use the source language also as meta-language, for sign language dictionaries that would be an almost impossible challenge, but the dictionary would benefit from having more information in DTS, e.g. sense definitions. Also some additional information types could be added, e.g. descriptions of inflected forms, or lesser studied fields like etymology and part-of-speech.

References

- Det Danske Sprog- og Litteraturselskab (2009-2011). Den Danske Ordbog [The Danish Dictionary]. Available at <http://ordnet.dk/ddo> (last accessed 18th September 2018)
- Kristoffersen, J. H. & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language dictionary. In S. Granger & M. Paquot (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-318.
- Kristoffersen, J. H. & Troelsgård, T. (eds) (2016). *Ordbog over Dansk Tegnsprog 3. udgave* [The Danish Sign Language Dictionary 3rd edition]. Center for tegnsprog, Copenhagen. Available at <http://www.tegnsprog.dk> (last accessed 18th September 2018)
- Stokoe, W.C.; Casterline, D.C.; Croneberg, C.G. (1965). *A Dictionary of American Sign Language on Linguistic Principles*, Washington D.C.: Gallaudet College Press.
- Zwitserlood, I, Kristoffersen, J.H. & Thomas Troelsgård (2013). *Issues in Sign Language Lexicography*. In: Jackson, H. (ed.): *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전



Records of Visible Language:
Sign Language Dictionary

주제 3 | 수어사전 편찬의 과제

요한나 메시
스웨덴, 스톡홀름 대학교

Johanna Mesch
Stockholm University, Sweden

2부 발표

수어사전 편찬의 과제

요한나 메시
스웨덴, 스톡홀름 대학교

본 글은 스웨덴 수어사전(Swedish Sign Language Dictionary)이 어떻게 개발되었는지 설명하고 수어사전을 편찬하는 데 오랜 시간이 걸리는 이유를 살펴볼 것이다. 스웨덴의 수어사전화 작업은 1998년 스톡홀름 대학에서 시작되었으며, 그 결과 2001년 첫 온라인 사전이 나오게 되었다. 스웨덴 수어사전(SSLD)은 2008년에 처음 만들어진 후 개발이 계속되고 있다. 2003년 말뭉치(corpus) 구축이 시작되고 ‘스웨덴 수어 말뭉치 2009~2011’ 3개년 프로젝트 덕분에 어휘 주석으로 말뭉치 데이터가 확대되자, 이를 온라인 스웨덴 수어사전의 지속적인 개발과 스웨덴 수어사전의 새로운 수어 및 어휘 변형을 위한 자료로 활용할 수 있는 방안에 대한 논의가 시작되었다.

키워드: 수어사전, 사전학, 스웨덴 수어

서론

수어사전은 여러 국가 내 수어의 지위를 높인다. 또한 학습자와 교수자, 가족 및 수어 연구자들의 요구에 부응하기 위해 필요한 도구이다(McKee & Vale, 2017). 책 형태의 스웨덴 수어사전은 구화주의 시대의 영향 아래 1916년 처음 출판되었다(Österberg, 1916). 가장 최근 책으로 출판된 사전은 스웨덴 전국 농인 협회가 1997년 10년 간의 작업 끝에 편찬한 것이다(Svenskt teckenspråkslexikon, 1997). 스웨덴 수어사전화 작업은 1988년 스톡홀름 대학교 언어학과 수어부가 처음 시작했다. 이 프로젝트의 결과로 3,132개의 수어 표제어가 담긴 첫 사전이 2001년 스웨덴 수어사전 디지털 버전이라는 제목으로 온라인에 공개되었다. 이 사전에는 스웨덴 전국 농인 협회가 출판한 책자 사전에서 따온 수어 동작 영상 파일이 포함되었다. 하지만 책자 사전의 기준을 따르는 것이 수어사전을 제한적인 자료로 만든다는 사실이 곧 명확해졌다. 목적이 다른 유연한 사전을 만들어낼 필요가 있었다.

그 이후 주요 어휘 데이터베이스인 스웨덴 수어사전(Swedish Sign Language Dictionary)이 2008년에

만들어졌으며 지금까지 개발이 계속되고 있다(Mesch, Wallin, & Björkstrand, 2012). 현재 스웨덴 수어 사전에는 17,800개의 수어 표제어와 4,070개의 예시 문장이 포함되어 있다(2018년 9월 기준). 이 프로젝트와 관련된 사전화 실무단은 스웨덴 수어의 단어를 기록하고 온라인 사전을 지속적으로 업그레이드한다. 사전 편찬자들은 자신들이 속한 언어 관련 기관뿐 아니라 다른 자원을 기반으로 작업하는 경우가 많다.

어떤 언어라도 언어 자원은 해당 언어나 그 언어에 포함된 사용역(register)을 대표적으로 보여주는 것이 중요하다. 그렇기 때문에 언어 자원 또한 언어의 기록 문서로 사용될 수 있는 것이다. 더 중요한 것이 있다. 수어사전은 다양한 목적에 적합한 자료여야 하고 수어 학습자, 통역사, 교수자, 농인 커뮤니티 및 다운증후군이나 기타 언어장애를 가진 사람들 등 여러 집단에게 좋은 자원이 될 수 있어야 한다.

사전화 작업과 작업자

사전팀은 농인인 3명의 상근 직원으로 구성되어 있으며 스웨덴 수어 단어를 기록하고 온라인 사전을 계속해서 업데이트하는 일을 한다. 사전화 작업은 장애가 있는 학생들을 위해 스톡홀름 대학교가 대학 차원에서 국가 예산을 통해 자금을 지원한다. (내부) 자문단도 함께 작업해 왔다. 사전의 규모가 계속 커지고 있기 때문에 외부 자문단이 필요한 상황이다. 따라서 언어 협의회와 함께 자문단이 구성될 예정이다.

데이터베이스는 파일메이커 프로(FileMaker Pro) 데이터베이스 파일로 만들어졌고 구축이 계속되고 있다. 스웨덴에서는 사인뱅크(Signbank)가 사용되지 않는데, 가까운 시일 내에 사인뱅크를 연구 도구로 사용할 계획이 있기는 하다. 현재 사전화 실무단은 파일메이커와 ELAN을 기술적으로 연결할 수 있는 방법을 찾는 중이다.

문장 구성 방법과 관련해서, 작업자(외부 자문단 위원)는 스웨덴어 문자 없이 사람이 수어 동작을 보여주는 정지 화면만을 보게 된다. 이는 서로 의미가 다른 여러 수어 동작 간 연관성을 구축한다. 그 후 수어 작업자는 자신만의 사용 예시를 문장의 형태로 고안해야 한다. 이러한 유도도 좋기는 하지만, 앞으로는 이 방식에 사전 편찬자가 말뭉치에서 관련 문장을 검색하고 검색된 문장을 사용에 적절한 형태로 변경하여 사전에 추가하는 방식이 더해질 것이다.

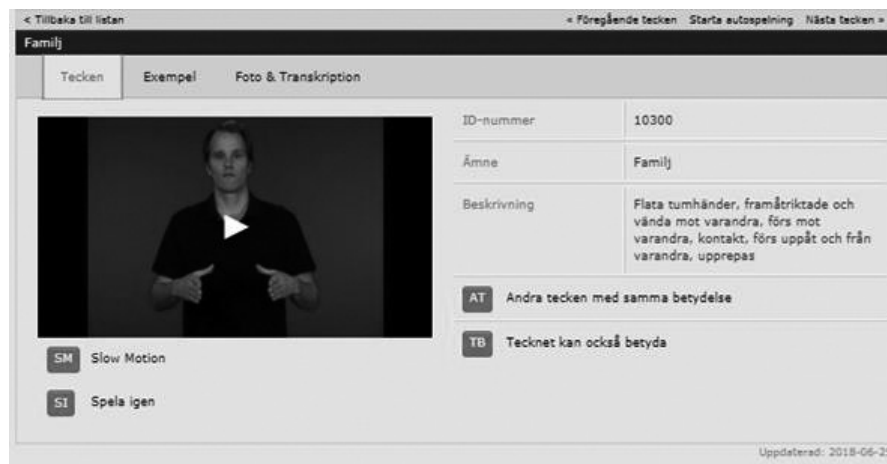
사전의 설계

주요 어휘 데이터베이스인 스웨덴 수어사전은 2008년 만들어졌으며 그 이후 인터페이스 개발이 계속되고 있다. 스웨덴 수어사전 웹사이트는 <https://teckensprakslexikon.su.se/>이다. 기술적 지원의 목적은 스웨덴 수어사전이 좀 더 사용자 친화적이고 아이폰이나 안드로이드 앱 등 기타 플랫폼에서도 사용 가능하도록 개발하는 것이었다. 스웨덴 수어사전에서 수어 동작은 수형, 수향, 수위, 수동, 스웨덴어 번역

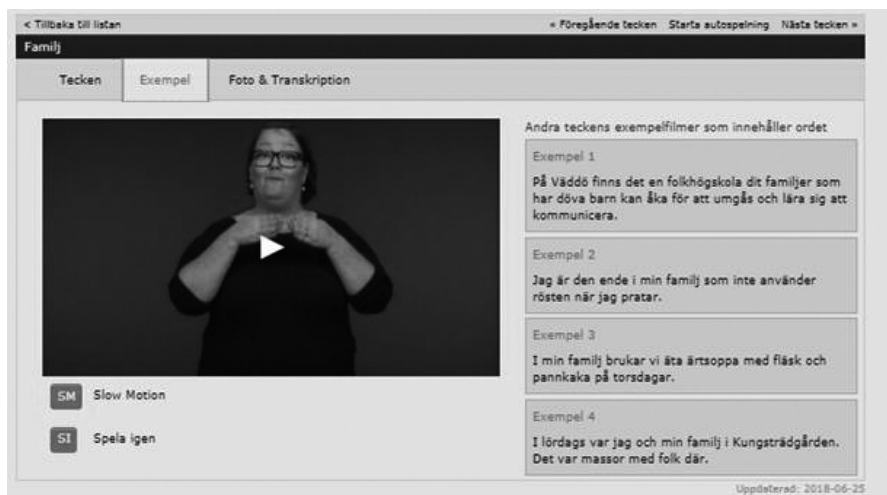
2부 발표

과 수어 표기로 설명된다. 사전의 각 항목은 단어 (또는 구절) 영상, 스웨덴어 번역, 음운 정보 및 음운 또는 의미상 동등한 수어 동작과의 사전 내 교차 연결(동음어와 동의어 등, 사진 1-3)로 대표된다(Svenskt teckenspråkslexikon, 2018). 스웨덴 수어 말뭉치 주석자를 지원하기 위해 ID-글로스가 추가된다.

수어 동작은 세 가지 방식(사진 4)이나 주제별로(46개의 하위분류) 검색 가능하다(가족, 스웨덴 도시와 지역, 의료 서비스나 지화 등). 또한 각 수어(일반 및 특수 용어) 동작에 ID 번호가 있다. 하지만 ID 글로스만으로는 수어를 완벽하게 설명할 수 없다. 각 수어 동작에 대한 설명과 번역도 있다. 스웨덴 수어사전에는 또한 첫 스웨덴 수어 책자 사전(Österberg, 1916)에 포함된 415개의 표제어와 함께 구 수어 동작이 추가되었다. <https://teckensprakslexikon.su.se/kategori/osterberg-1916> 빈도는 특정 기한 내에 방문자가 얼마나 자주 사전 사이트를 재방문했는지를 알려준다.

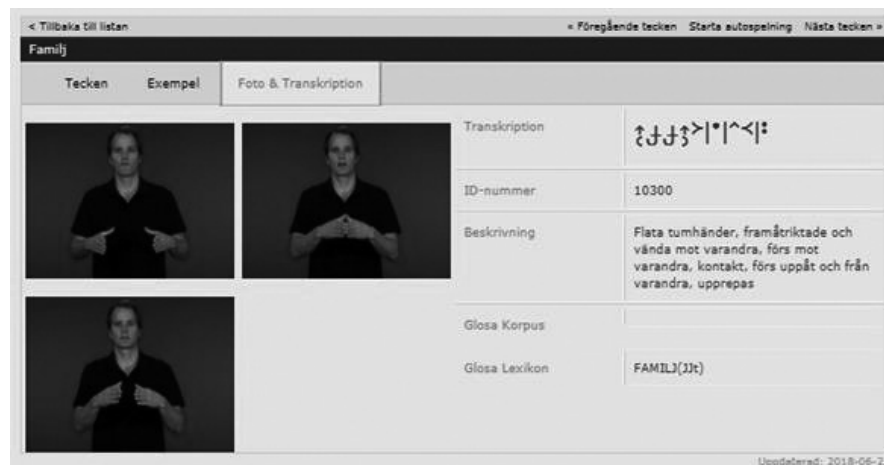


[사진 1.] 느린 재생, 다시 재생, 같은 의미를 가진 다른 동작, 이 동작의 다른 의미 옵션과 함께 동작을 보여준다



[그림 2.] 문장의 예를 보여준다

수어사전 편찬의 과제 | 요한나 메시



[그림 3.] 정지된 화면사진(프린트용)을 스웨덴 수어사전용 표기, ID 글로스, 동작 설명, 말뭉치 주석용 글로스, 사전용 글로스와 함께 보여준다



[그림 4.] 수형, 수향, 수위 및 수동으로 검색한다.

스웨덴 수어사전은 L2 교육과 자기학습의 표준화된 자료로 사용된다. 학습 목적으로 사전을 사용하는 것은 학습자, 가족 및 교육적 환경에 있어 중요하다. 농인 커뮤니티 역시 페이스북 그룹 덕분에 사전을 유용하게 이용하고 있다. 학습자를 위한 퀴즈 또한 만들어졌다. <https://tspquiz.se/app/#/start> 트위터 봇인 “Ett techken varje dag”(하루에 한 동작)이 2018년 1월에 만들어졌으며 학습자들 사이에서 반응이 좋다.[<https://twitter.com/allatecken>]. 이 트위터 봇은 스웨덴 수어사전에서 무작위로 선택한 수어 동작을 하나씩 올린다(그림 5).

2부 발표



[그림 5.] 하루에 한 동작 트위터 봇

수어사전을 지원하는 언어 자원 - 스웨덴어 수어 말뭉치

현재 일반 대중에게 공개된 스웨덴 수어의 주요 언어 자원은 두 가지로 사전과 말뭉치이다(Mesch, 2012; Mesch & Wallin, 2012, 2015; Mesch, Wallin, & Björkstrand, 2012; Mesch, Wallin, Nilsson, & Bergman, 2012). 두 언어 자원 모두 스웨덴 수어를 위한 기록 문서 역할을 한다. 스웨덴 수어의 첫 말뭉치는 2004년 ECHO 프로젝트와 연계되어 만들어졌다(O. Crasborn et al., 2007).

이 말뭉치는 스웨덴 수어 말뭉치(SSLC)로 24시간 분량의 영상 데이터 대화, 서술, 프레젠테이션으로 구성되어 있으며 2009~2011년 수어 사용자 42명으로부터 수집된 자료로 릭스방켄스 유빌레움스뱅크(Riksbankens Jubileumsbank)가 자금을 지원했다(Mesch, Wallin, Nilsson, et al., 2012; Wallin & Mesch, 2018). 말뭉치 구축, 특히 주석 작업은 매우 오랜 시간이 걸린다. 수집된 24시간 분량의 데이터 중 약 절반 정도에 주석이 달렸다(2018년 9월 기준). 그 외 스웨덴 수어의 말뭉치에는 촉각 수어 말뭉치(Mesch, 2016)와 스웨덴 수어 학습자 말뭉치(Schönström & Mesch, 2017)가 있다. 스웨덴 촉각 수어 말뭉치는 8명의 수어 사용자로부터 2007년과 2011년 수집된 대화와 유도로 구성되어 있다. 수집된 3시간 30분 분량의 자료 중 약 40%(수어 토큰 6,000개)에 주석이 달렸다. 스웨덴 수어 말뭉치와 다른 말뭉치의 주석 작업은 주석자 10명(농인 8명 비농인 2명)이 비상근으로 참여하고 있으며 여전히 진행 중이다.

스톡홀름 대학 언어학과 수어부가 수어사전과 스웨덴 수어 말뭉치를 고안 및 유지하고 있지만, 처음에는 서로 독립적으로 만들어졌다. 이로 인해서 두 자료 간 기능적 불일치가 생겨났다. 하지만 두 프로젝트는 작년에 통합되기 시작했고 현재는 이미 서로를 보완하는 수준이다. 예를 들어, 스웨덴 수어 말뭉치의 수어 토큰에는 스웨덴 수어사전 ID 번호가 주석으로 달리고(Riemer Kankkonen, Björkstrand, Mesch, & Börstell, 2018) 수어 동작이 스웨덴 수어 말뭉치에 포함되면 스웨덴 수어사전에도 추가된다. 또한 스웨덴

수어사전 온라인 인터페이스는 동의어 검색 결과를 스웨덴 수어 말뭉치 어휘 빈도수 인터페이스를 통해 계산된 스웨덴 수어 말뭉치 빈도수에 따라 분류한다(Börstell & Östling, 2016). 스웨덴 수어 말뭉치 데이터 빈도수 도구는 경쟁적인 이형 수어 간 상대적 빈도수를 보여준다(예: 수어 동작 TIO와 TIO(Y) '10'). 스웨덴 수어 말뭉치가 확대되면 좋은 언어 자원이 될 것이다.

<https://www.ling.su.se/teckenspr%C3%A5ksresurser/verktyg/frekvens>

스웨덴 수어 말뭉치는 현재 약 113,000개의 수어 토큰을 포함하는데, 말뭉치 데이터 수집 단계에서 구체적인 어휘 변형 유도 작업이 없었기 때문에(Stamp et al., 2014), 많은 경우 동의어나 수어 동작의 변형이 말뭉치 데이터 사용만으로는 조사가 불가능하다. 113,000개의 수어 토큰은 언어의 변형을 체계적으로 조사하기 위해 사용되는 구화 말뭉치보다는 훨씬 작지만 수어 말뭉치로는 꽤 큰 규모다.

클라우드소싱 - 언어 자원 개선

사전, 즉 스웨덴 수어사전을 개발하고 개선하기 위해 우리는 농인 커뮤니티를 참여시켜 새로운 수어 동작을 수집하고 동의어와의 유사성 및 동의어들 간 차이에 대한 데이터를 수집한다. - 바로 우리의 언어 자원, 즉 스웨덴 수어사전을 개선하기 위한 클라우드소싱이다. 두 가지 방식으로 작업이 진행된다: a) 스웨덴 수어 사용자와 수어 동작 사용, 어휘의 차이, 그리고 수어 동작 구성에 대해 에스엔에스(SNS)상에서 토론하고 b) 스웨덴 수어 사용자와 여러 장소에서 직접적으로 교류하며(전국 농인 기관 및 농인 커뮤니티 사업 및 활동 연례 참석) 수어와 수어에 대한 판단을 수집한다(Riemer Kankkonen et al., 2018).

Teckenspråkslexikon ('수어사전')이라는 이름의 페이스북 그룹은 2014년 10월에 시작되었으며 사전팀이 관리한다. 2,642명의 회원이 있으며(2018년 5월 기준) 회원들은 다양한 방식으로 교류하면서 어떤 수어 동작을 사용해야 하고 수어 동작 변형 간의 차이가 무엇인지 등에 대해 토론한다. 회원들은 스웨덴어나 스웨덴 수어 영상을 이용해서 질문을 올리거나 답글을 단다. 지난 2년간(2016년 1월부터 2018년 1월까지) 593개의 글이 올라왔고 5,817건의 교류가 있었다. 여기에는 X를 수어로 어떻게 하나요?(53%), Y사람에 대한 수어 동작이 있나요?(20%) 및 사용 방법에 대한 일반적인 질문(10%)과 특정 수어 동작의 어원(-0.2%) 등의 질문이 포함된다.

결론

수어사전을 개발하고 유지하는 것은 많은 시간을 요하는 자료 집약적인 작업이다. 비디오 파일을 업로드하거나 고장 및 해킹을 방지하면서 기술적인 발전을 지속하는 것이 중요하다. 사전은 또한 양질의 내용과

2부 발표

사용자 친화적인 자원을 유지하기 위해 연구 및 기술적 지원이 필요하다. 수어사전을 업데이트하고 개발하는 방법에는 언어 지식, 이전 영상 내용 등의 기타 자원, TV 프로그램과 영상 메시지, 그리고 국가별 수어에 대한 데이터를 에스엔에스(페이스북 그룹 등) 상에서 농인 커뮤니티로부터 직접 클라우드소싱하기 등이 포함된다. 수어의 말뭉치 규모는 아직 작은 수준이라 말뭉치 데이터만으로는 수어 형태의 변형을 조사할 수 없다. 사전을 위한 클라우드소싱은 언어 자원을 개선하는 유용하고 빠른 방법이다. 오프라인 데이터 수집은 사전팀에게 개인적으로 농인 커뮤니티를 만나 교류할 수 있는 기회를 제공하는데 이는 언어 기록화 작업의 중요한 요소다. 그 뿐만 아니라 개인과의 직접적 교류로 수어 동작 이형에 대해 좀 더 질적인 접근이 가능할 수도 있다. 각 수어 사용자와 관련된 메타 데이터는 말뭉치 등 기록화 작업에 추가될 필요가 있는 특정 개인이나 집단을 대상으로 한다. 말뭉치를 기반으로 한 사인뱅크는 수어 동작의 분류 정리화를 고려하는 연구자들에게 필요한 도구다. 예를 들어 사인뱅크에 수어 동작 글로스가 변경되거나 새롭게 추가될 경우, 연결되어 있는 ELAN 주석 파일을 통해 자동으로 업데이트되어 주석자와 연구자들의 작업이 쉬워진다. 초기 사전이나 현재 어휘 데이터베이스(스웨덴 수어 어휘 데이터베이스)는 말뭉치의 분류 정리화를 위한 말뭉치 프로젝트와 함께 사용되면 충분히 좋은 자원이 될 수 있다. 구어/문어 복합의 경우(의학, 기술 또는 직업 관련 주제에서 매우 흔함.) 수어사전에 표제어를 어떻게 넣을지는 아직도 명확하지 않으며, 빈도수가 높은 생산적 수어(묘사어나 분류사)도 사전에 추가되어야 한다.

참고문헌

- Börstell, C., & Östling, R. (2016). Visualizing lects in a sign language corpus : Mining lexical variation data in lects of Swedish Sign Language. In E. Efthimiou, S. E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining (pp. 13–18). Paris: ELRA.
- Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., Van Der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, 12(4).
- McKee, R., & Vale, M. (2017). Sign language lexicography. In P. Hanks & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–22). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-45369-4_34-1
- Mesch, J. (2012). Swedish Sign Language Corpus. *Deaf Studies Digital Journal*, 3. Retrieved from http://dsdj.gallaudet.edu/index.php?issue=4§ion_id=2&entry_id=128

- Mesch, J. (2016). Dataset. Tactile Swedish Sign Language Corpus. Department of Linguistics, Stockholm University.
- Mesch, J., & Wallin, L. (2012). From meaning to signs and back: Lexicography and the Swedish Sign Language Corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* [Language Resources and Evaluation Conference (LREC)] (pp. 123–126). Paris: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf
- Mesch, J., & Wallin, L. (2015). Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1), 103–121. <http://doi.org/10.1075/ijcl.20.1.05mes>
- Mesch, J., Wallin, L., & Björkstrand, T. (2012). Sign language resources in Sweden: Dictionary and corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon* [Language Resources and Evaluation Conference (LREC)] (pp. 127–130). Paris: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf
- Mesch, J., Wallin, L., Nilsson, A.-L., & Bergman, B. (2012). Dataset. Swedish Sign Language Corpus project 2009--2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University. Retrieved from <http://www.ling.su.se/teckensprakskorpus>
- Österberg, O. (1916). *Teckenspråket*. Uppsala: P. Alfr. Persons förlag.
- Riemer Kankkonen, N., Björkstrand, T., Mesch, J., & Börstell, C. (2018). Crowdsourcing for the Swedish Sign Language Dictionary. In M. Bono, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, ... Y. Osugi (Eds.), *Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community* [Language Resources and Evaluation Conference (LREC)] (pp. 171–174). Paris: European Language Resources Association (ELRA).
- Schönström, K., & Mesch, J. (2017). Dataset. The project From speech to sign – learning Swedish Sign Language as a second language. Department of Linguistics, Stockholm University.
- Svenskt teckenspråkslexikon. (1997). Leksand: Sveriges Dövas Riksförbund.
- Wallin, L., & Mesch, J. (2018). Annoteringskonventioner för teckenspråkstexter. Version 6, januari 2018. [Annotation guidelines for sign language texts].

Session 2

Challenges of creating a Sign Dictionary

Johanna Mesch,
Stockholm University, Sweden

The article will describe how the Swedish Sign Language Dictionary has developed, and why it takes a long time to establish such a dictionary. A lexicographic work of Swedish Sign Language was initiated in 1988 at Stockholm University, and it resulted, in 2001, the first dictionary online. The Swedish Sign Language Dictionary was created in 2008 and has been in development since. When the direction of the corpus construction started in 2003, and when the corpus data, thanks to the three-years project of the Swedish Sign Language Corpus 2009-2011, expanded with gloss annotations, a discussion has arisen about how the Online Swedish Sign Language Dictionary should continue to be in its development and in which direction, and how to use the SSL Corpus as a source of input for new signs and lexical variation in the SSL Dictionary.

Keywords: Sign language dictionary, lexicography, Swedish Sign Language

Introduction

A sign language dictionary elevates the status for existence of sign language(s) in many countries. It is also a necessary tool for serving the needs of learners, teachers, families and researchers of signed languages (McKee & Vale, 2017). The first Swedish Sign Language Dictionary in a book format was published in 1916 in the shadow of the oralism period (Österberg, 1916). The latest dictionary in printed book format was published by the Swedish National Deaf Association in 1997 after ten years' work (Svenskt teckenspråkslexikon, 1997). The lexicographic work of Swedish Sign Language was initiated in 1988 by the Sign Language Section of the Department of Linguistics at Stockholm University. In 2001, the first dictionary resulting from the project went online, entitled the Digital version of Swedish Sign Language Dictionary and included 3,132 sign entries. This dictionary includes video files for signs taken from a printed dictionary of the Swedish National Deaf Association. It soon became clear that following criteria of the printed dictionary was making it a limited resource. It needed to be a flexible dictionary for different purposes.

A follow-up dictionary, the main lexical database, the Swedish Sign Language Dictionary was created in 2008 and has been in development since then (Mesch, Wallin, & Björkstrand, 2012). Today, it has approximately 17,800 sign entries and 4,070 sentence examples (September 2018). The lexicographic work group associated with this project documents the vocabulary of Swedish Sign Language and continuously updates the web-based dictionary. The work of the lexicographers is often based on their intuitions concerning language, but also on different sources.

An important aspect of any language resource is that it is representative of the language or register that it covers. Because of this, the resources also serve as language documentation. This is perhaps even more important. A sign language dictionary should be a good resource for different purposes and different groups such as sign language learners, interpreters, sign language teachers, the deaf community, and other groups such as people with Downs syndrome and other language disorders.

The lexicographic work and actors

The dictionary team consists of three full-time employees, all Deaf, who document SSL vocabulary and continuously update the web-based dictionary. The work is financed by Stockholm University through national financing to support for disabled students at university level. There has been the (internal) advisory group. Because the dictionary has been growing bigger, an external advisory group is needed. With The Language Council, an advisory group will be established.

The database has been created as FileMaker Pro database files and its construction is ongoing. Signbank is not used in Sweden, although it is planned to have a Signbank as research tool in near future, and the group is trying to find a possible technical connection between FileMaker and ELAN.

With respect to the method for creating sentences, the actors (members of the external advisory group) are given still photos only, in which a person is presenting a sign, without any written Swedish words. This creates associations with many signs of different meanings. Then the sign actors have to devise their own usage examples in the form of sentences. This elicitation may be good, but in the future it will be combined with the possibility for lexicographers to look up related sentences in the corpus and change them to a suitable form for use in the dictionary.

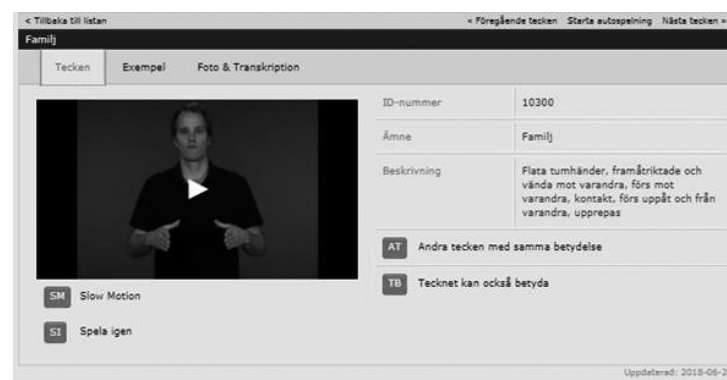
Dictionary design

The main lexical database, the SSL Dictionary was created in 2008 and its interface has been in development

Session 2

since. The website for Swedish Sign Language Dictionary is <https://teckensprakslexikon.su.se/> Technical support aimed to develop the dictionary to become more user-friendly and usable on different platforms such in as apps for iPhone and Android. In the SSL Dictionary, a sign is described in terms of its handshape, orientation, location, movement, Swedish translations and sign transcription. Each dictionary entry is represented by a video of the sign (or phrase), a Swedish translation, phonological information, and internal cross-links to phonologically or semantically equivalent signs – i.e. homophones and synonyms, see Figures 1-3 (Svenskt teckensprakslexikon, 2018). ID-glosses are added for supporting SSL corpus annotators.

A sign can be searched for using these ways (see Figure 4) or by subject (46 with subcategories), e.g. family, Swedish towns and provinces, healthcare or fingerspelled signs. There are also ID numbers for signs (in general and special terms). However, ID glosses alone cannot describe sign language perfectly. There are also descriptions and translations for each sign. The old signs with 415 sign entries from the first printed dictionary in SSL (Österberg, 1916) are also added to this lexical database. <https://teckensprakslexikon.su.se/kategori/osterberg-1916> Frequency refers to how often visitors return to the dictionary site within a time frame.

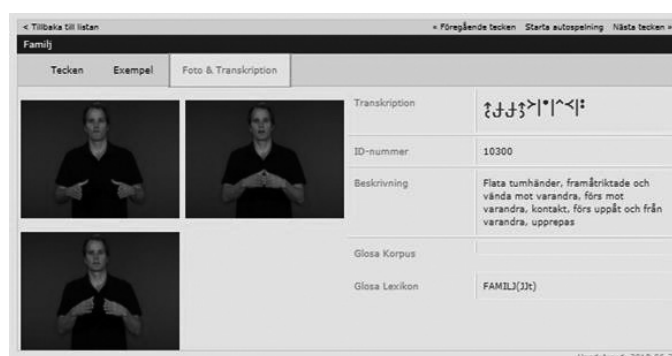


[Figure 1.] Showing a sign with options for slow motion, play again, other signs with same meaning, this sign has also other meaning



[Figure 2.] Showing some sentence examples

Challenges of creating a Sign Dictionary | Johanna Mesch



[Figure 3.] Showing still photos (for printing) with sign transcription for SSL, ID gloss, sign description, gloss for corpus annotation, gloss for dictionary



[Figure 4.] Searching via handshape, orientation, location and/or movement

The SSL Dictionary is used as standardized learning resource for L2 teaching and self-learning. The use of the dictionary for learning purposes is important for learners, families and educational settings. The Deaf community has also found it valuable thanks to a Facebook group. A Quiz has also been created for learners, <https://tspquiz.se/app/#/start>. A twitter bot “Ett tecken varje dag” (One sign every day) was created in January 2018, and it has been well-liked by learners [<https://twitter.com/allatecken>]. It publishes a randomly selected sign from the SSL Dictionary, see Figure 5.



[Figure 5.] Twitter bot One sign every day

Session 2

Language resources supporting for the sign language dictionary – Swedish Sign Language Corpus

There are currently two main language resources of SSL publicly available, a dictionary and a corpus (Mesch, 2012; Mesch & Wallin, 2012, 2015; Mesch, Wallin, & Björkstrand, 2012; Mesch, Wallin, Nilsson, & Bergman, 2012), both of which also serve as a form of language documentation for SSL. The first corpus of Swedish Sign Language was created in 2004 in connection to the ECHO project (O. Crasborn et al., 2007).

The resource is the Swedish Sign Language Corpus (SSLC), which consists of 24 hours of video data (conversations, narratives, and presentations) from 42 different signers collected in the years 2009–2011, financed by Riksbankens Jubileumsbank (Mesch, Wallin, Nilsson, et al., 2012; Wallin & Mesch, 2018). Corpus work, especially annotation work, is very time-consuming. Approximately half of the collected 24 hours of data have been annotated (September 2018). Other corpora in SSL are the Tactile Sign Language Corpus (Mesch, 2016) and the Learner Corpus in SSL (Schönström & Mesch, 2017). The Tactile SSL Corpus consists of dialogues and elicitation from 8 signers collected in 2007 and 2011. Approximately 40 % (6,000 sign tokens) of the collected four hours and thirty minutes have been annotated. Annotation work of the SSL corpus and other corpora is still ongoing with ten (eight deaf and two hearing) annotators working on it part-time.

Although both the SSL Dictionary and the SSL Corpus are designed and maintained by the Sign Language Section at the Department of Linguistics, Stockholm University, they were initially set up independently from each other. This has led to some discrepancies between the functions of the two resources. However, the two projects have started to converge in the last year, and now they are already supporting each other. For example, sign tokens in the SSLC are annotated with the SSLD ID number (Riemer Kankkonen, Björkstrand, Mesch, & Börstell, 2018), signs are added to the SSLD as they appear in the SSLC, and the SSLD online interface sorts synonyms search hits according to SSLC frequencies calculated through the SSLC lexical frequency interface (Börstell & Östling, 2016). This frequency tool for the SSLC data shows the relative frequencies of two competing sign variants, e.g. the signs TIO and TIO(Y) 'ten'. It will be a good resource when the SSLC has been expanding.

<https://www.ling.su.se/teckenspr%C3%A5ksresurser/verktyg/frekvens>

The SSLC contains currently around 113,000 sign tokens, and, as there was no explicit lexical variation elicitation task during the collection of the corpus data (Stamp et al., 2014), many sign synonyms or form variations cannot yet be investigated solely with the use of corpus data. A set of 113,000 sign tokens is fairly large for a sign language corpus, although it remains much smaller when compared to spoken language corpora used to investigate variation systematically.

Crowdsourcing - Improving language resources

To develop and improve the dictionary, specifically the SSLD, we involve the Deaf community to collect new signs and to gather data on familiarity with and variation within sign synonyms – that is, crowdsourcing to improve our language resources, specifically the SSLD. This works in two ways: a) discussing sign usage, lexical variation, and sign formation with SSL signers on social media; and b) interacting with SSL signers directly at various venues (the annual participation in Deaf events of national Deaf organizations and Deaf community business and activities) collecting signs and judgments about signs (Riemer Kankkonen et al., 2018).

A Facebook group entitled Teckenspråkslexikon ('sign language dictionary') was started in October 2014 and is administrated by the dictionary team. There are 2,642 members (in May 2018), who interact in different ways, discussing which signs are to be used, or the difference between sign variants. They can write questions or comments in Swedish or in SSL as video comments. In the last two years (i.e., January 2016 to January 2018), there have been 593 posts and 5,817 interactions. This includes questions about signs, such as How do you sign X? (53%), Is there a sign for person Y? (20%), general questions about the usage (10%) and etymology (<0,2%) of specific signs.

Conclusions

Developing and maintaining the sign language dictionary is time-consuming and resource-intensive. It is necessary to keep up with technical development, for example by updating video files, and protecting against crashes and hackers. The dictionary also needs research and technical support to maintain a good quality and user-friendly resource. The methods for updating and developing the sign language dictionary include language knowledge, other sources such as old video stories, TV programmes and video messages, and crowdsourcing data about the national sign language(s) directly from the Deaf community using social media, e.g. Facebook group. The corpus size for signed languages is still small, so form variations cannot yet be investigated solely with the use of corpus data. Crowdsourcing for the dictionary is a useful and rapid method for enhancing the language resource. Data collected through offline methods may provide a more qualitative approach to sign variants, interacting with individuals directly – aside from giving the dictionary team an opportunity to personally meet and interact with the Deaf community, which is an important aspect of any language documentation work. Metadata about each signer target specific individuals or groups that are needed to add to the documentation work, for example, from the corpus. The corpus-based Signbank is a necessary tool for researchers considering sign lemmatization. When changes or additions of new sign glosses, for example, have been made to the Signbank, they are automatically updated through linked ELAN annotation files, which makes annotators' and researchers' work easier. Early dictionaries or current lexical databases, e.g. the SSL lexical database, are good enough sources, if they are used together with the corpus projects for lemmatization of a corpus. How to put a

Session 2

headword on a sign dictionary is still not clear when thinking about compounds from spoken/written language (very common in subject areas such as medicine, technical or vocational words), and the high frequency part of productive signs (depicting or classifier signs) should be added to the dictionary.

References

- Börstell, C., & Östling, R. (2016). Visualizing lexis in a sign language corpus : Mining lexical variation data in lexis of Swedish Sign Language. In E. Efthimiou, S. E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Workshop Proceedings: 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining* (pp. 13–18). Paris: ELRA.
- Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., Van Der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, 12(4).
- McKee, R., & Vale, M. (2017). Sign language lexicography. In P. Hanks & G.-M. de Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–22). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-45369-4_34-1
- Mesch, J. (2012). Swedish Sign Language Corpus. *Deaf Studies Digital Journal*, 3. Retrieved from http://dsdj.gallaudet.edu/index.php?issue=4§ion_id=2&entry_id=128
- Mesch, J. (2016). Dataset. Tactile Swedish Sign Language Corpus. Department of Linguistics, Stockholm University.
- Mesch, J., & Wallin, L. (2012). From meaning to signs and back: Lexicography and the Swedish Sign Language Corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 123–126). Paris: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/workshops/24.Proceedings_SignLanguage.pdf
- Mesch, J., & Wallin, L. (2015). Gloss annotations in the Swedish Sign Language Corpus. *International Journal of Corpus Linguistics*, 20(1), 103–121. <http://doi.org/10.1075/ijcl.20.1.05mes>
- Mesch, J., Wallin, L., & Björkstrand, T. (2012). Sign language resources in Sweden: Dictionary and corpus. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]* (pp. 127–130). Paris: European Language Resources Association (ELRA). Retrieved from <http://www.lrec-conf.org/proceedings/lrec2012/>

workshops/24.Proceedings_SignLanguage.pdf

Mesch, J., Wallin, L., Nilsson, A.-L., & Bergman, B. (2012). Dataset. Swedish Sign Language Corpus project 2009--2011 (version 1). Sign Language Section, Department of Linguistics, Stockholm University. Retrieved from <http://www.ling.su.se/teckensprakskorpus>

Österberg, O. (1916). Teckenspråket. Uppsala: P. Alfr. Persons förlag.

Riemer Kankkonen, N., Björkstrand, T., Mesch, J., & Börstell, C. (2018). Crowdsourcing for the Swedish Sign Language Dictionary. In M. Bono, E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, ... Y. Osugi (Eds.), Proceedings of the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community [Language Resources and Evaluation Conference (LREC)] (pp. 171–174). Paris: European Language Resources Association (ELRA).

Schönström, K., & Mesch, J. (2017). Dataset. The project From speech to sign – learning Swedish Sign Language as a second language. Department of Linguistics, Stockholm University.

Svenskt teckenspråkslexikon. (1997). Leksand: Sveriges Dövas Riksförbund.

Wallin, L., & Mesch, J. (2018). Annoteringskonventioner för teckenspråkstexter. Version 6, januari 2018. [Annotation guidelines for sign language texts].

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전



Records of Visible Language:
Sign Language Dictionary



주제 4 | 대한민국 수어사전의 현황과 발전 방향

최혜원, 이현화
한국, 국립국어원

Hyewon Choi, Hyunhwa Lee
National Institute of Korean Language, Korea

2부 발표

대한민국 수어사전의 현황과 발전 방향

최혜원 · 이현화
한국, 국립국어원

1. 머리말

〈한국수어언어법〉(2016)이 시행된 지 2년여의 시간이 흘렀다. 한국 사회에서 한국수어가 언어로 인정되고 한국수어의 사용 권리를 보장받기 위한 농인의 오랜 염원을 담은 〈한국수어언어법〉은 한국수어의 발전을 위한 다양한 제도와 사업 활성화의 근거가 되었다. 2016년 〈한국수어언어법〉이 제정된 이래 대한민국은 한국수어의 언어로서의 발전을 위해 다양한 연구와 사업을 시행하고 있다. 5년마다 한국수어발전기본계획을 체계적으로 수립하고, 한국수어 사용 환경을 개선하고자 한국수어의 사용 실태 조사, 한국수어 말뭉치 구축, 문법에 관한 기초 연구 등을 본격적으로 가동했다. 또한 한국수어의 교육 및 보급을 위해 한국수어교원자격제도를 시행하면서 대상별 수어 교재를 개발하고 한국수어문화학교 프로그램을 운영하고 있다. 이와 함께 기존의 《한국수어사전》을 정비하는 과정에서 한 언어의 총체를 담은 한국수어사전으로의 도약을 위한 발전적인 방향을 모색하고 있다.

수어사전의 편찬은 수어의 독립적 지위를 알리기 위한 가장 기초적인 작업이다. 음성 언어의 종속적인 언어, 보조적인 의사소통 수단이라는 통념을 딛고 언어로서의 수어를 기술한 수어사전은 농인들이 독립된 언어 공동체를 이루고 있다는 증거가 된다. 실제로도 여러 나라에서 언어 사전으로서의 수어사전을 편찬하여 농인 공동체의 언어적, 문화적 지위 향상의 계기를 마련하였다. 이때, 수어사전의 형태는 그간의 음성 언어 중심의 관점으로 음성 언어에 대응하는 수어를 보여 주는 단순한 대역어 형태의 사전을 뛰어넘어야 함은 물론이다.

조선이 일본에 주권을 빼앗긴 일제 강점기 시대, 국어학자들을 중심으로 한글로 된 최초의 조선어 사전을 편찬하려 하였으나 결국 세상에 먼저 나오게 된 것은 조선총독부가 편찬한 《조선어사전》(1920)이었다. 사전 명칭과는 달리 이 사전은 조선어를 대상화한 일본어 중심의 사전일 뿐이었다.¹⁾ 조선어가 독립 국가의 언어가 아닌 상태에서 편찬된 사전은 일본어 표제어와 이에 대응하는 한글 표기, 그리고 일본어 뜻풀이

1) 이후 문세영이라는 국어학자의 치열한 노력으로 《조선어사전》(1938)이라는 한글로 표기된 한국어 사전이 만들어져 우리 민족에게 큰 희망을 주었다.

로 구성되어, ‘조선인의 일본어 교육, 일본인의 조선어 교육’이라는 식민지 지배 정책에 활용되었다(최경봉, 2014).

위의 사례와 마찬가지로 대한민국에서 그간 만들어진 한국수어사전들도 진정한 의미의 언어 사전으로서 틀을 갖추지 못하였다. 한국수어에 대한 이해가 깊지 않은 청인 교육자들이 한국어 텍스트로 된 교과목을 가르치기 위한 수단으로 만든 대역어 형태의 수어사전은 그 태생적 한계를 극복하지 못한 채 현재까지 답습되고 있는 실정이다. 농교육을 목적으로 한 사전의 출발점은 서구²⁾에서도 동일하지만 90년대 전후 유럽을 중심으로 언어학적 접근을 통해 수어의 자연 언어로서의 체계를 담은 진일보한 수어사전을 편찬하고 있다. 우리나라는 <한국수어언어법>을 통해 한국수어를 언어로 선언하였지만, 한국 사회에서 명실공히 언어로서 인정되기 위해서는 선언을 뛰어넘어 한국수어의 언어로서의 체계를 밝히 보이는 사전이 필요하다. 수어의 무한한 가능성을 보이고 한국 농사회의 정체성과 자부심을 뒷받침할 수 있는 사전 편찬을 위해 이번 학술행사는 매우 의미 깊은 역할을 할 것이라 기대한다.

이 글에서는 그간 한국에서 편찬된 수어사전들의 현황과 문제점을 살펴보고 향후 농인이 실제 사용하는 한국수어 자료를 충실히 담은 새로운 한국수어사전 편찬을 위해 국립국어원에서 수립한 중장기 계획을 밝히도록 하겠다.

2. 한국수어사전의 편찬 현황

수어사전이 없던 시기에 수어 어휘 목록은 주로 수어 보급을 위해 만든 수어 교재를 통해 볼 수 있었다. 1982년에야 한국에서는 처음으로 수어사전이 편찬되었는데 《표준수어사전》이라는 사전의 명칭에서도 확인할 수 있듯이 이 사전은 ‘수어 표준화’³⁾에 기여하기 위한 목적에서 출발하였다(이준우, 남기현, 조준모, 2018). 이 목적에 따라 농학교 교사들을 중심으로 구성된 ‘한국표준수어사전편찬위원회’는 일상생활 수어 어휘, 초·중등 국어 교과서의 국어 어휘에 대응하는 수어 어휘를 사전에 담아냈다.

아래의 표에서 보듯이 한국에서는 그간《표준수어사전》(1982)을 시작으로 총 6종의 수어사전이 편찬되었다.⁴⁾

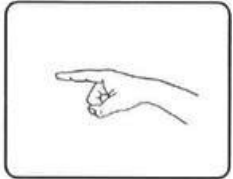




연번	사전/편찬기관(편찬자)	출판 연도	표제어	미시 구조
1	《표준수어사전》 표준수어사전 편찬위원회	1982	5,494개	

2) 1989년 호주 수어사전에서도 영어의 문법 형태(과거형 분사인 ‘-ed’)에 대응시킨 수어 표현의 흔적이 남아 있다.

3) 여기서 말하는 ‘표준화’는 일반적인 의미의 언어 표준화와는 다르다. 광범위한 수어 자료의 수집 과정이 없었기 때문에 지역적 변이형 등 여러 형태로 나타나는 어형 가운데 가장 일반적이면서 널리 통용되는 어휘를 대표형으로 선정하는 방식은 아니었다. 한국어 대응어가 없는 것은 새로 만들어 내고, 동형이의어는 변별을 위해 보조 사인을 첨가하는 등 음성 언어의 관점에서 수어를 ‘규격화’하는 작업이 대거 포함되었다.

4) 어원 사전, 전문용어 사전 등은 목록에서 제외하였다.

2부 발표





연번	사전/편찬기관(편찬자)	출판 연도	표제어	미시 구조
2	《수어사전》 단국대학교 출판부	1983	5,966개	구분(區分) (區 區分) = "구별". 구석(모퉁이의 안쪽) (區 區分) 두 손의 손가락 끝을 맞대어 그 끝이 밖을 향하게 하였다가 오른손만 빼면 다음 오른손의 1지표 왼손의 손가락 끝을 가리킨다. 구:세-주(救世主) (區 區分) = "구원" + "주인".
3	《표준수어》 김승국	1993	5,966개	가리키다 (區 區分) 오른 주먹의 1지를 펴서 그 끝으로 대상을 가리키는 자를 말한다. = "지목".  가리 (가리키다) (區 區分) = "가리키다".
4	《한국수어사전》 국립국어원 · 한국농아인협회	2005	6,800여 개	출발(出發) (區 區分) 목적지를 향하여 나아가는 행동을 시각에 맞춰 나타낸다. [표준어: 여는 동작] = 시작점, 착수.  손끝이 앞으로 손등이 앞으로 향하게 해서 두 손의 손바닥을 앞뒤로 맞붙여 나타낸다.
5	《한국수어사전》 김삼찬 외	2005	9,818개	—도다 발달(發達) (區 區分) 용인의 머리에 붙어 '갈탄'의 뜻을 나타내는 종언이며 예: 끝: 머무르는 일만 나타내준다. 종언어: —구나, —구 발달(發達) (區 區分) 1자관 및 조관손을 열 머리에서 아래로 내린다.  도달(到達) (區 區分) arrival 발달(發達) (區 區分) 목적지나 도달한 수준에 다다름 예: 끝: 도착한 것에 도달하면 요령하다. 종언어: 상하 발달(發達) (區 區分) 상이 오른쪽으로 향하게 한 왼손바닥에 한 오른 손끝을 얹어 낸다.
6	《한국수어사전》 국립국어원 (온라인)	2015	12,793개	한국어 정보 영향 (影響) (區 區分) [영향] 어떤 사물의 효과나 작용이 다른 것에 미치는 일. 영향력적 영향. 수령 (影響) (區 區分) 5지를 펴서 오른 주먹의 밑에서 오른 주먹을 손 바닥이 위로 향하게 하면서 앞으로 내린다. 앞어 정보(영향 그림) [영향(區 區分)]이 미치는 영향을 나타내는 상징적 동역 동형어 = 파급  영향 (影響) (區 區分) [영향] 어떤 사물의 효과나 작용이 다른 것에 미치는 일. 영향력적 영향. 수령 (影響) (區 區分) 5지를 펴서 오른 주먹의 밑에서 오른 주먹을 손 바닥이 위로 향하게 하면서 앞으로 내린다. 앞어 정보(영향 그림) [영향(區 區分)]이 미치는 영향을 나타내는 상징적 동역 동형어 = 파급 

〈표 1〉 국내에서 편찬된 한국수어사전

위에서 보듯이 한글 자모 가나다순으로 배열된 6개 사전은 한국어 표제어와 이에 대응하는 수어를 담은 대역어 사전의 형태이다. 대역어조차도 실제 사용되는 수어만 담은 것이 아니라 한국어 어휘를 표현하는 데에 필요하다는 이유로 수어 표현에는 없는 것을 새로 만들어 올린 경우가 적지 않았다. 이러한 결과로 과도한 수어 표제어가 산출되었고, 그중에는 한국어 어미와 조사, 합성어에 대응하는 다수의 인위적인 수어가 포함되었다.⁵⁾

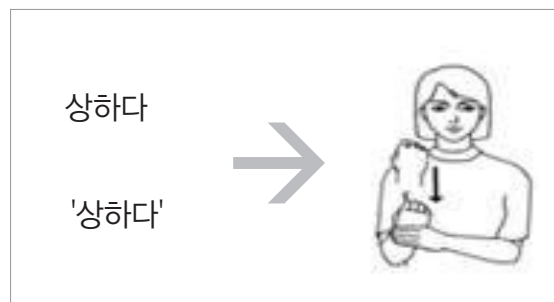
일반적으로 대역어 사전은 A 언어 표제어의 여러 의미를 B 언어의 어휘나 대체 표현으로 다양하게 제시한다. 예를 들어 표제어가 한국어 ‘상하다’일 경우 ‘상하다’의 여러 의미와 이에 대응하는 한국수어 대역어 정보를 제시해야 한다.

5) 1998년 네덜란드 정부는 네덜란드 수어의 법적 지위를 인정하기 위한 선결 과제로 네덜란드 기본 어휘 표준화(STABOL, Standardisation of the Basic Lexicon of NGT)를 내세웠다. 농사회에 수어 연구자들의 거센 반대에 부딪히기도 한 이 프로젝트에서 표준화된 어휘 수는 약 14,000개를 기록한다. 이 중 33%가 의료나 법률, 학교 교과서 용어 등 새로운 어휘이다.(Roland Plau, Markus Steinbach & Bencie Woll, 2012)

한국어 표제어	의미/한국수어 대역어			
	'다치다'	'아위다'	'망가지다'	'상하다'
상하다				

〈표 2〉 한국어 '상하다'에 대응되는 여러 개의 수어 표현

그러나 《한국수어사전》은 한국어 표제어의 대응 수어를 하나로 국한하여 한국어 표제어의 뜻을 한국수어 단어 하나로 제시하거나(1, 5), 한국어 표제어의 사전 뜻을 거의 단일하게 압축하여 보이고 있다(4, 5, 6). 이는 사전 이용자에게 한국어에 하나의 수어가 대응되고, 한국어와 한국수어의 의미 갈래가 일치한다는 한국수어에 대한 잘못된 이해를 불러일으킬 수 있다.



〈그림 1〉 한국어 표제어 '상하다'와 글로스가 일치하는 한국수어의 제시

대역어 사전의 틀에서 기술되면서 사전에서 다루는 수어 정보는 수어의 형태적 특징에 한정될 뿐 그 외 수어의 언어적 특징은 기술되지 못하였다. 이마저도 한국어에 한국수어를 무리하게 연결시키다 보니 한국수어가 왜곡되는 현상을 초래하였다. 예를 들어 5의 '목적지나 일정한 수준에 다다름.'의 뜻을 지니는 '도달'에 대응하는 수어의 동작을 보자. 수향이 수직 방향으로 위로 향해 있으나 '목적지'는 반드시 높은 곳에 있지 않을 수 있다. '목적지'의 위치 특성에 따라 '도달'을 표현하기 위해 수어 화자는 수향의 변화를 다양하게 주는데 사전에서는 이를 반영하지 못하고 있다. 또한 구체화된 표현이 많은 한국수어의 특성상 한국어 표제어의 뜻풀이와 수어가 대응되더라도 한국어 용례와는 불일치하는 경우가 발견된다. 한국어 '가장자리(둘레나 끝에 해당되는 부분)'에 대응되는 수어는 한국어의 일반적인 뜻에서 크게 벗어나지 않지만 “눈 가장자리가 빨갛게 되었다.”와 같은 용례를 표현할 때는 부적절한 표현이 된다.

2부 발표



〈그림 2〉 표제어 '가장자리'에 대응되는 수어 표현

한국수어사전이 처음 만들어진 시기나 사전의 총수, 표제어 수는 다른 어느 나라에 비교해 봐도 크게 뒤지지 않는다.⁶⁾ 그러나 사전 콘텐츠의 충실성은 이처럼 별다른 개선 없이 답보 상태에 머물고 있다.⁷⁾ 이는 2005년 《한국수어사전》까지 진행된 사전 편찬 작업이 '수어 표준화'의 기본 취지에서 크게 벗어나지 못하였기 때문이다. 이러한 한계는 2005년 《한국수어사전》의 형태를 그대로 반영한, 현재까지 서비스되고 있는 온라인 형태의 사전인 국립국어원의 《한국수어사전》⁸⁾(2015)에서도 마찬가지이다.

3. 한국수어사전의 발전 방향

앞서 살펴본 바대로 한국수어 표준화 사업의 결과물인 《한국수어사전》은 한국수어 자료를 모아 연구하고 수어 연구 방법론에 대한 환기를 시켰다는 점에서 의의를 찾을 수 있다. 그러나 구축 당시의 시대적 한계로 인하여 이 사전은 사용자들에게 지속적인 비판을 받아 왔다. 이에 국립국어원은 그간의 비판을 겸허히 수용하고 기존 사전의 한계를 극복한 새로운 한국수어사전의 편찬을 계획하고 있다.

효과적인 사전 설계는 사용자의 목적과 능력의 현실적인 이해에서부터 시작된다고 한 Atkins & Rundell(2008)의 주장처럼 국립국어원은 《한국수어사전》 개선 방향에 대한 올바른 설계를 위하여 여러 사업을 통해 농사회의 의견을 수렴하였다.⁹⁾ 2014년에 구성된 '한국수어 연구 자문 위원회'에서는 전국의 한국수어 사용자 553명을 대상으로 '수어 전문 용어 제정 및 표준화 수요 조사'를 진행하였다. 이 조사에 따르면 참여자들은 수어 표준화가 추진된 전문 분야 수어를 사용하지 않는 이유로 '실제 사용되는 수어와 다름(28.1%)'에 제일 많이 응답하였다.

'수형 기반 한국수어사전 구축'(국립국어원, 2016) 사업에서는 농인 15명을 대상으로 초점 집단 인터뷰

6) 초기의 뉴질랜드 수어사전(The Dictionary of New Zealand Sign Language, 1986)은 1,200개, 호주 수어사전(Auslan Dictionary, 1989)은 3,500개, 영국 수어사전(Dictionary of British Sign Language, 1992)은 1,800개의 표제어가 수록되어 있다.

7) 1960년대 스토킨을 중심으로 한 갤러티 대학 연구자들의 미국 수어에 대한 언어학적 연구 방법론은 1970년대 중반 영국과 유럽 전역의 연구자들에게 영향을 주었고, 1980년대 이후 수어는 전 세계적인 연구 분야의 중심에 섰다.

8) 《한국수어사전》(2015)에는 《한국수어사전》(2005)의 어휘 목록과 함께 이후 동일한 사전 편찬 방향하에 연차별 작업으로 추가된 6천여 개의 어휘가 포함되어 있다.

9) 이 글에서는 조사 결과 중에서 기술에 필요한 의견 일부를 선택적으로 제시하였다.

를 실시하였는데 그 결과 기존 사전이 ‘농사회에서 자주 사용되는 수어가 부재’한 상태이며 ‘한국어 대응 수어 표현이 중심이 되다 보니 농인들로부터 외면’받고 있다고 하였다. 그리고 이 사전이 ‘수어통역사들의 교육용으로 활용’되고 ‘청인들의 수어 학습용으로 활용’되고 있지만 ‘농인조차 이해하기 어려운 수어사전’이라 ‘농인과 청인 수어통역사 간 소통의 문제가 발생’한다고 하였다. 농인들은 현 사전을 ‘필요할 때 이용하지만 큰 도움을 주지 않는 수어사전’이며 ‘농인이 이해하기 어려운 표현의 한국어 정보’로 한국어를 이해하기 위해 사전을 이용하는 것에도 제한이 있는 것으로 평가하고 있었다.

이어 2017년 진행된 ‘한국수어사전 수어 뜻풀이와 용례 구축 및 정비’(국립국어원) 사업을 통해 수어학계와 농사회의 수어 관련 전문가 30인을 대상으로 설문 조사를 실시하였다. 조사 결과 《한국수어사전》의 단점으로 거시 구조 측면에서는 낮은 접근성과 인위적으로 만들어진 수어가 등재되어 있다는 점이 주로 지적되었고, 미시 구조 측면에서는 정보 항목 간의 낮은 일치도와 수어를 활용하기 위해 반드시 필요한 용례, 문법 정보 등의 부재가 지적되었다.

더불어 2016년과 2017년에 각각 진행된 두 조사에서는 《한국수어사전》의 발전 방향에 대해 제언을 받기도 하였는데 주요 내용은 다음과 같다.

- 수행 기반 한국수어사전 구축(국립국어원, 2016)
 - 농인 사용자 중심의 새로운 사전 제작
 - 한국어 정보를 농인의 관용적 수어 표현으로¹⁰⁾ 쉽게 풀이해서 설명
 - 수어사전의 목적과 이용 대상에 대한 명확한 설정 필요
 - 접근성 향상을 위한 방안
- 한국수어사전 수어 뜻풀이와 용례 구축 및 정비(국립국어원, 2017)
 - 농인과 청인이 모두 사용할 수 있는 사전
 - 현 사전에서 인위적인 표제어의 삭제 및 정비
 - 농인이 실제로 사용하는 수어를 수록하고 수어 관련 정보를 충실히 담은 사전
 - 수어의 다양한 변이형을 담은 사전
 - 사전의 접근성을 높이기 위한 표제어 분류 및 검색 방법(움직임 인식 등) 개발

이상의 결과로 볼 때 현재 《한국수어사전》의 이용자는 한국수어 학습자와 한국어 학습자로 나뉘며 각 그룹은 다른 목적으로 사전을 이용하고 있다. 또한 이 사전이 한국수어 표현을 인위적으로 만들어 사전에 등재하였으나 이것이 농사회에 받아들여지지 못해 사전으로서의 역할을 하기에 부족한 상황임을 알 수 있다. 따라서 이러한 의견을 바탕으로 《한국수어사전》을 정비하여 사전의 일치도와 완성도를 높이는 한편 이용자들이 목적에 따라 사전을 이용할 수 있도록 새로운 사전을 편찬할 필요가 있다.

10) Cokely, D. & Baker-Shenk, K. C. (1980)는 미국수어(ASL)의 설명에서 ‘관용어(idioms)’로 표현된 것들의 대부분은 사실 관용어가 아니고 관용어 정의에 맞지 않는다고 하며 단순히 몇 개의 다른 용법을 가진 수어들이라고 하였다. 이정옥, 이준우(2005)도 한국의 이와 같은 현상을 정리하며 ‘자연 수어’ 혹은 ‘농식 수어’라는 명칭으로 관용 표현을 받아들이고 있음을 지적하고 있다. 따라서 본고에서 인용한 조사의 결과도 같은 맥락으로 해석할 수 있다.

2부 발표

Hartmann(1983)은 외국어를 학습하고 교육하는 보조 도구로서 이언어 사전이 유용하게 사용될 수 있다고 하였다. 또한 한국수어 기술 사전 편찬의 필요성에 대해 논의한 이현화(2017)는 한국수어사전이 인위적으로 가공된 자료가 아닌 한국 농인의 자연스러운 발화를 담고 있는 실제 자료에 근거하여 표제어를 추출하고 이를 바탕으로 편찬되어야 한다고 하였다. 이러한 논의들을 바탕으로 볼 때 한국수어사전의 편찬 방향은 다음 두 가지로 귀결된다.

첫째, 사전은 사용 목적을 분명히 하여 ‘한국수어-한국어 사전’, ‘한국어-한국수어사전’으로 이원화 구축되어야 한다.

둘째, 변이가 많은 수어의 특성을 담은 사전은 기술 언어학에 근거하여야 하며 따라서 한국수어의 실증적인 자료를 모은 한국수어 말뭉치를 바탕으로 한 사전 편찬이 이루어져야 한다.

3.1. 양방향 이중 언어 사전 중장기 구축 계획

《한국수어사전》의 사용자와 전문가 집단의 의견과 제언 등을 종합적으로 고려하여 국립국어원이 수립한 새로운 한국수어사전의 중장기 구축 계획은 다음과 같다.

3.1.1. 양방향 이중 언어 사전의 구조

(1) 거시 구조

- 한국수어-한국어 사전
 - 표제어의 선정
 - 한국수어 말뭉치(‘15~)와 추가로 구축할 사전용 말뭉치(예정)를 바탕으로 표제어 선정
 - 한국수어의 고정된 어휘부(frozen lexicon)를 중심으로 생활용어, 지역어, 신어 등 수록
 - 표제어 표기 방법
 - 한국수어의 형태를 기록할 수 있는 수어 문자(SignWriting)로 수어를 표기하여 한국수어가 한국어와 다른 독자적인 언어라는 인식 제고
 - 글로스(gloss) 사용으로 인해서 발생하는 여러 문제점 개선
 - 표제어의 배열
 - ‘수형 기반 한국수어사전 구축(국립국어원, 2016)’의 수어소 분류를 참조하여 배열
- 한국어-한국수어사전
 - 표제어의 선정
 - 한국 농인의 한국어 학습에 필요한 표제어 10,000개 평정
 - 한국수어-한국어 사전에 한국어 대응 표현으로 수록된 한국어 표제어를 중심으로 수록
 - 표제어 표기 방법
 - 표제어는 한글로 표기하되 한국어 표제어에 대응하는 한국수어 표현은 썸네일(Thumbnail)로 제시

- 표제어의 배열
- 가나다순

(2) 미시 구조

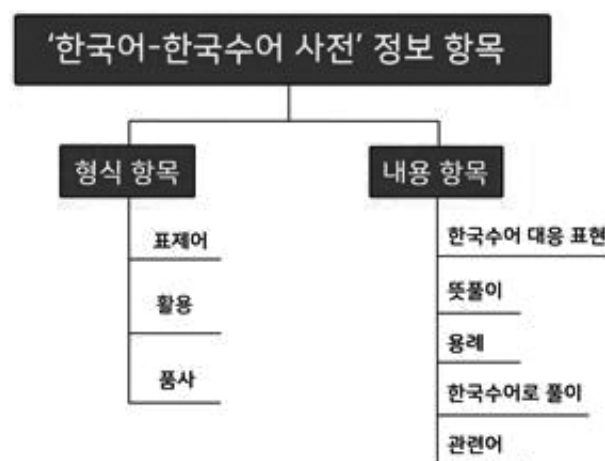
□ 한국수어-한국어 사전

○ 형식 항목

- (표제어) 기본형과 변이형 모두 확인할 수 있도록 영상으로 제시
- (마우스 제스처) 수어를 할 때 필수적으로 나타나는 마우스 제스처를 그림으로 제시
- (수어 문자) 표제어의 형태를 수어 문자로 기록
- (수형 그림) 수어소 검색 시 활용할 수 있도록 수형 그림 제시

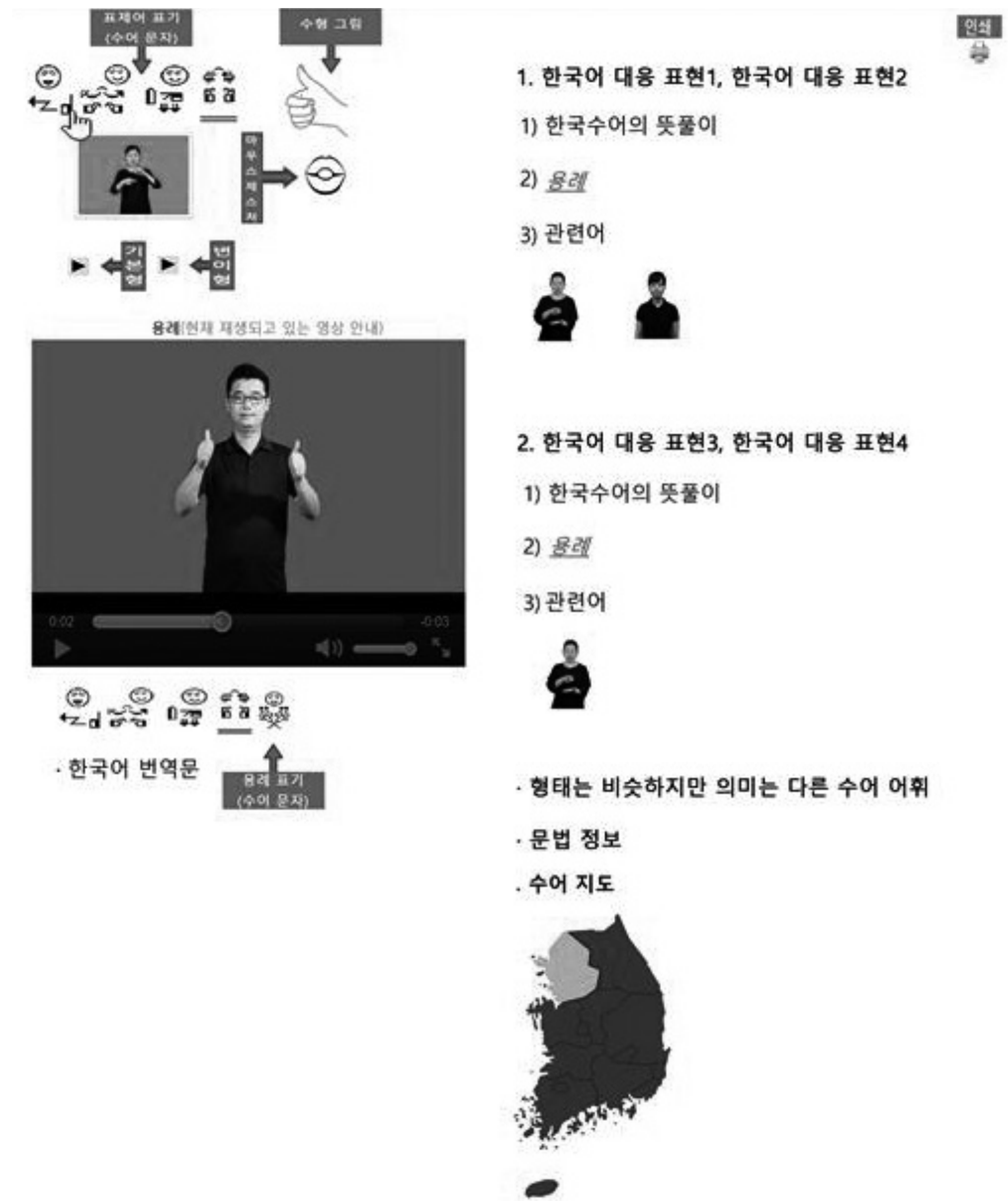
○ 내용 항목

- (대응 표현) 한국수어에 대응하는 복수의 한국어 표현을 제시
- (뜻풀이) 한국수어 말뭉치를 바탕으로 추출한 한국수어의 의미를 한국어로 쉽게 이해할 수 있도록 기술
- (용례) 표제어의 다양한 활용 예를 수어 동영상으로 제시
 - 제시된 용례를 수어 문자로 표기하고 이를 클릭할 경우 해당 표제어로 이동할 수 있도록 연동 기능 구현
- 정확하고 자연스러운 한국어로 번역
- (관련어) 표제어와 일정한 의미 관계를 맺는 한국수어를 제시
- (형태는 비슷하지만 의미가 다른 수어) 한국수어 학습자들이 빈번하게 나타내는 학습 오류 유형으로, 수어 학습의 효과를 높이기 위해 관련 정보 제공
- (문법 정보) 수어의 복수형, 부정형, 활용형 등 문법 정보 제시
- (수어 지도) 지역 변이가 매우 발달한 수어의 특성에 따라 수어 지도를 통해 지역어 정보 제공



〈그림 3〉 한국수어-한국어 사전 정보 항목

2부 발표



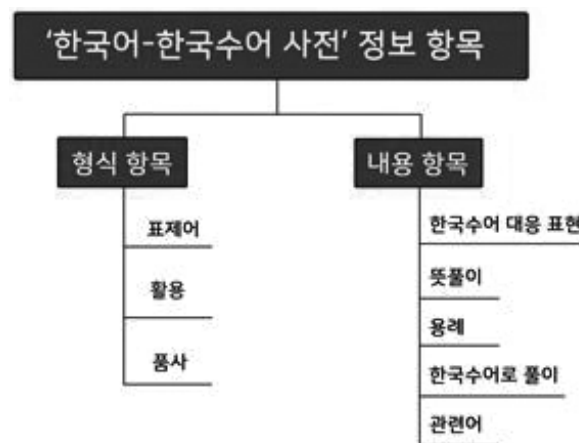
〈그림 4〉 한국수어-한국어 사전 시범 화면

□ 한국어-한국수어사전

○ 형식 항목

- (표제어) 1만 개의 규모로 표제어 제시
- (활용) 표제어의 활용형을 제시하여 한국어 학습 효과 제고
- (품사) 한국어 표제어의 품사 제시

- 내용 항목
 - (대응 표현) 표제어에 대응하는 복수의 한국수어 표현을 썸네일(Thumbnail)로 제시
 - (뜻풀이) 한국어 기초사전의 뜻풀이 제시
 - (용례) 한국어 기초사전의 용례 제시
 - (한국수어로 풀이) 번역이 아닌 농인의 이해를 기반으로 하여 한국어 표제어의 뜻풀이와 용례를 하나의 영상 안에서 설명
 - (관련어) 한국어 표제어와 유의, 반의 관계를 맺고 있는 단어를 관련어로 제시



〈그림 3〉 한국수어-한국어 사전 정보 항목

이익(利益)

[활용] 이익이, 이익도, 이익만

1. 물질적으로나 정신적으로 보탬이나 도움이 되는 것(한국어 표제어의 뜻풀이)

한국어 대응 표현

- 눈 앞의 이익(한국어 표제어의 용례)
- 작은 이익
- 이익이 된다.
- 이익이 없다.
- 그는 더 큰 일을 이루기 위해서 눈 앞의 작은 이익들을 희생해야 했다.

현재 재생되고 있는 영상 안내

반대말 손해

2. 일정 기간 벌어들인 돈에서 그 돈을 버는 데 쓴 비용을 뺀 나머지 금액

한국어 대응 표현

- 보장한 이익
- 큰 이익
- 이익이 남다.
- 이익이 높다.
- 모든 기업은 이익을 최대화하기 위하여 생산 비용을 최소화하려고 노력한다.

현재 재생되고 있는 영상 안내

유의어 이윤

〈그림 6〉 한국어-한국수어사전 시범 화면

2부 발표

3.1.2. 양방향 이중 언어 사전 구축 계획

구분	1단계('20~'22)	2단계('23~'25)	3단계('26~'28)	4단계('29~'31)
한국 수어 자료 수집	<ul style="list-style-type: none"> 한국수어 자료 수집 및 가공 한국수어 말뭉치 구축 <ul style="list-style-type: none"> 서울 외 지역 말뭉치 구축 기본 전사 의미 전사 사전용 말뭉치 기초 연구 및 자료 구축 	<ul style="list-style-type: none"> 한국수어 말뭉치 구축 <ul style="list-style-type: none"> 의미 전사 변이형 전사 문법 정보 전사 마우스 제스처 전사 사전용 말뭉치 구축 <ul style="list-style-type: none"> 기본 전사 의미 전사 멀티미디어 자료 구축 	<ul style="list-style-type: none"> 한국수어 말뭉치 구축 <ul style="list-style-type: none"> 기본 전사(계속) 	<ul style="list-style-type: none"> 한국수어 말뭉치 구축 <ul style="list-style-type: none"> 기본 전사(계속)
한국 수어 지식 관리 시스템 구축	<ul style="list-style-type: none"> 한국수어 자료 통합 관리 시스템 고도화 한국수어-한국어 용례 구축 시스템 개발 및 기능 고도화 	<ul style="list-style-type: none"> 한국수어-한국어 사전 편찬기 개발 및 기능 고도화 한국수어-한국어 사전 서비스 시스템 구축 및 기능 고도화 <ul style="list-style-type: none"> 한국어 어휘망 활용 검색 도입 및 수형 검색 기능 고도화 사용자 화면 편집 기능 개발 모바일 웹 인터페이스 구축 	<ul style="list-style-type: none"> 한국어-한국수어사전 편찬기 개발 및 기능 고도화 한국어-한국수어사전 서비스 시스템 구축 및 기능 고도화 한국수어 뜻풀이 영상 관리 시스템 개발 및 기능 고도화 	<ul style="list-style-type: none"> 한국어-한국수어사전 서비스 시스템 구축 및 기능 고도화 (이원화된 사전 간 연결)
양방향 사전 편찬	<ul style="list-style-type: none"> 〈한국수어-한국어 사전〉 <ul style="list-style-type: none"> 사전 시범 제작 한국수어 표제어 구축 및 지침 마련 한국수어 의미 구축 및 지침 마련 한국수어 뜻풀이 구축 및 지침 마련 한국수어 용례 구축 및 지침 마련 한국어 대응 표현 구축 및 지침 마련 한국수어 뜻풀이 구축 및 지침 마련 수어 동영상 제작 및 감수 	<ul style="list-style-type: none"> 〈한국수어-한국어 사전〉 <ul style="list-style-type: none"> 한국수어 표제어 추가 구축 한국수어 변이형 구축 및 지침 마련 수어 문자 구축 및 지침 마련 수형 정보 구축 및 지침 마련 한국수어 의미 구축 확장 한국수어 뜻풀이 확장 구축 한국수어 용례 확장 구축 한국어 대응 표현 구축 관련어 구축 및 지침 마련 수어 동영상 제작 및 감수 	<ul style="list-style-type: none"> 〈한국수어-한국어 사전〉 <ul style="list-style-type: none"> 문법 정보 구축 및 지침 마련 수어 지도 구축 및 지침 마련 마우스 제스처 정보 구축 및 지침 마련 사용성 평가 및 오류 정비 수어 문자 교육 및 보급 오류 정비 〈한국어-한국수어사전〉 <ul style="list-style-type: none"> 한국어 표제어 1만 개 평정 한국수어 뜻풀이, 용례 선정 및 지침 마련 한국어의 뜻풀이와 용례를 한국수어로 풀이 한국수어 대응 표현 구축 및 지침 마련 관련어 구축 	<ul style="list-style-type: none"> 〈한국수어-한국어 사전〉 <ul style="list-style-type: none"> 수어 문자 교육 및 보급 오류 정비 〈한국어-한국수어사전〉 <ul style="list-style-type: none"> 한국어의 뜻풀이와 용례를 한국수어로 풀이 한국수어로 한국어의 뜻풀이 확장 구축 한국어의 용례 수어 번역 확장 구축 한국수어 대응 표현 구축 한국수어 대응 표현 동영상 제작

〈표 3〉 양방향 이중 언어 사전 중장기 구축 계획

오랜 시간과 대규모의 작업 범위로 농사회의 자원이 총 투입되는 이 계획에 따르면 새로운 한국수어사전은 2031년에 완성될 예정이다. 향후 편찬될 사전은 다음과 같은 부분에서 이전의 사전과는 차별화된 모습으로 도약할 것이라고 기대해 볼 수 있다.

○ 이용자의 사용 목적에 따른 사전 이원화 편찬

수어사전을 편찬한 대부분의 나라들이 이중 언어 사전을 편찬하였지만 현실적인 제한(사업 기간, 예산 등)으로 인해 하나의 사전으로 두 목적을 이루는 방향을 취하고 있다. 이러한 사전은 주로 수어 사용자에게는 수어에 해당하는 음성언어 대역어를 제공하고, 수어 학습자에게는 수어와 관련된 정보를 제공한다. 물론 이것은 두 가지 목적을 동시에 이룬다는 장점이 있으나 여전히 하나의 사전으로 모든 정보를 제공하기에는 제한적일 수밖에 없다. 따라서 학습자 사전으로서의 효용성을 높이기 위해서는 두 목적을 분리한 사전의 편찬이 필요하다. 국립국어원에서 계획하고 있는 사전은 목적에 따라 이원화한 것으로 ‘한국수어-한국어 사전’은 한국수어 사용자를 위한 표현용 사전이며, 한국수어 학습자를 위한 이해용 사전이다. 반면에 ‘한국어-한국수어’ 사전은 한국어 학습자를 위한 이해용 사전이며 한국수어 학습자를 위한 표현용 사전이다.

○ 증거가 있는 사전

지금까지 한국에서 편찬된 6종의 수어사전은 대체로 비슷한 방법으로 자료를 수집하였는데 그 방법은 다음과 같다. 첫째, 한국어 표제어 목록을 선정하고 이에 대응하는 한국수어를 수집하거나 수어 표현이 존재하지 않는 경우에는 수어를 만든다. 둘째, 기존에 나와 있는 수어 교재 또는 사전에서 수어 단어를 표집한다. 이렇게 수집한 수어를 사전으로 엮었지만 이용자 입장에서는 어느 수어가 만들어진 것인지, 어느 수어가 농인이 실제로 사용하는 것인지 등을 확인할 수 없어 혼란이 가중되었고 결국 사전의 신뢰도가 하락하였다. 이에 수어를 기술하는 사전으로 편찬 방향을 설정하고 말뭉치에 근거하여 표제어 등 사전 정보를 추출하고자 한다.

수어를 정의하는 부분에서 중요한 것은 수어를 제1언어로 사용하는 사람(native signer)이 해당 수어를 맥락 안에서 어떻게 사용하는지를 보는 것이다(Johnston, 1998). 따라서 새로운 사전이 이용자에게 단어의 적절한 용법을 가르쳐 줄 수 있도록 용례를 제공하고자 한다. 용례는 말뭉치에서 추출한 용례를 부분적으로 가공한 가공용례와 농인의 직관에 의해 제작한 작성용례로 구분한다. 이를 제시할 경우 그 출처를 표시하여 사전 이용자들이 참고할 수 있도록 할 예정이다.

○ 변이 수용

Johnston(2003)은 수어에 관해서, 어휘적 변이(lexical variation, 다양한 수어 사용자들이 같은 의미를 완전히 다른 수어로 나타내는 것)는 몇몇의 수어사전 편찬자들에게 주요한 문제가 된다고 하였다. 그러나 이러한 변이를 없애기 위해 추진하는 수어 표준화가 얼마나 실현 가능성이 낮은지 여러 이유를 근거로 제시하였다. 역사적으로 우리나라뿐만 아니라 다른 나라에서도 시도되었던 음성언어중심의 수어 표준화는 농사회로부터 외면을 받아왔다. 그러므로 이러한 변이를 없애려는 시도보다 변이를 모두 수용하고 기록하

2부 발표

여 사전이 수어언어의 총체적 자원이 되도록 해야 할 것이다.

○ 수어문자(SignWriting) 도입

수어는 문자 체계가 없는 언어로 이것은 사전 편찬을 더 어렵게 만드는 요인이 된다. 수어사전 편찬자들은 이러한 어려움을 극복하고자 여러 접근을 하고 있는데 Kristoffersen & Troelsgård(2010)에 따르면 편찬된 수어사전들은 수어 표제어를 나타내기 위해 동영상, 그림, 사진, 형태 표기(HamNoSys, SignWriting 등), 글로스(glosses), 숫자, 대역어, 어떻게 수어가 산출되는지 글로 묘사하기와 같은 다양한 방법을 사용하고 있다. 대부분은 글로스를 사용하는데 이것은 수어의 의미를 나타낸 음성언어의 어휘이며, 수어를 기억하기 위해 사용된다. 음성언어의 어휘를 하나의 라벨처럼 수어 표제어에 부여하고 관리하는 것은 사전 편찬자와 이용자 모두에게 상당한 편의를 제공하지만 해당 수어가 음성언어 어휘의 의미와 완전히 동일하다거나 혹은 그 어휘가 가지고 있는 여러 의미 중에서 수어는 단 하나의 의미만을 가지고 있다는 오해를 일으키기도 한다. 또한 음성언어라는 라벨을 보고 수어에 접근하는 것은 결국 음성언어를 통해 수어를 인지하도록 해서 수어 학습에 방해 요소가 된다.

수어를 정확하게 기술하고, 학습자로 하여금 이를 시각적으로 인식할 수 있도록 하기 위해서 향후 나올 사전에서는 수어문자(SignWriting)를 사용해 수어를 표기하고자 한다. 이미 미국, 호주, 벨기에, 브라질, 캐나다 등 여러 나라에서 수어문자를 사용하고 있으며¹¹⁾ 체코수어사전¹²⁾에서는 수어문자를 사용하여 표제어를 표기하고 있다. 수어문자는 한국의 국가사업으로는 처음 도입되는 것으로 이를 계기로 하여 수어로 생산되는 많은 정보가 기록, 전송되는 발판을 마련하게 될 것이다.

○ 번역이 아닌 설명

한국수어가 하나의 언어로 인정받기 이전의 시기에는 이에 대한 이해 부족으로 한국어에 한국수어 어휘만 대응시킨 일대일 번역이 주를 이루었다. 그러나 ‘청각장애학생을 위한 국어교육용 한국어-수어 동영상 사전 개발(윤점룡 외, 2009)’에서 지적한 바처럼 시각 언어인 수어는 청각 음성언어인 한국어와는 다른 문법적 특성을 가지고 있으므로 단어를 일대일로 대응시켜 번역하는 것만으로는 의미를 제대로 전달할 수 없다. 이와 같은 오류를 피하기 위하여 국립국어원은 2017년에 한국어의 뜻을풀이를 한국수어로 시범 번역하였고¹³⁾ 문헌 고찰과 전문가 의견 수렴을 통해 지침을 수립하였다. 그러나 번역 시 한국수어가 한국어에 강하게 견인되고, 한국어와 한국수어 간의 설명 방식 차이가 상당하여 수어 사용자들이 번역문을 이해하지 못하는 경우가 발생하였다. 이와 같은 경험에 근거하여 수어 사용자가 사전의 뜻을풀이를 직관적으로 이해할 수 있도록 번역이 아닌 설명 방식의 뜻을풀이를 제공하고자 한다. 또한 대부분의 경우 예시를 들어 설명하는 농인의 어법을 충실히 따라 뜻을풀이와 용례를 분리하지 않고 하나의 흐름으로 제작하여 상호보완적 역할을 하도록 하며 이를 통해 수어 사용자의 접근성과 활용도를 높일 것이다.

11) <http://www.signwriting.org/about/who/>

12) <http://www.dictio.info/czj?action=page&page=front&lang=en>

13) 한국수어사전 수어 뜻을풀이와 용례 구축 및 정비 사업

4. 결론

지금까지 한국수어사전의 편찬 현황을 개관하고 발전 방향을 제시하였다. 현 시점에서 그간 편찬된 한국 수어사전은 진정한 의미의 언어 사전으로서 틀을 갖추지 못하였다고 할 수 있다. 그러나 국립국어원은 <한국수어언어법> 제정(2016) 이후에 관련 정책과 연구를 위한 기틀을 마련하고 있다. 더불어 한국수어사전이 언어 지식의 총체로 거듭날 수 있도록 다각도로 접근하고 있다. 이러한 계획이 적절히 실현되기 위해서는 대규모의 인력이 필요하나 그간 농사회에서 연구 인력 양성과 그에 대한 투자가 충분하지 못했다. 이러한 제한으로 사전 구축 초기에는 다소 시행착오를 겪을 수도 있겠으나 지속적으로 연구 인력 양성이 병행된다면 농인의 언어 그리고 농인과 청인이 공존하는 방향을 담은 사전에 한 발 더 다가갈 수 있을 것이라 기대해 본다.

참고문헌

- 김삼찬, 이선호 외 36인(2005), 《한국수어사전》, 형설출판사.
- 김승국(1993), 《표준수어》, 오성출판사, 서울.
- 국립국어원(2014), 《한국수어 연구 자문 위원회 회의 결과 보고서》, 국립국어원.
- 국립국어원(2016), 《수형 기반 한국수어사전 구축》, 국립국어원.
- 국립국어원(2017), 《한국수어사전 수어 뜻풀이와 용례 구축 및 정비》, 국립국어원.
- 국립국어원, 한국농아인협회(2005), 《한국수어사전》
- 김승국(1983), 《수어사전》, 단국대학교출판부
- 윤점룡, 정희선, 김선영, 원성욱, 김지숙, 장은숙, 김경진, 임진이, 하길중, 양국종(2009), 《청각장애 학생을 위한 국어교육용 한국어-수어 동영상 사전 개발》, 한국연구재단.
- 이정옥, 이준우(2005), 한국수어 관용표현에 관한 탐색적 연구, 《특수교육 저널: 이론과 실천》 제6권 제3호, 한국특수교육문제연구소, 335-354쪽
- 이준우, 남기현, 조준모(2018), 한국수어사전 편찬 방향을 위한 제언, 《특수교육 저널: 이론과 실천》 제19권 제1호, 한국특수교육문제연구소, 137-156쪽
- 이현화(2017), 한국수어 정비 사업: 한국수어사전(한국수어의 기록), 《새국어생활》, 제27권 제2호, 67-83쪽

2부 발표

최경봉(2014), 원고로 남은 최초의 우리말 사전, 《말모이》, 《새국어생활》 제24권 제3호, 국립국어원, 20-37

표준수어사전편찬위원회(1982), 《표준수어사전》, 금옥학술문화재단

허일(2003), 한국 농인의 언어 사전 편찬에 있어서의 쟁점 연구, 제5회 한국수어학회 학술 대회.

Atkins, B.T.S. & M. Rundell. (2008), The Oxford Guide to Practical Lexicography, Oxford/New York: Oxford University Press.

Cokely, D. & Baker-Shenk, K. C. (1980), American Sign Language: A Teacher's Resource Text on grammar and Culture, Gallaudet University Press.

Johnston, T.(1998). Signs of Australia. A new dictionary of Auslan(Australian Sign Language) (the sign language of the Australian deaf community), North Rocks, NSW, Australia: North Rocks Press, Royal Institute for Deaf and Blind Children.

Johnston, T.(2003). Language Standardization and signed Language Dictionaries, Sign Language Studies, vol 3, Published by Gallaudet University Press, 431-468.

Kristoffersen, J. H., & Troelsgård, T. (2010), Compiling a Sign Language Dictionary: Some of the Problems faced by the Sign Language Lexicographer In M. Mertzani(Ed.), Sign Language Teaching and Learning: Papers from 1st Symposium in Applied Sign Linguistics(Applied Sign Linguistics), Bristol: University of Bristol, Centre for Deaf Studies.

Roland Pfau & Markus Steinbach & Bencie Woll(2012), Sign Language: An International Handbook, de Gruyter Mouton

R.R.K. Hartmann (1983), 《사전 편찬의 원리와 실제》(서태길, 김양진, 도원영, 이상혁, 권오희, 옮김). 서울: 제이앤씨.(원서출판 1983)

국립국어원 한국수어사전 (<http://sldict.korean.go.kr/front/main/main.do>)

수어문자 누리집 (<http://www.signwriting.org/about/who/>)

한국민족문화대백과사전 (<http://encykorea.aks.ac.kr/>)

체코수어사전 (<http://www.dictio.info/czj?action=page&page=front&lang=en>)

Current Status and Development Direction of Korean Sign Language Dictionary

Hyewon Choi, Hyunhwa Lee
National Institute of Korean Language, Korea

1. Introduction

It has been approximately 2 years since the <Korean Sign Language Act> (2016) was implemented. The <Korean Sign Language Act>, embodying a long-standing desire of the deaf who had hoped Korean sign language to be recognized as a language in Korean society and to be granted with the right to use Korean sign language, became the basis for various systems and projects that support the development of Korean sign language. Since the enactment of the <Korean Sign Language Act> in 2016, the Republic of Korea has carried out a wide range of research and projects for the development of Korean sign language including systematic establishment of Korean sign language development plan in every 5 years; research on the use of Korean sign language to improve the environment for the use of Korean sign language; creation of Korean sign language corpus; and full-scale basic research on grammar. In addition, a qualification system for Korean sign language teachers has been put in place along with development of sign language textbooks for different targets and Korean sign language cultural school programs for education and distribution of Korean sign language. The Republic of Korea is also seeking for the ways to improve 《Korean Sign Language Dictionary》 to a dictionary that conveys the summation of the language.

The compilation of a sign language dictionary is the most fundamental project to publicize the independent status of the sign language. A sign language dictionary that describes sign language as a language, against a common notion that sign language is a sub-language of a spoken language and a subsidiary communication means, proves that the deaf form an independent linguistic community. In fact, many countries have provided opportunities to enhance the status of the deaf community culturally and linguistically by publishing a sign language dictionary. Of course, the form of a sign language dictionary in this case should be more than just a translation dictionary that simply presents sign language equivalents to spoken language words from a spoken language-oriented perspective.

When Joseon, an old Korean dynastic kingdom, was a colony of Japan, Korean linguists tried to compile the

Session 2

first Joseon language dictionary in Korean alphabet. However, 《Joseon Language Dictionary》(1920) became the first dictionary published by the Government-General of Joseon. Unlike the title, the dictionary was centered around Japanese language, targeting Joseon language. The Joseon language¹⁾ dictionary, which was compiled while Joseon was not an independent country, consisted of Japanese entries, Korean equivalents in Korean alphabet and definition of Japanese entries. The dictionary was used for a colonization policy to ‘teach Japanese language to Koreans and Korean language to Japanese (Choi Kyeong-bong, 2014).

Just as the case above, Korean sign language dictionaries published in the Republic of Korea so far did not have the right frame as a true language dictionary either. Sign language dictionaries in the form of a target language dictionary were made by hearing teachers without deep understanding of Korean sign language as a means to teach a curriculum in Korean texts. This type of dictionaries still continue to follow the old format without being able to overcome their inherent limitations. Sign language dictionaries in western countries²⁾ had a similar starting as they were for education of the deaf. However, more advanced sign language dictionaries began to be compiled mostly in Europe around the 1990s to present the system of sign language as a natural language from a linguistic approach. The Republic of Korea also announced the Korean sign language as a language through the <Korean Sign Language Act>. Nevertheless, to ensure that a sign language is recognized as a legitimate language in Korean society, the country needs to go beyond publicizing and should compile a dictionary that clearly shows the system of Korean sign language as a language. This conference is expected to play a pivotal role in presenting infinite potential of sign language and creating a dictionary that supports the identity and pride of the Korean deaf community.

This article will look into the current status and challenges of the sign language dictionaries published in the Republic of Korea and the mid- & long-term plan established by the National Institute of Korean Language for the compilation of a new Korean sign language dictionary that faithfully includes actual Korean sign language data used by the deaf.

2. Current Status of Korean Sign Language Dictionary


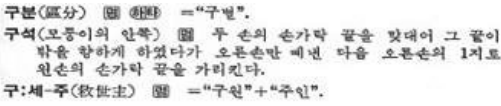
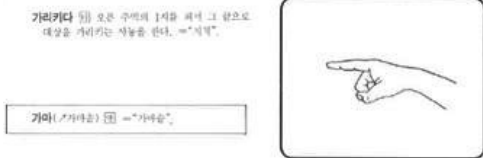

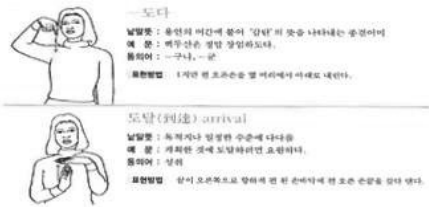

When there was no sign language dictionary, a list of sign language words was included in sign language textbooks created for the distribution of sign language. It was not until 1982 that the first sign language dictionary was published in the Republic of Korea under the title of 《Standard Sign Language Dictionary》. As the title

-
- 1) After the publication, a Korean dictionary titled as 《Joseon Language Dictionary》(1938) in Korean alphabet was created thanks to the hard work by Moon Se-young, a Korean linguist, and it gave a great hope to the Koreans.
 - 2) An Australian sign language dictionary from 1989 has a trace of applying English language grammar (‘-ed’ for past participle) to sign language expressions.
 - 3) The ‘standardization’ here is different from general standardization of a language. As there was no collection of a broad range of sign language documents, the standardization was not about selecting the most widely used form as the standard from variations. Rather, it focused on ‘regularizing’ sign language from a perspective of spoken language by creating new sign language equivalents of Korean language if there was no equivalent and adding supplementary signs to distinguish same signs with different meanings.

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

suggests, this dictionary was created to contribute to ‘standardizing sign language’³⁾(Lee Jun-woo, Nam Ki-hyun & Cho Jun-mo, 2018). The ‘Korean Standard Sign Language Dictionary Compilation Committee’ consisting of school teachers was formed to this end and the committee compiled the dictionary which contained sign language words for daily life and the sign language equivalents of Korean words used in elementary and middle school textbooks.

As seen in the table below, a total of 6 sign language dictionaries have been published in the Republic of Korea starting from <Standard Sign Language Dictionary> (1982).⁴⁾

No.	Dictionary/ Institution(compilerr)	Year	No. of entries	Micro structure
1	《Standard Sign Language Dictionary》 Standard Sign Language Dictionary Compilation Committee	1982	5,494	
2	《Sign Language Dictionary》 Dankook University Press	1983	5,966	
3	《Standard Sign Language》 Kim Seung-kuk	1993	5,966	
4	《Korean Sign Language Dictionary》 NIKL- Korea	2005	Around 6,800	
5	《Korean Sign Language Dictionary》 Kim Sam-chan and others	2005	9,818	
6	《Korean Sign Language Dictionary》 National Institute of Korean Language (online)	2015	12,793	



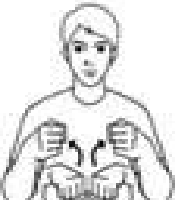

<Table 1> Korean sign language dictionaries published in the Republic of Korea

4) Dictionaries for etymology and terminology are excluded.

Session 2

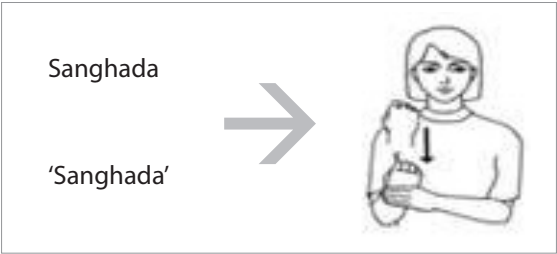
As seen above, the 6 dictionaries presented in Korean alphabetical order are translation dictionaries that contain Korean entries and sign language equivalents. Even the translation in these dictionaries included artificially created signs in addition to actually used sign language simply because it was necessary in expressing Korean words. As a result, an excessive number of sign language entries were compiled including many artificial sign language expressions corresponding to suffixes, prepositions and compound words of Korean language.⁵⁾

A translation dictionary in general provides various meanings of an A language entry in B language as words or alternative expressions. For instance, if an entry is ‘Sanghada (go bad)’ in Korean, a translation dictionary should propose different meanings of ‘Sanghada’ along with the information on Korean sign language translations.

Korean language entry	Meaning / Korean sign language translation			
	‘Get hurt’	‘Become thin’	‘Get ruined’	‘Go stale’
Go bad				

<Table 2> Various sign language expressions equivalent to ‘Sanghada’ in Korean language

However, Korean sign language dictionaries either limit the sign language equivalent of a Korean language entry to one and suggest only one Korean sign for a Korea entry (1,5) or simplify the definition of a Korean language entry (4,5,6). This could cause a misunderstanding about Korean sign language among the dictionary users that one sign is the equivalent of one Korean word and that Korean language and Korean sign language share the identical semantic domain.



<Picture 1> Korean language entry ‘Sanghada’ and Korean sign language with the same gloss

5) The government of the Netherland in 1998 has set Standardisation of the Basic Lexicon of NGT (STABOL) as a priority task to recognize the legal status of Dutch sign language. The project, opposed greatly by sign language researchers in the deaf community, standardized approximately 14,000 words. 33% of them were new words including terminologies for medical and legal sectors as well as for school textbooks. (Roland Pfau, Markus Steinbach & Bencie Woll, 2012)

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

As these dictionaries were based on the frame of a translation dictionary, the information on sign language was limited to the morphological features of sign language rather than embracing linguistic characteristics. Furthermore, the unnatural linkages of Korean sign language with Korean language in these dictionaries resulted in distorting Korean sign language. For instance, the sign language for ‘Dodol’, meaning ‘reaching a destination or a certain level’ in Korean in dictionary 5 shows that the hand orientation is vertically upward although ‘destination’ is not always above. Sign language users change the hand orientation to express ‘Dodol’ depending on the locational features of a ‘destination’, but the dictionary does not reflect such usage. Also, due to the characteristics of Korean sign language that has a great number of detailed expressions, a sign language gesture may not correspond to Korean language examples even if the gesture is equivalent to the definition of the Korean entry provided in the dictionary. For example, the sign language for ‘Gajangjari’ (circumference or edge) in the dictionary is not particularly wrong in general meaning, but is inadequate to be used in an expression as, “The edge of an eye is red.”



<Picture 2> A sign language expression equivalent to an entry ‘Gajangjari’

Korean sign language dictionaries are not far behind compared to the dictionaries of other countries in terms of the time it was first compiled, the number of dictionaries and the number of entries.⁶⁾ Nevertheless, Korean sign language dictionaries are in a stalemate without much improvement when it comes to how faithfully the content reflects the actual language.⁷⁾ This is because the compilation process that had continued up to the creation of 《Korean Sign Language Dictionary》 in 2005 did not develop much from the basic intention of the ‘standardization of sign language’. This limitation is also seen in 《Korean Sign Language Dictionary》⁸⁾(2015) compiled by the National Institute of Korean Language, which is still in service online, as it reflects the format of 《Korean Sign Language Dictionary》 created in 2005.

- 6) The first sign language dictionary in New Zealand(The Dictionary of New Zealand Sign Language, 1986) included 1,200 entries while the first sign language dictionaries in Australia(Auslan Dictionary, 1989) and Britain(Dictionary of british sign language, 1992) contained 3,500 and 1,800 entries, respectively.
- 7) The linguistic research methodology for American sign language by researchers of Gallaudet University led by Stokoe in the 1960 affected the researchers in the UK and Europe in the mid-1970s, and sign language became the major subject of research after the 1980s.
- 8) 《Korean Sign Language Dictionary》(2015) includes the lexicon from 《Korean Sign Language Dictionary》(2005) and 6,000 other words that were added each year under the same direction of compiling dictionary.

Session 2

3. Development Direction of Korean Sign Language Dictionary

As has been described, 《Korean Sign Language Dictionary》, which was a result of the Korean sign language standardization project, is meaningful in that it collected and studied Korean sign language materials and called attention to the methodologies of sign language research. However, because of the limitations of the time when it was published, the dictionary has continued to receive criticism from its users. The National Institute of Korean Language has humbly accepted the criticisms and is now planning to compile a new Korean sign language dictionary that overcomes the limitations of the previous dictionary.

According to Atkins & Rundell(2008), designing an effective dictionary begins from the realistic understanding of the objectives and capabilities of users. As such, the National Institute of Korean Language has collected opinions from the deaf community through various projects to set the right direction for the improvement of 《Korean Sign Language Dictionary》⁹⁾. The ‘Korean Sign Language Research Advisory Committee’, established in 2014, conducted a ‘Survey on the demand for enactment and standardization of sign language terminology’. According to the survey, respondents answered that the biggest reason for not using standardized sign language terminologies was because ‘it was different from the sign language actually being used (28.1%)’.

The ‘Creation of Korean sign language dictionary based on hand orientation’ project (National Institute of Korean Language, 2016) performed a targeted group interview with 15 deaf, and the result showed that the existing dictionaries were ‘lacking sign language expressions frequently used in the deaf community’ and ‘disregarded by the deaf as they are centered around sign language expressions that are equivalents to Korean language’. The interviewees also answered that the existing dictionaries were ‘used for educating language interpreters and teaching sign language to hearing people’, but were ‘difficult even for the deaf to understand’ which could cause ‘communication problems between the deaf and hearing sign language interpreters’. The deaf saw the existing dictionaries as ‘dictionaries they use when necessary but not of a great help’ and evaluated that even using the dictionaries to understand Korean language was limiting because “the information provided in Korean language was difficult for the deaf to understand”.

The ‘Establishment and modification of definitions and examples of sign language in Korean sign language dictionaries’ project (National Institute of Sign Language) in 2017 performed a survey targeting 30 sign language experts in the sign language academia and deaf community. The result suggested that the drawbacks of 《Korean Sign Language Dictionary》 were poor accessibility and artificially created sign language expression from a macro-structure perspective. For a micro structure, poor matching of information items and lack of examples and information on grammar that are absolutely necessary for the use of sign language.

The two studies performed each in 2016 and 2017 also collected suggestions for the development direction of 《Korean Sign Language Dictionary》, and the major details are as follows.

9) This article selectively addresses the opinions that are necessary from the survey results.

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

- Creation of Korean sign language dictionary based on hand orientation (National Institute of Korean Language, 2016)
 - Creation of a new dictionary centered around the deaf users
 - Easy explanation of Korean language information in sign language expressions commonly used by the deaf¹⁰⁾
 - Clear designation of the objectives and target users of a sign language dictionary
 - Measures to improve accessibility
- Establishment and modification of definitions and examples of sign language in Korean sign language dictionaries (National Institute of Korean Language, 2017)
 - Dictionaries that can be used by both the deaf and hearing people
 - Deletion and modification of artificial entries in existing dictionaries
 - A dictionary that contains sign language actually used by the deaf and faithfully includes sign language-related information
 - A dictionary that contains diverse variations of sign language
 - Development of methods to categorize entries and search for words (motion recognition, etc.) in order to enhance accessibility of dictionary

Considering the results described above, the users of the current 《Korean Sign Language Dictionary》 can be divided into Korean sign language learners and Korean language learners, and the two groups are using the dictionary with different objectives. Also, the dictionary created and included artificial Korean sign language expressions, but these were not accepted by the deaf community. As such, the dictionary can not perform its role as a dictionary. It is necessary to revamp 《Korean Sign Language Dictionary》 based on these opinions to ensure that it reflects the actual language better and it is more complete while compiling a new dictionary so that the users can use a dictionary according to their objectives.

According to Hartmann(1983), a bilingual dictionary can be useful as a supportive tool to learn and teach a foreign language. In addition, Lee Hyun-hwa (2017), who discussed the need for publishing a descriptive dictionary for Korean sign language, argued that a Korean sign language dictionary should be compiled based on the entries extracted from the materials that contain natural dialogue of the Korean deaf rather than artificially-made materials. These suggestions lead to a conclusion that the direction of a Korean sign language dictionary should follow the two below.

First, the Korean sign language dictionary should have a clear objective and should be dualized into ‘Korean sign language-Korean language dictionary’ and ‘Korean language-Korean sign language dictionary’.

5) Cokely, D. & Baker-Shenk, K, C. (1980) noted, in explaining American sign language (ASL), that most of expressions that are described as ‘idioms’ are actually not idioms and do not fit the definition of idiom and that they are simply sign language gestures that have a few different uses. Lee Jeong-ok & Lee Jun-woo (2005) also summarize the same case in Korea, pointing out that idiomatic expressions are accepted as ‘natural sign language’ or ‘sign language of the deaf’. The result of the survey mentioned in this article can be interpreted within the same context.

Session 2

Second, the dictionary that embodies the characteristics of a sign language with a wide range of variations should be based on descriptive linguistics, and, as such, it should be centered around the Korean sign language corpus that has a collection of empirical materials of Korean sign language.

3.1. Mid- & long-term plan to establish a two-way bilingual dictionary

The mid- & long-term plan to publish a new Korean sign language dictionary established by the National Institute of Korean Language in consideration of the opinions and suggestions by the users of 《Korean Sign Language Dictionary》 and expert groups is as follows.

3.1.1. Structure of two-way bilingual dictionary

(1) Macro structure

- Korean sign language – Korean language dictionary
 - Entry selection
 - Select entries from the Korean sign language corpus ('15~) and an additional corpus to be established exclusively for dictionary (tentative)
 - Include daily vocabularies, dialects and new words focusing on the frozen lexicon of Korean sign language
 - Notation of entry
 - Notate sign language in sign writing (SignWriting) that could document the form of Korean sign language to raise an awareness that Korean sign language is an independent language from Korean language
 - Address various issues caused by the use of gloss
 - Entry arrangement
 - Refer to the classification of chereme in 'Creation of Korean sign language dictionary based on hand orientation (National Institute of Korean Language, 2016) for arrangement
- Korean language – Korean sign language dictionary
 - Entry selection
 - Evaluate 10,000 entries necessary for the Korean deaf to learn Korean language
 - Focus on Korean entries included in the Korean sign language-Korean language dictionary as Korean language equivalents
 - Notation of entry
 - Notate entries in Korean letters (Hangul) but suggest Korean sign language expressions equivalent to Korean language entries in thumbnails
 - Arrangement of entry
 - In Korean alphabetical order

(2) Micro structure

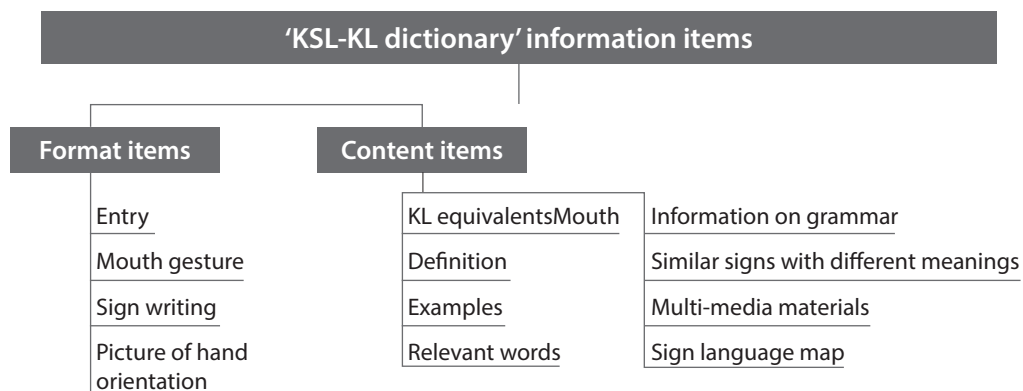
□ Korean sign language - Korean language dictionary

○ Format items

- (Entry) Provide via a video to show both default and variations
- (Mouth gesture) Present mouth gestures, essential for sign language, in a picture
- (Sign writing) Document the form of entries in sign writing
- (Hand shape picture) Provide pictures of hand shape to be used for searching for chereme

○ Content items

- (Equivalent expression) Propose multiple Korean language expressions that are the equivalents of a Korean sign language expression
- (Definition) Describe the meaning of Korean sign language
extracted from the Korean sign language corpus in easy Korean language
- (Example) Suggest various examples of entries via a video
 - Notate the provided examples in sign writing and link the notation with the entry concerned so that users can click the link to find the entry
 - Translate examples into accurate and natural Korean language
- (Relevant words) Suggest Korean sign language gestures that are semantically related to an entry
- (Morphologically similar signs with different meanings) Provide information on the signs that are frequently mistaken by Korean sign language learners for more effective sign language learning
- (Information on grammar) Present information on grammar such as plural, negative, conjugation, etc. of sign language
- (Sign language map) Include information on dialects via sign language map considering the characteristics of a sign language that has a wide range of regional variations.



<Picture 3> Korean sign language-Korean language dictionary information items

Session 2

표제어 표기 (수어 문자)

수형 그림

인쇄

1. 한국어 대응 표현1, 한국어 대응 표현2

1) 한국수어의 뜻풀이

2) 용례

3) 관련어

2. 한국어 대응 표현3, 한국어 대응 표현4

1) 한국수어의 뜻풀이

2) 용례

3) 관련어

· 형태는 비슷하지만 의미는 다른 수어 어휘

· 문법 정보

· 수어 지도

· 한국어 번역문

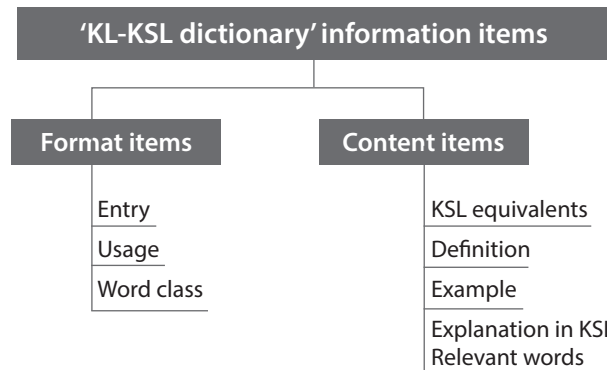
· 용례 표기 (수어 문자)

<Picture 4> A demonstration of Korean sign language-Korean language dictionary

- Korean language-Korean sign language dictionary
 - Format items
 - (Entry) Suggest around 10,000 entries
 - (Usage) Propose usages of entries for more effective Korean language study
 - (Word class) Provide word class of Korean language entries

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

- Content items
- (Equivalent expression) Propose multiple Korean sign language expressions that are the equivalents of an entry in thumbnail
- (Definition) Present definitions from a basic Korean language dictionary
- (Examples) Provide examples from a basic Korean language dictionary
- (Explanation in Korean sign language) Explain the definitions and examples of Korean language entries via a video based on the understanding by the deaf rather than translation
- (Relevant words) Suggest vocabularies that are synonyms or antonyms of Korean language entries as relevant words



<Picture 5> Korean language-Korean sign language dictionary information items



<Picture 6> A demonstration of Korean Language-Korean Sign Language Dictionary

Session 2

3.1.2. Publication plan for two-way bilingual dictionary

	Phase 1 ('20~'22)	Phase 2 ('23~'25)	Phase 3 ('26~'28)	Phase 4 ('29~'31)
Collect KSL materials	<ul style="list-style-type: none"> • Collect and modify Korean sign language materials • Establish KSL corpus <ul style="list-style-type: none"> - Create corpus for Seoul and other regions - Basic transcript - Semantic transcript • Conduct research for the corpus for dictionary and accumulate materials 	<ul style="list-style-type: none"> • Establish KSL corpus <ul style="list-style-type: none"> - Semantic transcript - Transcript of variations - Transcript of grammatical information - Mouth gesture transcript • Create corpus for dictionary <ul style="list-style-type: none"> - Basic transcript - Semantic transcript • Accumulate multi-media materials 	<ul style="list-style-type: none"> • Create KSL corpus <ul style="list-style-type: none"> - Basic transcript (continued) 	<ul style="list-style-type: none"> • Create KSL corpus <ul style="list-style-type: none"> - Basic transcript (continued)
Establish KSL Knowledge management system	<ul style="list-style-type: none"> • Improve KSL material integration and management system • Develop a system to build examples of KSL – KL and enhance its functions 	<ul style="list-style-type: none"> • Develop KSL-KL dictionary compilation system and enhance its functions • Establish KSL-KL dictionary service system and enhance its functions <ul style="list-style-type: none"> - Introduce search functions utilizing Korean language lexical network and enhance search functions based on hand orientation - Develop user interface editing function - Establish mobile web interface 	<ul style="list-style-type: none"> • Develop KL-KSL dictionary compilation system and enhance its functions • Establish KL-KSL dictionary service and enhance its functions • Develop a management system for KSL definition videos and enhance its functions 	<ul style="list-style-type: none"> • Establish KL-KSL dictionary service system and enhance its functions (connect the two dictionaries)
Publish two-way dictionaries	<ul style="list-style-type: none"> • <KSL-KL dictionary> <ul style="list-style-type: none"> - Create a dictionary for trial - Establish KSL entries and guidance - Establish KSL meanings and guidance - Establish KSL definitions and guidance - Establish KSL examples and guidance - Establish equivalent KL expressions and guidance - Establish KSL definitions and guidance - Create and edit sign language videos 	<ul style="list-style-type: none"> • <KSL-KL dictionary> <ul style="list-style-type: none"> - Establish additional KSL ' entries - Establish KSL variations and guidance - Establish sign writing and guidance - Establish information on hand shape and guidance - Establish and expand meaning of KSL - Expand KSL definitions - Expand KSL examples - Establish equivalent KL expressions - Establish relevant words and guidance - Create and edit sign language videos 	<ul style="list-style-type: none"> • <KSL-KL dictionary> <ul style="list-style-type: none"> - Establish grammatical information and guidance - Establish sign language map and guidance - Establish information on mouth gesture and guidance - Evaluate usability and correct errors - Teach and distribute sign writing - Correct errors • <KL-KSL dictionary> <ul style="list-style-type: none"> - Evaluate 10,000 KL entries - Select KSL definitions and examples and establish guidance - Explain definitions and examples of KL in KSL - Establish equivalent KSL expressions and guidance - Create relevant words 	<ul style="list-style-type: none"> • <KSL-KL dictionary> <ul style="list-style-type: none"> - Teach and distribute sign writing - Correct errors • <KL-KSL dictionary> <ul style="list-style-type: none"> - Explain definitions and examples of KL in KSL - Expand KL definitions in KSL - Expand KSL translations of KL examples - Establish equivalent KSL expressions - Create videos for equivalent KSL expressions

<Table 3> Mid- & long-term plan for compilation of two-way bilingual dictionary

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

The plan, a long-term large-scale project resourced by the deaf community, states that the new Korean sign language dictionary will be completed in 2031. The dictionary that will be published in the future can be expected to be differentiated and more advanced from the previous dictionaries with the following reasons.

- Dualized compilation of dictionary according to objectives of users

The majority of the countries that have published sign language dictionaries have bilingual dictionaries, but these dictionaries attempt to fulfill two different objectives with one dictionary due to realistic limitations (project period and budget). Such dictionaries provide sign language users with the spoken language equivalents of sign language and sign language learners with sign language-related information. This satisfies two objectives simultaneously, but there are limitations in providing all information with only one dictionary. Therefore, to make it more useful as a dictionary for learners, two separate dictionaries with different objectives should be compiled. The dictionary planned by the National Institute of Korean Language is dual dictionaries - 'Korean sign language-Korean language dictionary' supports Korean sign language users for expression and Korean sign language learners for understanding. Meanwhile, 'Korean language-Korean sign language dictionary' is to support Korean language learners for understanding and Korean sign language learners for expression.

- Dictionary with evidence

The 6 dictionaries published in Korea collected materials in a similar method. First, a list of Korean entries was selected and Korean sign language equivalents were collected or created if there were none. Second, sign language words were sampled from existing sign language textbooks or dictionaries. The collected sign language words were included in a dictionary, but users could not tell which words were artificially created and which were actually used by the deaf. This has led to more confusion and undermined the credibility of a dictionary. Therefore, the National Institute of Korean Language is to compile a dictionary that describes sign language and extract information such as entries based on a corpus.

One important thing in defining a sign language is to observe how native signer who use the sign language as the first language use the sign language concerned in context (Johnson, 1998). Therefore, the new dictionary is to provide examples so as to teach adequate usage of a word to the users. Examples will be divided into adapted examples that are extracted from a corpus then partially processed and editorial examples that are intuitively created by the deaf. Examples will have their original sources marked so that the users can refer to them.

- Acceptance of variations

According to Johnston(2003), lexical variation (expression of a same meaning in completely different sign language gestures by sign language users) is a major issue for some sign language dictionary compilers. However, it has been proposed that standardization of sign language that aims to eliminate the variation is unrealistic due to a number of reasons. Historically, sign language standardization based on a spoken language

Session 2

has been disregarded by the deaf community not only in Korea but in other countries. As such, it is more desirable to accept and document all variations rather than remove them to ensure that a dictionary can be a comprehensive resource for sign language.

○ Introduction of SignWriting

Sign language is a language without a writing system, which makes it more difficult to publish a dictionary. Compilers of sign language dictionaries are approaching this issue in various ways to overcome the difficulty. According to Kristoffersen & Troelsgård (2010), the existing sign language dictionaries use different ways to present sign language entries such as video, picture, photo, transcript of form (HanNoSys, SignWriting, etc.), gloss, number, translation and written description of how a sign language gesture is created. In most cases, these dictionaries use gloss, which is a spoken language word that represents the meaning of a sign and is used to remember sign language.

Assigning a spoken language word to a sign language entry as if a label offers a great convenience to both compilers and users, but it can cause a misunderstanding that the sign language entry concerned is completely identical to the meaning of the spoken language word or that the sign language entry only means one particular meaning out of the various meanings of the spoken language word. Also, approaching a sign language based on the label of a spoken language makes learners recognize the sign language through the spoken language which impedes sign language learning.

For accurate description of sign language and visual recognition of the description by learners, the new Korean sign language dictionary will be using SignWriting to transcribe the sign language. It is already being used in many countries including the US, Australia, Belgium, Brazil and Canada¹¹⁾, and the Czech sign language dictionary¹²⁾ presents entries in SignWriting. SignWriting will be introduced for the first time in Korea's national project and it will lay a foundation for a wide range of information created in sign language to be documented and passed down.

○ Explanation, not translation

When Korean sign language was not recognized as a language, Korean language words were paired with the sign language lexicon for one-on-one translation due to the lack of understanding. However, as mentioned in 'Development of a Korean language – sign language video dictionary for Korean language education of students with hearing impairment (Yoon Jeom-ryong and others, 2009)', Sign language, which is a visual language, has different grammatical characteristics from Korean language, which is a spoken language. Therefore, one-on-one translation of the lexicon can not fully convey the meaning of the sign language. In order to avoid such errors, the

11) <http://www.signwriting.org/about/who/>

12) <http://www.dictio.info/czj?action=page&page=front&lang=en>

13) Establishment and modification project of sign language definitions and examples in Korean sign language

National Institute of Korean Language translated the definition of Korean language into Korean sign language in 2017 as a trial¹³⁾ and established a guidance after studying documents and collecting opinions from experts. However, the result showed that Korean sign language was strongly led by Korean language in translation, and there were some cases where sign language users could not understand translation due to a significant disparity in the ways Korean language and Korean sign language explain meanings. Based on this experience, the new dictionary will contain definitions that are explanations, not translations, so that sign language users can intuitively understand definitions in the dictionary. Also, it will ensure that definitions and examples are provided coherently, rather than separately, in order to faithfully follow the way the deaf present explanation with examples and enhance its accessibility and usability for sign language users.

4. Conclusion

This article has looked into the status of Korean sign language dictionary compilation and proposed its development direction. The Korean sign language dictionaries that have been published up to date are not equipped with the appropriate frame as true language dictionaries. However, the National Institute of Korean Language has established a foundation for relevant policies and research since the enactment of the <Korean Sign Language Act> (2016). The institute is in need of a large-scale manpower to materialize these plans, but the deaf community has not been able to generate a sufficient number of researchers or make investment in fostering researchers. Due to such limitations, the project to compile a dictionary may come across some difficulties in the beginning. Nevertheless, continued fostering of researchers in the future will allow us to take a step closer to the dictionary that embodies the real language of the deaf for coexistence of the deaf and hearing people.

Reference

- Kim Sam-chan, Lee Seon-ho & 36 others (2005), 《Korean Sign Language Dictionary》, Hyungseul Publishing Co.
- Kim Seung-kuk (1993), 《Standard Sign Language》, Osung Publishing House, Seoul.
- National Institute of Korean Language(2014), 《Korean Sign Language Research Advisory Committee Report》, National Institute of Korean Language.
- National Institute of Korean Language(2016), 《Creation of Korean sign language dictionary based on hand orientation》, National Institute of Korean Language.

Session 2

- National Institute of Korean Language(2017), 《Establishment and modification of definitions and examples of sign language in Korean sign language dictionaries》, National Institute of Korean Language.
- National Institute of Korean Language, Korea Association of the Deaf (2005), 《Korean Sign Language Dictionary》
- Kim Seung-kuk (1983), 《Sign Language Dictionary》, Dankook Press
- Yoon Jeom-ryong, Jeong Hee-seon, Kim Seon-young, Won Seong-ok, Kim Ji-sook, Jang Eun-sook, Kim Kyeong-jin, Lim Jini, Ha Gil-jong & Yang Kook-jong (2009), 《Development of a Korean language – sign language video dictionary for Korean language education of students with hearing impairment》, National Research Foundation of Korea.
- Lee Jeong-ok & Lee Jun-woo (2005), An Exploratory Study on the Idiomatic Expressions of Korean Sign Language, 《Special Education Journal: Theory and Practice》 Vol. 6, No. 3, Research Institute of the Korea Special Education, p.335-354
- Lee Jun-woo, Nam Ki-hyun & Cho Jun-mo (2018), A Proposal for the Direction of Korean Sign Language Compilation, 《Special Education Journal: Theory and Practice》 Vol. 19, No. 1, Research Institute of the Korea Special Education, p.137-156
- Lee Hyun-hwa (2017), Korean Sign Language Alignment Project : Korean Sign Language Dictionary (Documentation of Korean Sign Language), 《Sae kugo saenghwal》, Vol. 27, No. 2, p.67-83
- Choi Kyeong-bong (2014), The First Korean Language Dictionary in Manuscript, 《Malmoi》, 《Sae kugo saenghwal》Vol. 24, No. 3, National Institute of Korean Language, 20-37
- Standard Sign Language Dictionary Compilation Committee (1982), 《Standard Sign Language Dictionary》, Kumok Research and Culture Foundation
- Heo Il (2003), A Study on Challenges in Compiling Language Dictionary for the Korean Deaf, 5th Symposium of Korea Sign Language Society
- Atkins, B.T.S. & M. Rundell. (2008), The Oxford Guide to Practical Lexicography, Oxford/New York: Oxford University Press.
- Cokely, D, & Baker-Shenk, K, C. (1980), American Sign Language: A Teacher's Resource Text on grammar and Culture, Gallaudet University Press.
- Johnston, T.(1998). Signs of Australia. A new dictionary of Auslan(Australian Sign Language)(the sign language of the Australian deaf community), North Rocks, NSW, Australia: North Rocks Press, Royal Institute for Deaf and Blind Children.
- Johnston, T.(2003). Language Standardization and signed Language Dictionaries, Sign Language Studies, vol 3, Published by Gallaudet University Press, 431-468.

Current Status and Development Direction of Korean Sign Language Dictionary | Hyewon Choi, Hyunhwa Lee

Kristoffersen, J. H., & Troelsgård, T. (2010), Compiling a Sign Language Dictionary: Some of the Problems faced by the Sign Language Lexicographer In M. Mertzani(Ed.), Sign Language Teaching and Learning: Papers from 1st Symposium in Applied Sign Linguistics(Applied Sign Linguistics), Bristol: University of Bristol, Centre for Deaf Studies.

Roland Pfau & Markus Steinbach & Bencie Woll(2012), Sign Language: An International Handbook, de Gruyter Mouton

R.R.K. Hartmann (1983), 《Lexicography : principles and practice》(translated by Seo Tae-gil, Kim Yang-jin, Doh Won-young, Lee Sang-hyuk, Kwon Oh-hee). Seoul:J&C.(Wonseo Publishing Co. 1983)

National Institute of Korean Language's Korean Sign Language Dictionary
(<http://sldict.korean.go.kr/front/main/main.do>)

SignWriting website (<http://www.signwriting.org/about/who/>)

Encyclopedia of Korean Culture (<http://encykorea.aks.ac.kr/>)

Czech sign language dictionary
(<http://www.dictio.info/czj?action=page&page=front&lang=en>)

2018 국립국어원 국제 학술 대회

보이는 언어의 기록, 수어사전

Records of Visible Language:
Sign Language Dictionary

주제 5 | 미래의 수어사전

토마스 항케
독일, 함부르크 대학교

Thomas Hanke
Hamburg University, Germany

3부 발표

미래의 수어사전

토마스 항케
독일, 함부르크 대학교

서론

본 발표가 수어사전에 적용되는 어떠한 사이버 기술 또는 공상과학에 나올 법한 내용에 관해 소개할 것으로 예측했다면 실망할 수 있을 것이다. 대신 수어사전 설계 단계에서 현재의 상황을 발전시킬 수 있는 사용 가능한 기술 및 곧 적용할 수 있을 미래의 기술에 대해 소개하고자 한다.

본 주제는 현재 필자가 직접 참여하고 있는 프로젝트를 기준으로 작성하게 되었다. 필자는 2009년부터 독일 함부르크 과학 및 인문 학회(Academy of Sciences and Humanities in Hamburg)의 독일 수어(DGS) 말뭉치 구축을 위한 장기 연구 프로젝트를 진행하고 있으며 연구 기간은 2023년까지로 말뭉치 중심의 독일 수어사전을 편찬할 예정이다. 현재 6년 정도의 연구기간이 남았다.

따라서 본 발표에서 다루고자 하는 ‘미래’ 핵심기술은 ‘컴퓨터 비전’(Computer Vision)과 ‘아바타 기술’(Avatar Technology)이다. 미래의 수어사전과 관련성이 있다고 판단되는 그 외 기술은 이미 사용이 가능하며 적용만 하면 된다.

필자가 진행하고 있는 프로젝트를 소개한 후 ‘컴퓨터 비전’과 ‘아바타 기술’이 수어사전에 있어 필수 기술이라고 생각하는 이유 및 적용 방법에 대해 설명하도록 한다. 그리고 눈으로 확인할 수 있는 사전을 제공하기 위해 기존의 ‘표준 웹 기술’과 ‘컴퓨터 비전/아바타 기술’을 결합하는 방법에 관해 설명하는 것으로 발표를 맺고자 한다.

DGS-KORPUS: 말뭉치에서 사전으로, 그 긴 여정에 관하여

DGS-Korpus는 함부르크 과학 및 인문 학회와 함부르크 대학교의 장기 프로젝트이다. 연구 기간은 2009년에 시작해 2023까지 진행될 예정이다. 이 연구 프로젝트의 목적은 두 가지로, 독일 수어(DGS) 참조 말뭉치 구축 및 구축된 말뭉치 데이터를 기반으로 한 현대 DGS 사전 편찬이다. 또한 프로젝트의 모든

단계에서 농 공동체(language community)의 참여를 유도하는 것이 중요한 관심사이다.

수어를 위한 언어학적 말뭉치를 구축할 때 언어제공자(informant)로부터 데이터를 수집하고 주석을 다는 방법 외에는 현재 다른 방법이 없다고 판단된다. 인터넷에서 수집할 수 있는 DGS 영상데이터가 방대하고 일부는 공영방송사에서 제공하지만 이러한 출처로부터 데이터를 추출하여 수어사용자 및 수어가 사용되는 문맥을 대표할 수 있는 표본 데이터로 활용하기에는 거의 불가능하며 법적 및 윤리적으로 어려움이 많다.

데이터 수집 지역

과거 프로젝트 경험을 기반으로 독일 전역에 걸쳐 여러 지역에 이동식 스튜디오를 마련하기로 결정했다. 이는 각 촬영 지역과 해당 지역 출신의 참여자를 통해 ‘현지’의 분위기를 담는 동시에 전사에 필요한 고품질의 촬영을 하기 위해 내린 결정이었다. 물론 촬영 지역으로 선정된 장소는 앞서 언급한 ‘현지성’과 언어제공자들의 이동 시간 및 여행계획 모두를 고려하여 결정했다.

이에 따라, 현재 및 이전 청각장애 학교들의 통학 범위, 교육 환경 등을 결정하는 주(州) 경계(연방주, Bundesland), 특히 옛 동·서독 국경, 방언경계선으로 예상되는 경계뿐만 아니라 촬영 장소까지의 이동 시간 등 현실적 요소를 고려하여 13곳의 데이터 수집 지역을 선정했다. 선정 지역은 언어제공자 모집을 위해 또다시 최대 5개 소지역으로 구분했다. 대도시권은 도시·농촌 혼합지역 또는 농촌지역과 달리 그 자체로 소지역으로 구분했다.

농인 인구에 관한 인구조사가 충분치 않아 지역별 언어제공자 수(및 소지역별 분포) 목표를 해당 지역 전체 인구 수 기준으로 결정했으며 입증되지는 않았으나 농인들이 대도시에서 거주하는 것을 선호한다는 경험을 바탕으로 대도시에는 2배 적용하여 결정했다. 또한 각 지역별로 최소 16명(남녀 최소 각 2명*4개 연령 그룹)의 언어제공자 모집 기준을 적용하여 총 328명의 목표 참여자 수를 산출했다. 실제로는 330명의 참여자를 촬영했다.

언어제공자 선별기준

수어 말뭉치 구축에 이상적인 언어제공자는 누구인가라는 질문은 여전히 답을 찾지 못한 문제이다. 특히 농인 부모를 둔 농인 자녀(2세대 이상)만을 대상으로 할지 혹은 이 집단을 적어도 우선시할 것인지 하는 문제 또한 정해진 것은 없다. 연구계 내에서 심도 깊은 논의 후, 본 프로젝트에서는 주관적으로나마 수어를 의사소통의 주요 수단으로 사용하는 사람들을 대상으로 하기로 결정했다. 흥미로운 점은 이 정의(이후 언어제공자 모집 절차에 적용된 기준)에서 청각장애를 언급하지 않는다는 것이다. 말뭉치 구축에 참여한 한 언어제공자는 앞서 언급한 기준을 충족하는 건청(健聽)인이다.

지역별 및 방언별 변이에 유의하면서 우리가 정한 언어제공자 선별기준 중 최우선 기준은 바로 ‘현지성’(rootedness)이었다. 해당 지역에서 최소 10년 이상 거주한 사람을 선별했으며 같은 지역에서 태어나 자란 사람들을 우선시했다. 추가 기준으로는 사회적 지위(주로 ‘화이트칼라’ 사무직 종사자 및 ‘블루칼라’ 생산

3부 발표

직·서비스직 구분)와 수어 습득 연령이 있다. 수어 습득 연령의 경우, 특정 나이를 기준으로 적용 시 고령 참여자에서 문제가 발생하는 것을 발견했다.

언어제공자 모집

각 지역별로 현지 농인(‘연락담당자’, 지역별 1-2명)에게 자신의 개인 네트워크 및 농인 모임 등을 활용하여 1차적 선별 기준을 충족하는 언어제공자 후보 모집을 요청했다. 실제 그 지역에서 필요한 언어제공자 수보다 최소 30% 더 많은 후보를 모집했다. 그 다음 함부르크팀이 후보 명단에서 참여자를 선별했다. 후보자가 연락 담당자와 함께 작성한 설문지, 연락 담당자 추천을 고려하여 선별했으며 동시에 2차적 선별 기준과 연령 그룹, 성별 및 소지역 간 균형을 고려하여 결정했다. 한 지역 내 분포를 고르게 하기 어려운 경우 지역 간에 걸쳐 균형을 이루도록 보완했다.

사전 동의 및 메타데이터

프로젝트 종료 후에도 데이터를 보존하는 것이 필수 목적이자 동시에 현재 및 미래 저작권, 데이터 보호 권리 등 법적 요구 사항 및 윤리 기준을 충족하는 말뭉치 구축 프로젝트를 위한 연구 참여 동의서 마련이 어렵다는 점은 이미 널리 알려진 사실이다. 특히 익명성이 불가능한 수어 말뭉치 연구에서는 더욱 어렵다.

연구 초기 단계부터 동료 연구자들과 법률 전문가와 함께 논의했다. 이 논의를 바탕으로 데이터 수집에서 미성년자는 참여자에서 제외하기로 결정했다.

저작권법 법률 전문가의 도움으로 사전 동의서를 마련했으며 동의서는 독일어 및 DGS로 작성됐다. 선별자들에게는 촬영 일자보다 훨씬 이른 시기에 사전 동의서를 서면 및 DVD 형태로 제공하여 가정에서 논의할 수 있도록 했으며 또한 현장에서 연락 담당자와 동의서에 대해 의논할 수 있는 상담 시간을 진행한 후 동의서 내용을 충분히 이해했음을 확인하는 서면 서명 및 영상 서명을 요청했다.

또한 데이터 수집에 참여 요청을 받은 대상자는 각자 집에서 또는 촬영 시작 전 연락 담당자와 함께 2차 설문지를 작성했다.

유도 과제(ELICITATION TASKS)

언어제공자들에게 20가지의 과제를 요청했다. 과제 구성방식은 스토리텔링(수어, 사진 또는 영상 형태의 프롬프트), 주어진 주제에 관한 토론 및 자유로운 대화 등의 형태를 취했다. 꼼꼼하게 계획한 과제로 다양한 형태의 조합을 재미있게 구성하여 총 5시간의 촬영 시간에도 불구하고 대부분의 참여자들은 즐겁게 촬영했다.

이전의 구어 및 수어 말뭉치 프로젝트에서 사용했던 과제도 포함하여 비교 연구를 진행할 수 있는 구간을 마련했으며 새로운 형태의 과제도 포함했다. 향후 데이터 수집에도 활용하기 위해 새로 개발한 유도 자료

전부를 공개하지는 않았으나 요청 시 연구자들에게 제공 가능하다.

영상 촬영

이동식 스튜디오에는 2명의 언어제공자 및 사회자(대부분의 경우 연락 담당자)와의 데이터 수집 세션을 HD로 촬영할 수 있는 장비를 구비했으며 장소를 이전 및 유지하기 쉽고 임대한 장소에 맞춰 준비했다(스튜디오를 촬영 장소로 이송하여 해당 장소에서 서비스 제공업체가 설치함). 대부분의 촬영 장소에 광대역 인터넷 연결이 어려워 촬영 데이터는 특수 서류가방에 하드디스크 형태로 함부르크로 발송됐다.

각 지역별 프로젝트 연락 담당자는 언어제공자 모집뿐만 아니라 사회자로서 촬영 세션을 진행하는 방법에 관해 심도 있는 교육을 받았다. 지역별 언어제공자가 선별된 후 연락 담당자들의 첫 업무는 각 촬영일별로 언어제공자 그룹을 정하는 것이었다. 쌍으로 정해진 두 언어제공자는 되도록 동일한 소지역 출신으로 구성했다. 이렇게 선별된 언어제공자 그룹은 대부분 같은 연령 그룹 및 성별로 구성되었다.

사전에 언어제공자들에게 단색 의상을 입도록 요청했다. 이를 잊고 온 참여자들에게는 검은색 티셔츠를 제공했다. 촬영 당시 제공된 티셔츠는 사용 여부와 관계없이 가져갈 수 있도록 선물로 제공했다. 언어제공자들에게 촬영이 총 7시간 소요될 것으로 안내했다. 두 번의 휴식시간과 넉넉한 점심시간을 제외하고 약 5 - 5.5 시간의 촬영이 진행됐다. 촬영 프로그램을 변경해야 하는 경우도 발생했다(한차례 단전 발생, 폭설 또는 더위로 인한 대중교통 운행 지연). 언어제공자 한 명의 건강상의 이유로 촬영 일정을 이틀에 걸쳐 진행한 경우는 한 건 발생했다.

검토 절차

촬영 후 언어제공자들에게 각자 본인의 촬영 영상을 DVD 형태로 TV 또는 컴퓨터로 확인할 수 있도록 제공했다. 촬영 영상에 타입코드를 입력 프로젝트 또는 출판본에서 제외하고 싶은 부분을 적을 수 있도록 했다.

우리가 예상했던 것보다 언어제공자들이 콘텐츠를 이유로 제외를 요청했던 부분은 적었다.(오히려 언어제공자들보다 연구진에서 더 많은 부분을 출판본에서 삭제했다.) 몇몇 참여자들은 상당 부분에 대한 삭제를 요청했으나 주로 코를 긁는 행위 등과 같은 행위가 삭제 요청 사유였다.

총 330명 중 3명은 검토 서류를 제출하지 않아 동의를 철회한 것으로 처리될 예정이다.

말뭉치 주석(CORPUS ANNOTATION)

수어에 관한 기계 판독이 가능한 참조 말뭉치를 구축하기 위해서는 기본 주석으로 분할(segmentation) 및 표제화(lemmatisation)뿐만 아니라 문어로 번역(현실적인 이유로)하는 작업이 필요하다.

3부 발표

번역

다른 말뭉치 프로젝트 경험에 따라 번역을 먼저 진행하고 다음에 토큰(token) 분할 및 표제어 추출 작업을 진행한다. 호주 수어 말뭉치 프로젝트(Auslan Corpus Project)와 같은 여타 프로젝트와는 달리 의역을 추구하지 않았다. 대신 번역은 가능한 한 DGS 발화(원천 언어)에 가깝게 진행했다. 이러한 수어 텍스트 말모듬(chunk)은 결속성을 갖고 이해할 수 있는 발화 단위를 구성해야 한다. 발화 단위 구분('chunking' 또는 경계 표지 'boundary marker')을 위해 사용되는 신호는 수어에서 사용하는 '머리 끄덕이기', '몸 움직이기', '응시 또는 리듬을 사용하는' 등이 있다.

이렇게 작업된 짧은 독일어 번역문은 DGS 영상에 맞춰 배치되며 이는 몇 가지 역할을 한다. DGS를 모르는 이들도 수어 텍스트 내용을 이해할 수 있으며 DGS 이해 수준이 다양한 주석자 역할을 하는 학생(청인)들이 표제어 추출하는데 도움을 준다. 또한 검색이 가능하며 특정 수어 토큰이 사용된 문맥을 검색할 때의 발화 단위를 결정한다.

DGS 및 독일어를 모르는 연구자들이 본 말뭉치 데이터를 활용할 수 있도록 독일어 번역문은 다시 영어로 번역되었다.

분할 및 표제어 추출

수어 자동인식 프로그램 또는 도구가 없는 실정하기에 토큰화 및 표제어 작업은 여전히 주석자들에 의해 수동으로 진행되는 시간 소모가 큰 과정이다. 표제어 작업 전 수어 텍스트를 분할해야 한다. 즉, 각 수어의 발화 시간의 범위(time span)를 시작점과 종결점이 있는 토큰 표식(tag)으로 구분해야 한다. 이 작업은 광의적 또는 협의적 방법으로 할 수 있다. 광의적 방법이란 하나의 수어 토큰이 시작되는 지점이 바로 직전의 수어의 종결점이라는 것을 의미한다. 이 방법은 기호 간 중간 동작이 있다는 사실을 무시한다. 즉, 한가지 기호 유형의 토큰을 확인할 경우 시각적 노이즈가 존재한다. 수어 자동인식을 위해 머신러닝 모델을 학습시킬 때 매우 중요한 요소이다. 따라서 우리는 각 기호 토큰별로 시작점과 종결점을 표시하는 협의적 의미의 표식을 적용했다.

분할작업은 구어 말뭉치 작업에서의 토큰화에 해당된다. 어휘화(lexicalisation) 정도와 관련된 번역 차용과 합성어 간 구분 등에 대한 몇 가지 아직 풀리지 않은 문제들이 있다.

본 프로젝트에서의 주석 작업 환경에서 토큰이란 단순히 시간에 맞춰 배치된 일련의 텍스트 나열이 아닌(HamNoSys를 사용한) 유형(type)의 형태에 대한 정보를 담고 있거나 대체하는 유형과 연결된 데이터베이스 개체이다. 수어 어휘 주석(gloss)은 유형 표지 및 일부 텍스트의 토큰과 연결하는 표식 역할을 하나 구조적 연관성은 없다. Johnston의 확인용 어휘 주석 ID-gloss 개념은 기저의 데이터베이스 모델에서 이미 그 기능을 하기 때문에 이 맥락에서는 필요치 않다. 유형-토큰 매칭 결정은 확인하고자 하는 토큰 영상과 동일한 유형의 다른 토큰을 보여주는 비디오를 비교함으로써 쉽게 확인할 수 있다.

번역 및 표제어 작업 외에도 주석 작성 첫번째 단계에서 입 모양(mouthing) 주석 작업과 관찰 가능한 마

우스 제스처(mouth gesture)를 표시하기로 결정했다. 이런 결정의 배경에는 입 모양이 DGS 기호 토큰의 의미에 중요한 단서이자 기호의 형태와 함께 토큰을 매칭할 기호 하위 유형을 검색할 때 사용할 수 있기 때문이다. 또한 입 모양은 관습화된 수어(conventional sign)와 생산적인 수어(productive sign)를 구분할 때 사용된다. 입 모양은 소문자로 주석을 단다. 실제 발화는 무시하고 입술 모양에서 읽을 수 있는 의도했던 어휘(어간)를 주석 처리한다. 현재는 마우스 제스처의 경우, 마우스 티어(tier)에 '[MG]'를 더하는 단순한 방법으로 처리하고 있다. 마우스 제스처를 구분할 수 있는 '머신러닝 시스템'을 개발하기 위해 박사과정 프로젝트를 진행 중이다.

공개 말뭉치

50시간의 영상자료와 함께 공개 말뭉치를 제공하는 것은 본 프로젝트의 주요 목표 중 하나이다. 두 개의 대상 그룹인 언어학 및 기타 학문 연구자들과 '농 공동체'는 각각 '니즈(Needs)'가 다르기에 이를 충족하기 위해 우리는 두 개의 별도 포털을 제공한다. 출판된 자료는 '농 공동체'의 문화유산에 중요한 의의를 지닌다. 사람들은 데이터를 검색하면서 2차 세계대전 및 베를린 장벽 붕괴 당시 목격자들의 증언, 농인들의 학교생활을 포함한 개인적인 삶에 대한 이야기 등 재미있는 이야기 혹은 나라 전역에 걸쳐 다양한 방식의 수어를 확인할 수 있다. DGS를 유창하게 구사하지 못하더라도 농인들의 삶에 대해 알고 싶어하는 이용자 또는 수어를 배우는 사람들을 위해 독일어 자막도 함께 제공한다.

언어학적 관점 혹은 다른 목적으로 말뭉치를 살피고자 하는 연구자들을 위해서는 기본 주석과 함께 영상을 제공하고 독일어-영어 번역문도 함께 제공한다.

총 550시간 중 공개 말뭉치에 담을 50시간을 선별할 때 몇 가지 기준을 사용했다. 참여했던 모든 언어제공자들이 공개 말뭉치에 포함될 수 있도록 했으나 주로 스토리텔링과 같은 과제보다 자유로운 대화 형태의 영상을 담고자 했다. 스토리텔링 영상의 경우, 일부분을 공개하여 이 과제에 관심있는 사람들이 전체 영상을 확인하기 위해 연구진 측과 계약을 맺기 전 샘플로 활용할 수 있도록 했다. 자신을 소개하고 수어 이름을 설명하는 과제는 언어제공자들의 개인정보를 공개하지 않는 원칙에 상충되기에 공개 말뭉치에서는 제외되었다.

공개 말뭉치를 선택하는 과정에 있어 언어제공자들이 제시한 출판 제한사항을 존중할 뿐만 아니라 동시에 데이터 작업 시 전사자들이 한 표시도 반영했다. 이렇게 우리 측에서 제외한 자료의 주된 이유는 언어제공자들이 스스로 데이터를 검토하면서 문제 삼지 않았더라도 '수어 공동체'와 같이 작은 공동체에서 쉽게 누구인지 파악할 수 있는 제3자에 대한 내용과 미래에 부정적으로 영향을 미칠 수 있는 언어제공자 자신에 대한 매우 개인적인 정보 등을 담았기 때문이다. 이는 주로 법에 저촉되는 행위 및 직장에서의 부적절한 행위 등을 포함한다. 단, 정치적으로 부적합한 대화를 제외하고자 한 목적은 없었다.

곧 공개할 말뭉치는 농 공동체 포털의 경우 거의 50시간 분량, 언어학적 연구 목적의 데이터 포털의 경우 30시간 이상의 영상과 함께 공개될 예정이다.

두 포털은 영상에 적용되는 라이선스와 출판된 주석 데이터에 대한 설명자료를 함께 제공한다. 물론 라이

3부 발표

선스 내용이 언어 제공자들의 연구 참여 동의서상 선택한 사항을 반영하나 언어제공자들이 영상에 드러날 수밖에 없다는 사실을 감안하고 있다. 독일법에 따라 다른 말뭉치 프로젝트에 비해 본 영상물에 대한 라이선스가 더 제약적이다. 라이선스에서 허용하는 공개 말뭉치의 데이터 이상으로 데이터를 사용하고자 하는 연구자들은 함부르크 학회 측에 연락하여 계약 체결 후 희망하는 목적에 따라 데이터를 사용할 수 있다.

농 공동체의 참여

우리의 주요 고려사항 중 하나는 농 공동체와 협력하고 바로 자신의 언어에 대한 연구가 진행됨에 있어 참여하고 영향을 미칠 수 있도록 하는 것이다. 수어 공동체 사람들은 우리 프로젝트의 여러 단계에 걸쳐 참여했다. 데이터 수집 시 언어제공자 뿐만 아니라 프로젝트팀 구성원(직원 혹은 학생 보조), 연락 담당자, 사전 사용에 대한 이용자 설문조사 참여자, 웹 기반 DGS 피드백 제공자 및 포커스 그룹(Focus Group) 구성원 등의 형태로 참여했다.

포커스 그룹은 농 공동체에 소속된 11명의 농인으로 구성되어 있다. 농 공동체 내에서 프로젝트, 피드백 및 결과물을 홍보하며 프로젝트에 기여한다. 또한 개별 기호 및 기호의 사용에 대한 정보를 제공하여 프로젝트에 도움을 준다. 포커스 그룹은 우리에게 농 공동체의 관점을 알 수 있는 기회가 되어 생각을 공유할 수 있는 중요한 역할을 한다. 매년 열리는 포커스 그룹 총회에서 말뭉치에서 직접 뽑은 예시, 후보 표제어의 형태 및 내용, 레이아웃 기능 등 사전의 여러 측면에 대해 논의해왔다. 일례로, 포커스 그룹에서 논의한 내용을 토대로 사전 표제어에 기술하고 있는 각 의미별로 용례를 한 가지 이상 수록하여 이용자들이 서면 정보에 의존하는 대신 자신의 수어 능력을 사용하여 의미를 구별할 수 있도록 하자는 의견에 따라 이를 반영했다.

말뭉치의 미래 활용

DGS-Korpus가 함부르크와 그 외 지역에서 다수의 프로젝트가 진행될 수 있도록 근간이 되어 이 장기 사업을 통해 개발한 말뭉치 및 어휘적 자원이 충분히 연구되기를 바란다. 프로젝트 형태는 학사, 석사 학위 연구에서부터 다년간의 제3자 후원 프로젝트 등 다양할 수 있다. DGS-Korpus가 DGS에 대한 연구에 있어 중요한 자원이 되기를 바란다.

말뭉치 기반의 사전편찬

상당한 규모의 표제화된 말뭉치 없이는 사전편찬자들은 모어 사용자들의 직관력과 수어 설명에 관한 참여자 관찰에만 의존할 수밖에 없을 것이다. 상당 규모의 맥락이 있는 유창한 수어 사용을 검토 및 분석하지 못하는 상황에서 수어의 의미 및 용법에 대한 구체적이고 신뢰할 수 있는 설명을 하기란 거의 불가능에 가깝다. 바로 이것이 말뭉치의 힘이다.

이제 활용 가능한 표제화 DGS 말뭉치 데이터의 풍부한 규모 및 범위를 토대로 우리의 사전편찬 작업을 다음 단계로 발전시킬 수 있다. 적어도 사용빈도가 높은 수어의 경우, 모든 중요한 기호 정보를 실제 맥락에서 사용되는 기호의 용법 및 의미를 분석한 내용을 토대로 작성할 수 있게 되었다. 또한 구어 말뭉치 기반 사전편찬 과정에서 적용되는 방법을 수어사전에도 적용할 수 있을 것이다.

발생빈도를 기준으로 표제어 선정 시 말뭉치 데이터를 활용한다. 우리의 사전편찬을 위한 표제화 원칙('표제어 구축establishment of lemmas')에 따라 표제어의 범위가 결정된다. 그 다음, 표제어를 작성하기 위해 필요한 수어에 관한 모든 정보를 준비하기 위해 관련 말뭉치 데이터를 검토 및 분석한다.

사전 표제어는 수어 형태, 변이형, 문법 정보 및 관련 있는 경우 어휘의 분포 정보에 대한 전반적인 정보를 담고 있다. 동형이의어 상호 참조, 연관 수어 및 유사한 형태의 수어도 주어진다. 추가 설명 부분에서는 해당 수어를 포함하여 자주 발생하는 합성어 유사한 구성 및 관용표현을 기술한다.

사전 표제어의 핵심은 의미 나열 형태로 수어의 뜻과 용법을 설명하는 것이다. 뜻과 용법을 분석하기 위해 등재 후보 수어의 상당히 많은 양의 토큰 및 맥락을 검토하여 맥락적 의미 및 기타 관련 사실을 분석한다. 그 다음 유사한 용법 사례를 묶어 하나의 의미로 설명한다. 각 의미별로 그 의미를 파악, 설명 및 묘사하기 위해 여러 종류의 정보를 제공하고 그 외 관련 유용한 정보를 제공한다. 독일어로 기술한 사전학적 정의는 해당 수어의 의미를 설명하고 '공동 의미'(co-sense)와 구별한다. 말뭉치에서 인용한 예문은 의미를 더욱 설명한다. 어휘 변이, 동의어, 유사하거나 연관된 의미를 갖고 있는 수어 및 반의어를 기술하고 상호 참조한다. 이는 독일어로 작성한 정의를 읽고 이해할 필요 없이 특정 의미를 이해하는데 추가적으로 도움이 된다. 일반적인 입 모양 및 연어 패턴 등 기타 정보 또한 제공한다.

활용할 수 있는 말뭉치가 있다는 것은 우리에게 수어에 대한 새롭고 구체적인 통찰력을 제공하고 기호 사용에 대한 깊은 분석 및 정의가 가능해진다. 그 결과 우리의 사전 표제어에는 수어 및 기술된 의미에 대한 높은 수준의 정보와 다양한 범위의 정보를 담고 있다. 따라서 한 항목에서 한 수어에 관한 모든 가용한 정보를 포괄적으로 제공함에 따라 표제어 구성이 복잡한 편이다. 현재 단계에서는 정보를 글로스 사용 없이 제공하는 방법을 모색하고 있다. 대신 기호 형태는 마이콘(micon, "움직이는 아이콘moving icon") 형태와 마우스를 올려놓으면 미니어처 영상을 재생하는 썸네일(thumbnail) 형태로 나타내고 있다.

컴퓨터 비전: 이용자가 무엇을 검색하는가?

독일어 전통적 사전에서 표제어를 찾기 위해서는 이용자는 단어(적어도 앞부분이라도) 표기의 철자를 알고 알파벳 순서를 알아야 했다. 이렇듯 표제어를 찾기 위해 상당한 지식이 필요한데 알파벳 외 문자 체계를 사용하는 언어 혹은 문자 체계가 없는 언어의 경우 더욱 어려워진다. 수어의 경우 정해진 문어 체계가 없다. 사전편찬자들은 다른 (구어)언어에 의존하기를 원치 않아 수어의 음성학적 변수를 사용하여 사전 표제어 순서를 정했었다. 예를 들어, 수위 및 수형의 순서로 배치했던 것이다. 하지만 이용자들은 사전편찬자들 만큼 수형 또는 수위에 따른 분류라는 개념에 익숙하지 않아 사전을 사용하는데 또 한 가지 장벽이 생기는

3부 발표

것이다. 이 관점에서 전자사전도 놀랄 만큼의, 약간의 개선안을 더했다. 사용된 문자 체계의 알파벳 순서를 알 필요는 없지만 그 외 모든 문제점은 그대로 존재했다. 전반적으로 사전을 사용하는 방법을 아는 사람들이 적어지면서 사전편찬자들은 전통적인 접근방법으로 수어사전을 구축하기에는 이용자들이 이러한 ‘접근’을 위한 기술을 배워야 하는 등 어려움이 존재한다.

바로 여기서 ‘예시 검색(Search by example)’이 상당히 간단한 대안이 될 수 있다. 컴퓨터 카메라에 찾고자 하는 수어를 표현하는 것이다(사전은 비싸고 사용하기 어려운 기술 없이 언제 어디서든 사용 가능해야 하기에 사이버 슈트(cyber suit) 또는 데이터 글러브(data glove)는 고려하지 않는다.). 사실 2011년에 EU 프로젝트 ‘DictaSign’를 통해 이러한 인터페이스 시제품을 개발한 적이 있다. 카메라를 향해 수어를 표현하면 시스템은 수어 동작을 하는 아바타 형태로 가장 근접한 기호들의 명단을 보여준다. 이 명단(되도록 짧으면 하는) 중에서 찾고자 했던 기호를 클릭할 수 있다. 그 당시 시제품이 220 기호 기준으로 잘 작동되었으나 규모를 늘리기는 어려웠다. 또한 키넥트(Kinect) 센서가 다른 유형의 웹캠 또는 휴대폰 카메라로 대체된다면 더 개선될 것이다. 아직 개선의 여지가 있다는 것을 의미한다.

최근에는 미국 카네기 멜런 대학교(Carnegie Mellon University)에서 개발한 머신러닝 프레임워크인 ‘오픈 포즈(OpenPose)’의 활용 가능성을 검토하기 시작했다. 이름에서 시사하듯 영상 이미지로 사람 수와 무관하게 포즈(동작)를 인식하는 것이다. 영상 속 인물을 추적하지는 않으나 프레임별로 재인식한다. 간단한 신체 골격과 관절의 위치를 2D 형태로 보여주며 각 포즈가 얼마나 정확한지를 알려주는 신뢰 수치를 제공한다. 이 라이브러리의 최근 버전은 얼굴 및 손을 더욱 구체적으로 인식할 수 있다(처리시간은 더욱 길어졌다). 우선 오픈 포즈를 활용하여 적은 양의 데이터를 실험해본 결과 가능성이 있어 보인다. 수어 영상 처리에 관한 지식을 활용한다면 오픈 포즈 데이터를 토대로 3D정보도 재현할 수 있을 것이다. 오픈 포즈에 추가 지도학습(supervised learning) 층위를 더하고 학습에 사용하는 수단으로 주식 말뭉치 데이터를 활용하는 것이다. 과거의 접근방법에 비해 이 방법이 규모를 확장하기 수월할 것이라 판단되지만 데이터처리가 완료될 때까지 기다린 후에 공식 평가를 할 수 있을 것이다.

이러한 진전이 있음에도 불구하고 이용자가 달성하고자 하는 것에 따라 ‘예시 검색’의 대안 방법도 모색해야 할 필요가 있다는 점을 잊지 말아야 할 것이다. 한 가지 시나리오를 들어보자. 모르는 수어를 본 이용자가 있다. 불행히도 이 이용자는 수어를 구체적으로 보지 못했다(혹은 보았지만 일부 잊었다.). 이를 해결하기 위한 접근 방법은 두 가지가 있다. 첫번째는 이용자가 디폴트 ‘데이터’를 추가하여 수어를 완성하고 이를 카메라에 표현한 후 보기 중 제외하고자 하는 파라미터를 선택하는 것이다. 또 다른 방법으로는 찾고자 하는 수형, 수위, 움직임 등을 시스템에 전달하는 유저 인터페이스가 있어 이용자는 확실한 정보만을 입력하고 더 세부적인 유저 인터페이스를 사용하여 시스템이 총체적 접근을 채택할 때 보다 더 많은 디테일을 인식할 수 있도록 하는 것이다. 예를 들어, 웹캠을 통해 상체를 캡처하는 방법보다 더 구체적으로 수형을 인식하는 렙 모션(Leap Motion) 카메라를 활용하는 시스템 몇 가지가 개발 중이다. DictaSign에도 수위를 구체화하는 시제품이 있었다.

아바타 기술: 사전이 당신에게 수어하게 하라!

독자 모두가 수어 아바타를 본 적이 있을 것이다. 사실 전세계적으로 진행되고 있는 수어 아바타 프로젝트가 굉장히 많아 엄청난 양의 수동 작업을 통해 조율하지 않는 이상 이해할 수 있고 믿을 수 있는 수어 동작을 할 수 있는 아바타가 많지 않다는 사실이 놀라울 정도이다. 모든 수어 표제어를 수동으로 조율하는 것이 가능할 수 있겠지만 우리는 아바타가 말뭉치의 예시를 수어 동작으로 보여주는 것에 관심을 갖고 있다. 물론 이 경우 수동 조율은 불가능하다. 자동적 수어 표현은 모션 캡처 데이터 또는 HamNoSys의 음성학적 데이터 등을 토대로 한 합성 생성(synthetic generation)으로 재현한다. 첫 번째 방법은 특수 장비 및 상당한 후처리 시간이 소요된다. 또한 한 명의 혹은 여러 명의 수어 모션 캡처 데이터를 어떻게 결합할 것인지도 불명확하다. 합성 생성 방법은 어휘 데이터베이스에 있는 음성학적 데이터를 토대로 이루어져 큰 노력 없이 가능하다. 단, 표기법에 따른 수어는 로봇이 수어하는 느낌이 있다. 필자의 의견은 하이브리드 방식이 가장 잠재력이 있다. 사용 가능한 경우 표준 모션 캡처 데이터를 사용하나 필요한 경우 합성 생성으로 구현하는 것이다. 일반적인 지속시간(duration) 또는 실제 지속시간 등과 같은 말뭉치 데이터를 활용할 수 있으며(앞으로는 오픈 포즈 데이터도 활용하여) 조금 더 사람이 하는 수어 동작에 가까워질 수 있을 것이다.

시카고 드폴 대학교(DePaul University)의 폴라 프로젝트(Paula Project)에 폴라 아바타가 하이브리드 데이터를 기반으로 DGS 수어동작을 표현할 수 있도록 참여하고 있다.

이 시점에서 높은 품질의 영상이 있는데 말뭉치의 예문을 아바타가 보여주는 방식을 채택하고자 하는 이유에 대해 궁금해질 수 있다. 사전에 예시로 사용할 수 있는 공개 말뭉치 상의 좋은 수어 데이터를 보유하고 있다는 점을 자랑스럽게 생각한다. 하지만 말뭉치 규모가 제한적이라는 점을 감안했을 때(최근 문어 말뭉치에 비해 제한적인 규모), 타당한 이유로 공개 말뭉치에서는 제외됐지만 제외된 예시문도 제공하고 싶은 것이다.(개인의 삶 이야기를 많이 포함하고 있어 본 연구 말뭉치에서는 종종 발생한다.) 제3자에 대한 내용이라면 수어 모델을 활용하여 언어제공자가 스튜디오에서 한 수어를 따라하고 제3자 언급에 대한 내용은 대체하는 방식을 채택할 수도 있다. 하지만 자신의 수어가 아닌 수어 영상이 인터넷상에 공유되기 때문에 타인의 수어를 그대로 따라 하는 것을 불편해하는 수어 모델들도 있다. 바로 이러한 경우에 아바타가 완벽한 익명성을 보장하는 전략이 된다.

언어 패턴도 또 하나의 잠재 응용 분야이다. 모든 패턴 요소가 단 한 명의 수어를 기준으로 할 가능성은 거의 없다. 여러 사람이 언어 패턴을 표현하는 것은 이용자를 불편하게 할 수 있다. 따라서 마찬가지로 이 분야에서 수어 모델 또는 아바타에 의존해야 한다.

말뭉치 예문 대신 만들어진 예문을 사용하는 사전의 경우, 사전 편찬자는 아바타로 하여금 이용자에게 맞춤형 방언 변이형으로 수어 동작을 하도록 하여 사전을 맞춤화 할 수도 있다.

15년 전 디지털 영상 대신 아바타를 사용하면서 절감할 수 있는 광대역을 이유로 들어 아바타를 옹호하는 근거가 되었을 수 있다. 이제는 대부분의 공동체에서 이 주장은 더 이상 근거가 없지만 일부의 경우에는 아직 유효할 수 있다. 사전편찬자들에게 선택의 여지가 있다.

3부 발표

사전의 시각적 탐색(VISUAL EXPLORATION)

기호 형태를 매우 정확하게 검색하는 것이 가능해진다 하더라도 이용자들이 단일의 표제어를 찾는 목적 이상으로 사전의 콘텐츠를 탐색하고 활용하는 것이 사전편찬자들의 꿈일 것이다. 과거 종이사전 시대에는 찾고자 하는 표제어를 찾기 위해 종이를 뒤적이다가 자연스럽게 여러 항목을 확인하게 되고 그 중 일부가 시선을 끌었을 것이다. 오늘날 사전편찬자들은 이용자들이 다른 표제어도 찾아보도록 유도하기 위해서는 다른 표제어에 상호 참조를 제공하거나 다수의 표제어를 제공해야 한다. ASL-LEX 프로젝트는 잘 작성된 그래프와 함께 굉장히 매력적인 콘텐츠 시각화를 구현했다. 하지만 이 방법이 모든 사전에 적합한 것은 아니다. 우리가 목표했던 크기의 사전에 맞춰 적용했을 때는 ASL-LEX 프로젝트에 비해 시선을 끌지 못했다. 따라서 오늘 논의하는 사전 프로젝트별로 각각 다른 해결책이 필요할 것이다.

우리가 구축한 DGS 말뭉치에는 DGS에서 발견되는 변이형(지역별 및 연령별)의 상당수를 반영한다. 피드백에 따르면 지역별 변이형은 DGS 공동체에게 큰 관심사이며 우리가 맵(map)으로써 제공할 수 있는 데이터에 이용자들의 만족도가 높은 것으로 나타났다.

이때 잘 구현된 맵을 작성하기 위해 d3js 표준 웹 기술을 적용했다.

결론

본 발표를 마무리하기 전에 미래의 수어사전의 모습은 어떨지에 대해 전망해보도록 하자. 잭 인 더 박스(Jack-in-the-box) 장난감을 기억하는가? 그 아이디어를 수어에 적용해보자. 다시 말해서, 아바타를 통해 수어를 확인하고 싶을 때 박스를 열어 바로 확인하는 것이다. 대신 박스 속 잭(Jack)이 아닌 가상의(혹은 실존의) 수어사용자의 홀로그램 디스플레이인 것이다. 그리고 박스일 필요가 없다. 삼성 20h 또는 iPhone 20xh를 책상에 얹어 놓고 홀로그램 수신기를 켜고 있다고 상상해 보라.

수어통역이 필요한 사람 옆에 프로젝터를 활용하여 수어통역사를 띄우는 모습을 상상해 보라. 미술관에서 안내하고 있는 작품 바로 옆에 가상의 안내원이 있는 모습을 상상해 보라. 만약 홀로그램이 너무 ‘스타워즈’와 같은 것으로 들린다면 증강현실(AR)을 떠올리면 될 것이다. 특수 안경을 통해 볼 수 있도록 컴퓨터가 생성한 콘텐츠를 현실세계에 이미지로 투사할 수 있도록 하는 새로운 기술이다. 물론 우리가 원하는 증강현실 시나리오 속 가상의 미술관 안내원은 전시회 안내 동안 발화만이 아닌 수어로 안내하는 시나리오일 것이다! 안내원이 사용한 수어 중에 이해하지 못한 기호가 있다면 간단하게 눈을 깜빡여 바로 사전의 관련 페이지로 연결되는 모습을 상상해 보라. 마치 오늘날 킨들(Kindle)로 외국도서를 읽을 때처럼 말이다.

추신, ‘사전 페이지’라고 했던가? 사전은 종이사전에서 하이퍼텍스트(hypertext) 형태로 변했다. 미래 사전의 모습을 상상할 때는 페이지 개념을 무시해야 할지도 모른다. 가상의 3D 세상에서는 사전 표제어가 마치 방이 여러 개 있고 이웃과 연결되는 문이 많은 아파트와도 같은 구성일지도 모른다. 경청해 주셔서 고맙습니다.

SIGN LANGUAGE DICTIONARIES OF TOMORROW

Thomas Hanke
Hamburg University, Germany

INTRODUCTION

If you expect this talk to be on cybertechnology or any kind of science fiction applied to sign language dictionaries, I have to disappoint you. Instead, this talk focuses on technologies already available or becoming available in the very near future that would help us to evolve the current state of the art in sign language dictionary design.

This focus is motivated by the project I am currently involved in: I am leading a long-term research project of the Academy of Sciences in Hamburg which started in 2009 with collecting a corpus of DGS (German Sign Language) and shall end in 2023 with a corpus-oriented dictionary of DGS, i.e. in slightly more than six years.

So the “future” core technologies I will cover here are Computer Vision and Avatar Technology. Any other technologies that I see relevant for future sign language dictionaries are already available today and “only” need to be applied.

After giving a bit of background on our project, I will outline why I consider Computer Vision and Avatar Technology essential for sign language dictionaries and how they fit. Finally, I will give some ideas how these technologies combine with standard web technologies to provide visually explorable dictionaries.

DGS-KORPUS: THE LONG JOURNEY FROM CORPUS TO DICTIONARY

DGS-Korpus is a long-term project of the Academy of Sciences and Humanities in Hamburg and Hamburg University. It started in 2009 and is scheduled to end in 2023. The project has two major objectives: To build a reference corpus of German Sign Language (DGS) language data and to publish a dictionary of contemporary DGS based on the corpus data. Beyond the output, it is a major concern of the project to involve the language community in all phases of the project.

Session 3

We believe that when building a linguistic corpus for a sign language, there currently is no way around collecting data from informants and to annotate these. While there is a huge number of DGS video data available on the internet and some from public broadcasters, it is practically impossible and both legally and ethically very difficult to sample data from these sources to be representative for the language with respect to the users of the language and contexts in which the language is used.

DATA COLLECTION REGIONS

From experiences in earlier projects, it was one of the key decisions to have a mobile studio to be set up in different places across Germany. The idea was to have quite a “local” spirit with the recording location in the region and all persons involved coming from that region while still ensuring high-quality recordings needed for transcription. Obviously, the number of locations selected for recordings needs to be a compromise between localness in the above sense, but also relevant for the informants’ travel times, and the logistics.

Our solution was the definition of thirteen data collection regions, trying to respect the catchment areas of current and former Schools for the Deaf, state (Bundesland) borders determining a. o. educational settings, especially the former border between West and East Germany, suspected dialectal borders, but also practical considerations such as travel times to the recording locations. The regions were further subdivided into up to five sub-regions relevant for informant selection. Large metropolitan areas form their own sub-regions, in contrast to others with mixed or more rural structures.

Due to the lack of census data on the Deaf population, the target number of informants per region (and a suggestion for the distribution onto the sub-regions) was determined from the population figures of the community at large, with a weight of 2 for larger cities to reflect (unproven) experience that Deaf people often prefer to live in larger cities. Together with a set minimum of 16 informants per region (to cover four age groups times two sexes with at least two informants each), this resulted in a target number of 328 participants. We actually filmed 330 participants.

INFORMANT SELECTION CRITERIA

It was and still is an open question who would be the ideal informants for a corpus of a sign language – this is especially true for the question whether one should exclusively target Deaf of Deaf (2nd or higher generation deaf) or at least prioritise this group. After intensive discussions within the research community, our decision was to consider people who (subjectively) use sign language as their primary means of communication. Interestingly, this definition (that was then implemented by the recruiting procedures) does not even mention deafness, and in fact we have one hearing informant in the corpus who completely satisfies the criteria defined.

With an eye on regional/dialectal variation, rootedness in the region was set a high-priority criterion: People

needed to be resident in a region since at least 10 years, with a preference for people born and raised at the same place. Social background (primarily just white vs. blue collar) and sign language acquisition age were additional criteria. In the latter case, it turned out that any fixed cut-off age would have turned out problematic especially for elderly people.

INFORMANT RECRUITING

For each region, we contracted local deaf persons (“contact persons”, one or two per region) who used their personal networks as well as events in Deaf clubs etc. to recruit informant candidates fulfilling the primary criteria, at least 30% more people than actually needed for each region. It was then the team in Hamburg to decide whom to invite from the list of candidates. These decisions were based on a questionnaire filled in by the candidate together with the contact person, recommendations by the contact persons, also considering the secondary criteria and the balance between age groups, sexes and sub-regions. What could not be balanced within a region was then compensated across regions.

INFORMED CONSENT & METADATA

It is well known by now how difficult it is to set up an informed consent for participation in a linguistic corpus project where it is an integral aim to preserve the data beyond the end of the project, satisfying legal requirements such as current and upcoming copyright, data protection right etc. as well as ethical standards. This is even more true for sign language corpora where anonymisation is close to impossible.

We had started discussions with colleagues and legal experts early on. This led us to the decision not to include minors in the data collection.

Our informed consent was developed in cooperation with a legal expert on copyright law und formulated in both German and DGS. The invitees received the informed consent on paper and on DVD for consultation at home well before their recording day and had time on-site to discuss the informed consent with the contact person before being asked to sign the paper and sign into the camera that they had signed the contract fully understanding what it meant.

People invited to participate in the data collection also received a second questionnaire to be filled out at home or together with the contact person right before the recording sessions started.

ELICITATION TASKS

We used a set of 20 different tasks for the informant. The formats ranged from story-retelling (with prompts in sign, picture, or movie) to discussions on a given topic as well as free conversations. With careful planning, it

Session 3

was possible to make the mix of formats diverting enough that most participants enjoyed the recording session despite a net length of 5 hours.

The set contains a number of tasks previously used in other corpus projects, both on spoken and sign languages, to lay a basis for cross-linguistic research, as well as new formats. Not all details of the newly developed elicitation materials are available in the publications in order to keep the material suitable for future data collections. The materials are, however, available to other researchers upon request.

RECORDING

The mobile studio was equipped to record the data collection sessions with two informants and a moderator (in most cases the contact person) in HD quality while remaining easy to relocate and maintain, and adaptable to the rooms rented. (The studio was transported to the recording locations and built up there by a contracted service provider.) As most recording locations did not offer broadband internet connections, data was shipped to Hamburg on hard disk sets in special suitcases.

The project's contact persons in the region had not only received extensive training on informant recruitment, but also on how to conduct the recording sessions as moderators. Their first task after the informants for the region had been selected was to arrange pairs of informants for each recording day. Preferably the two persons should be from the same sub-region when possible. The majority of pairs filmed are same-age group and same-sex.

Informants had been asked beforehand to wear single-colour clothing. Where people had forgotten this request, we offered them black T-shirts. (This was a gadget to take home whether they had used it in the recording session or not.)

Informants were asked to plan in 7 hours on-site. This included two coffee breaks and a longer lunch break, leaving 5 to 5.5 hours for the recording. In some cases, the recording programme had to be adapted (one power outage, public transport delays due to snowfall and heat). There was just one case where the recording session had to be split into two days due to the health condition of one of the informants.

REVIEW PROCEDURE

After the recordings, informants received “their” recordings as a set of video DVDs to be viewed on a TV set or a computer. The recordings were overlaid with timecodes, allowing the informants to write down time spans from the recordings they wanted to be excluded from any usage in the project or just from publication.

Informants marked fewer spans to be excluded for contents reasons than we had expected (and in fact we exclude more passages from publication than the informants themselves). Some people did mark quite a number of time spans, but they were more concerned about things like scratching their noses.

Three out of 330 people did not return their review sheets, now to be treated as if they had retracted their

informed consent.

CORPUS ANNOTATION

In order to build a machine-readable reference corpus of sign language, segmentation and lemmatisation as well as, for pragmatic reasons, translation into a written language are minimally required as the core of any basic annotation.

TRANSLATION

Following experiences from other corpus projects, we start with the translation first, then segment and lemmatise tokens. Unlike other projects e.g. the Auslan Corpus Project, we do not aim for a free translation. Instead, the translation should be as close to the DGS utterance (source language) as possible. These chunks of signed text should form a coherent and comprehensible meaning or utterance unit. Further cues for chunking (or ‘boundary markers’) are signals inherent to sign languages like head nod, body shift, gaze or rhythm.

The more or less short written German sentences are aligned with the DGS video and serve several functions: They give access to the content of the signed texts even to those not knowing DGS and guide our (hearing) student annotators – with a varying competence of DGS – when lemmatising. In addition, they are searchable and define preliminary utterance units when looking for the context in which a sign token is used.

In order to make the corpus data accessible to researchers knowing neither DGS nor German, the German translations are also translated into English.

SEGMENTATION AND LEMMATISATION

With no automatic sign recognition software or tools at hand, tokenisation and lemmatisation still remain a time consuming step done by human annotators. Before lemmatising, the signed text has to be segmented, i.e. the time span of each articulation of a sign must be determined by creating token tags with start and end points. This can be done in a broad or narrow way. The broad way means that one defines the starting point of a sign token being the ending point of the previous sign or vice versa. This approach ignores the fact that there are transitional movements between signs, i.e. when checking sign tokens of one sign type there is more visual noise. This is crucial when trying to train machine-learning models for automatic sign recognition. We therefore opted for a narrow segmentation in defining start and end for each sign token.

Segmentation roughly corresponds to tokenisation in vocal language corpus processing. There are a number of open questions such as the distinction between loan translation and compounds, closely related to the degree of lexicalisation.

In our annotation environment, tokens are not mere time-aligned text strings, but database entities linked to a

Session 3

type, inheriting or superseding the type's form information (using HamNoSys). Glosses are labels for types and linked tokens convenient in some contexts, but they have no structural relevance. Johnston's notion of ID-glosses is superfluous in this context as its function is already guaranteed by the underlying database model. Decisions in type-token matching can be easily verified by comparing the token video in question with videos showing other tokens of the same type.

Besides translation and lemmatisation, we decided to also annotate mouthings – and give an indication of observable mouth gestures – in the first phase of annotation. The reason is that mouthings are an important clue to the meaning of a DGS sign token which, in combination with the sign form, can be used to search for the appropriate sign sub-type the token should be matched to. Furthermore, mouthings are used to distinguish between conventional and productive uses of sign types. Mouthings are annotated in lower case. Disregarding the actual articulation, at least the intended word (word stem) to be lip-read is annotated. For the time being, mouth gestures are annotated in a simplistic way by just adding “[MG]” in the mouth tier, as we have a PhD project running to develop a machine learning system to tell mouth gestures apart.

PUBLIC CORPUS

A public corpus with 50 hours of video is one of the key deliverables of the project. In order to best serve two target groups with rather different needs, i.e. researchers from linguistics and other disciplines on the one hand and the language community on the other, we provide two separate portals: For the language community, the published material is an important contribution to cultural heritage. People are invited to browse through the data to find interesting stories like eye-witness reports from WW II to the fall of the Berlin wall, personal stories from Deaf life, especially school experience, or just different signing styles from across the country. For people interested in learning more about Deaf life without being fluent in DGS or for people learning sign language, German subtitles are provided.

For researchers interested in exploring the corpus linguistically or otherwise, the videos are provided together with basic annotation as well as translation into German and English.

There are a number of criteria we used to identify the 50 out of 550 hours to become part of the public corpus. While we wanted everyone who participated to be visible in the public corpus, we strongly preferred relatively free conversations over tasks such as story-retelling. Of the latter category, we only included a couple of examples so that people interested in these tasks can view an example before entering an agreement with us to have access to the full set. The tasks on presenting oneself and explaining one's sign name was completely excluded from the public corpus as it conflicts with the rule not to make informants' contact data public.

The public corpus selection obviously respects the publication restrictions as set up by the informants, but also marks set by our transcribers when working on the data. Primary reasons for such exclusions from our side were conversations about third persons who could be easily identified from the content within a community as small as the sign language community, very private detail as well as facts about the informants themselves that

might shed a bad light on them in the future, even if the informants themselves obviously did not consider these passages problematic when reviewing the data. This includes conflicts with law and inappropriate behaviour at the working place. It was not our aim, however, to exclude politically incorrect talk.

The upcoming new release of the public corpus will almost reach the 50 hours for the language community portal and more than 30 hours for the linguistic research data portal.

The portals also explain the licences applicable to the video and annotation data published. Of course, the licenses respect the options the informants chose in their informed consent, but they also take into account that the informants are inevitably visible in the videos. Following the constraints of German legislation, these licenses are more restrictive than those found with other corpus projects. Researchers interested to use data beyond what is possible with the licenses on the public corpus are of course invited to contact us and arrange contractual agreements hopefully allowing their intended uses.

INVOLVEMENT OF THE LANGUAGE COMMUNITY

One of our key concerns is to cooperate with the language community and to make sure that it has some influence on and part in the research on their language. Members of the signing community are involved at many levels of our project: They are not only informants in the data collection, but also members of the project team (as staff and student assistants), contact persons, participants of the user survey on dictionary use, participants of the web-based DGS-Feedback and members of the Focus Group.

The Focus Group currently consists of eleven Deaf members who are well rooted in the Deaf Community. They support the project within the Deaf Community by advertising the project, the feedback and our products. Also, they support our work in answering questions on individual signs and their uses. The group is an important place for us to discuss some ideas because it makes the Language Community's perspectives on relevant issues available to us. In yearly meetings with the focus group we have discussed many aspects of the dictionary such as examples taken directly from the corpus, form and content of preliminary entries as well as layout functions. One decision following a discussion in the focus group was to generally include more than one example for each sense displayed in a dictionary entry to enable users to differentiate senses using their sign language competence rather than having to rely on written information.

FUTURE USE OF THE CORPUS

DGS-Korpus hopefully lays the ground for a multitude of satellite projects, in Hamburg and elsewhere, fully exploring the corpus and lexical resources developed in this long-term enterprise. These projects may range from BA and MA theses projects up to multi-year third-party funded projects. We expect the DGS-Korpus to be an important language resource for generations of research on DGS.

Session 3

CORPUS-BASED LEXICOGRAPHY

Without a lemmatised corpus of a considerable size, dictionary makers have to rely solely on native speakers' intuition and participant observation for the description of signs. It is at least very difficult if not almost impossible to achieve a detailed and reliable description of the signs' meanings and uses without the possibility of reviewing and analysing a considerable amount of natural fluent signing in context – that is corpus evidence.

With the substantial size and scope of the lemmatized DGS Corpus data now available, we are finally able to take our lexicographic work to the next level. At least for signs of higher frequency we can now base all crucial sign entry information on analyses of actual uses and meanings of the signs in context and we can start to employ methods well established in corpus-based lexicography of spoken languages.

The corpus data is used for lemma selection on the basis of frequency. Following our lemmatisation rules for the dictionary ('establishment of lemmas'), the scope of an entry is defined. Then relevant corpus data are reviewed and analysed to prepare all information on the signs relevant for writing an entry.

Dictionary entries contain general information on the sign form and its variants and some grammatical and, if relevant, distributional information. Cross-references to homonyms, related sign and signs of similar form are given. In the run-on section, frequent compound-like constructions and idiomatic phrases that contain the sign are listed.

The core of a dictionary entry is a description of the attested meanings and uses of a sign in the form of a list of dictionary senses. For the analysis of the meanings and uses, we review a substantial number of tokens of each sign candidate and their contexts and take notes on their contextual meanings and other relevant facts. Similar instances of use are then grouped together and described as individual senses. For each sense several kinds of information are given to identify, describe and illustrate this particular sense and to provide other useful information relating to it. A lexicographic definition written in German explains the given sense and distinguishes it from its co-senses. Example sentences taken directly from the corpus illustrate the sense. We list and cross-reference lexical variants, synonyms and signs of similar and related meanings as well as antonyms. This provides an additional approach to grasp a particular sense without reading and understanding a definition written in German. Other information types like typical mouthings and (collocation) patterns are also provided.

Having a corpus to work with provides us with new and more detailed insights on signs and allows for more detailed analyses and description of their uses. Consequently, our dictionary entries contain a high degree of detail and a large variety of different kinds of information on the signs and their listed senses. Therefore, the entry structure is rather complex as it provides a comprehensive view of all the available information on one sign in one place. At the present stage we are experimenting with ways to present the information structure without using any glosses. Instead, forms are represented by micons ("moving icons"), thumbnails that play a miniature movie showing the sign as soon as the user hovers over the thumbnail.

COMPUTER VISION: WHAT IS THE USER SEARCHING FOR?

In order to find an entry in a traditional dictionary of German for example, the user has to know how to spell the word's citation form (at least the beginning), and has to know the sorting order of the Latin alphabet. This is quite a lot of knowledge just to find an entry, but it gets worse with languages for a non-alphabetic writing system or even no writing system at all. For sign languages, there are no established writing systems. Dictionary makers not wanting to rely on another (spoken) language often used selected sign language phonetic parameters to order the entries in a dictionary, such as by location height and handshape. But as users not used to the level of abstraction used by the dictionary maker in grouping handshapes or locations, this means another hurdle to access the dictionary. Electronic dictionaries are only a surprisingly small step forward in this respect: While they eliminate the need to know the alphabetic ordering of the writing system used, all other issues remain valid. As lexicographers in general are faced with fewer and fewer people having the knowledge how to use a dictionary, a traditional approach is not likely in the case of sign language dictionaries where people interested would need to learn these access techniques.

This is where the "Search by example" provides a seemingly straightforward alternative: You just sign the sign you are interested in into the computer's camera. (I am not considering cybersuits or datagloves as a dictionary needs to be accessible anywhere anytime without expensive or difficult-to-use technology.) In fact, we developed a prototype of such an interface in the EU project DictaSign back in 2011. You sign into a camera, the system replies with a list of the closest matches represented by an avatar signing them. This allows you to select the sign you are looking for by simply clicking onto it in a (hopefully short) list. While that prototype worked quite well with a set of 220 signs, it did not scale well. Furthermore, it would be nice if the Kinect could be replaced by any kind of webcam or phone camera. So there is quite some work to be done!

We have recently started to explore the possibilities of OpenPose, a Machine Learning framework from Carnegie Mellon University (CMU, USA). As the name implies, it is meant to recognize the poses of any number of persons in video images. It does not track persons, but re-recognizes them frame by frame. The output is 2D positions of joints in a rather simplified skeleton, together with confidence values telling you how much you can rely on each of the positions. The latest version of this library allows more detailed recognizers for the face and the hands to be added (at the cost of even longer processing times). Our experiments with a small set of data already processed with OpenPose are rather promising. Using our knowledge about the sign language videos being processed, we can even reconstruct 3D information from the OpenPose data. The idea is to build a second-layer supervised learning system on top of OpenPose, with the training set being the annotated corpus data. We are confident that this system will scale a lot better than previous approaches but of course have to wait for the data processing to be completed until we can start formal evaluation.

Despite this progress we should not forget that it is also necessary to consider alternatives to "Search by example", depending on what the user wants to achieve. Let us assume the user has seen a sign s/he has does not know. Unfortunately, the user could not see the sign in all detail (or has forgotten some detail). There are two

Session 3

approaches to cover this scenario: In the first one, the user completes the sign by adding default default “data” and signs that to the camera and then selects which parameters of the exemplar are to be ignored. Alternatively, there are user interfaces to tell the system which handshape you are looking for, what would be the location, what would be the movement trajectory. So the user would only input the detail s/he is sure about, using more specific user interfaces potentially allowing the system to recognize more detail than in a holistic approach. E.g. there are several systems being developed using Leap Motion cameras to recognize the handshape in more detail than probably possible with current webcams capturing the upper half of the body. In DictaSign there was also a prototype to specify the location.

AVATAR TECHNOLOGY: HAVE THE DICTIONARY SIGN TO YOU!

I am sure everyone in the audience has seen signing avatars before. In fact, there have been so many signing avatar projects around the world that it is surprising that most are miles away from being able to sign in an understandable and believable way except when fine-tuned by massive amounts of human resources. While it might be possible to hand-tune every headword signed, we are also interested in having the avatar sign examples from the corpus. Clearly, in that case hand-tuning is no longer an option. Automatic sign generation either relies on motion capture data or on synthetic generation, such as from HamNoSys phonetic data. The first approach requires special equipment and lots of post-processing time, and it remains unclear how to best glue signing mocap data from one single or even several people together. Synthetic generation works from phonetic data available in the lexical database anyway, so it kind of comes for free. However, signing from notation often gives the impression of a robot signing. In my eyes, hybrid approaches are most promising: They use normalized mocap data where available, but fall back to synthetic generation where necessary. Corpus data such as typical or actual durations can be used (and hopefully OpenPose data in the future) allow to come closer to human signing.

We are cooperating with the Paula Project at DePaul University in Chicago to enable the Paula avatar to sign DGS from hybrid data.

At this point, you might ask why we want to have an avatar sign example sentences from the corpus if we have good-quality videos available. In fact, we are proud to have lots of nice signing data from the public corpus that can be used as examples in the dictionary. However, given the limited size of the corpus (compared at least to recent written language corpora), we would like to also show examples not part of the public corpus for good reasons. (This happens quite often in our corpus due to the high percentage of life stories included.) If it is just reference to a third person, we might also use sign models copying the informant’s signing in the studio and replacing that third-person reference. But there are cases where the sign models feel uneasy about exactly repeating someone else’s signing as they would become visible in the internet with signing that they consider not to be their own. In that case, the avatar is a perfect anonymisation strategy.

Collocation patterns is another potential area of application: It is highly improbable that you find exemplars of

all pattern elements signed by one single person. Having the pattern signed by different persons might irritate the user. So once again you have to rely on either sign models or avatars.

For dictionaries using invented examples instead of corpus examples, avatars might allow the dictionary maker to personalize the dictionary by signing dialectal variants fitting to where the user is from.

Fifteen years ago, bandwidth saving from using an avatar instead of digital video would have been the major argument in favour of avatars. While this no longer is a point for many communities, even in the mobile world, it might still be a valid point for others. Definitely, dictionary makers should have the choice.

VISUAL EXPLORATION OF A DICTIONARY

Even if it will become feasible to implement very precise search for sign form, it remains the dictionary maker's dream that the users explore the dictionary's content beyond looking up single entries. In the days of printed dictionaries, when you had to browse through the pages to find the entry you were looking for, you automatically scanned several articles and might be caught by some. Nowadays, the dictionary maker has to explicitly offer cross references to other entries or provide a universe of entries inviting the user to explore it. The ASL-LEX project has a particularly attractive visualisation of its contents as a nicely laid-out graph. Unfortunately, their approach is not the right one for every size of a dictionary. It looks much less attractive when scaled to the dictionary size we want to end up with, so the problem probably needs to be addressed separately for each dictionary project discussed today.

Our DGS corpus reflects the high degree of variation found in DGS, both in the regional and age dimension. As it turns out through the feedback we get, regional variation is of high interest to the DGS community, and people are very satisfied with the data we can provide as maps.

Here we use standard web technologies, in this case d3js, to generate nicely drawn maps.

CONCLUDING REMARKS

Now, before closing, let us speculate about what sign language dictionaries of the day after tomorrow might look like: You remember Jack-in-the-box? Apply that idea to sign language: Whenever you need an avatar to show signs, you get it right where you need it by opening a box. But instead of Jack out of the box, it is a holographic display of a virtual (or even real) signer out of the box. And it need not be a box: Just put your Samsung 20h or iPhone 20xh face-down on the table to start the holographic transmitter...

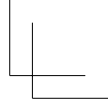
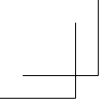
Think about interpreters projected into the room where you need them, next to the person you want to interact with. Think about the virtual museum guide side by side with the objects it is presenting. If holographic displays sound too much Star Wars to you, just replace them with Augmented Reality (AR). This is upcoming technology

Session 3

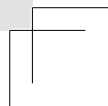
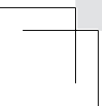
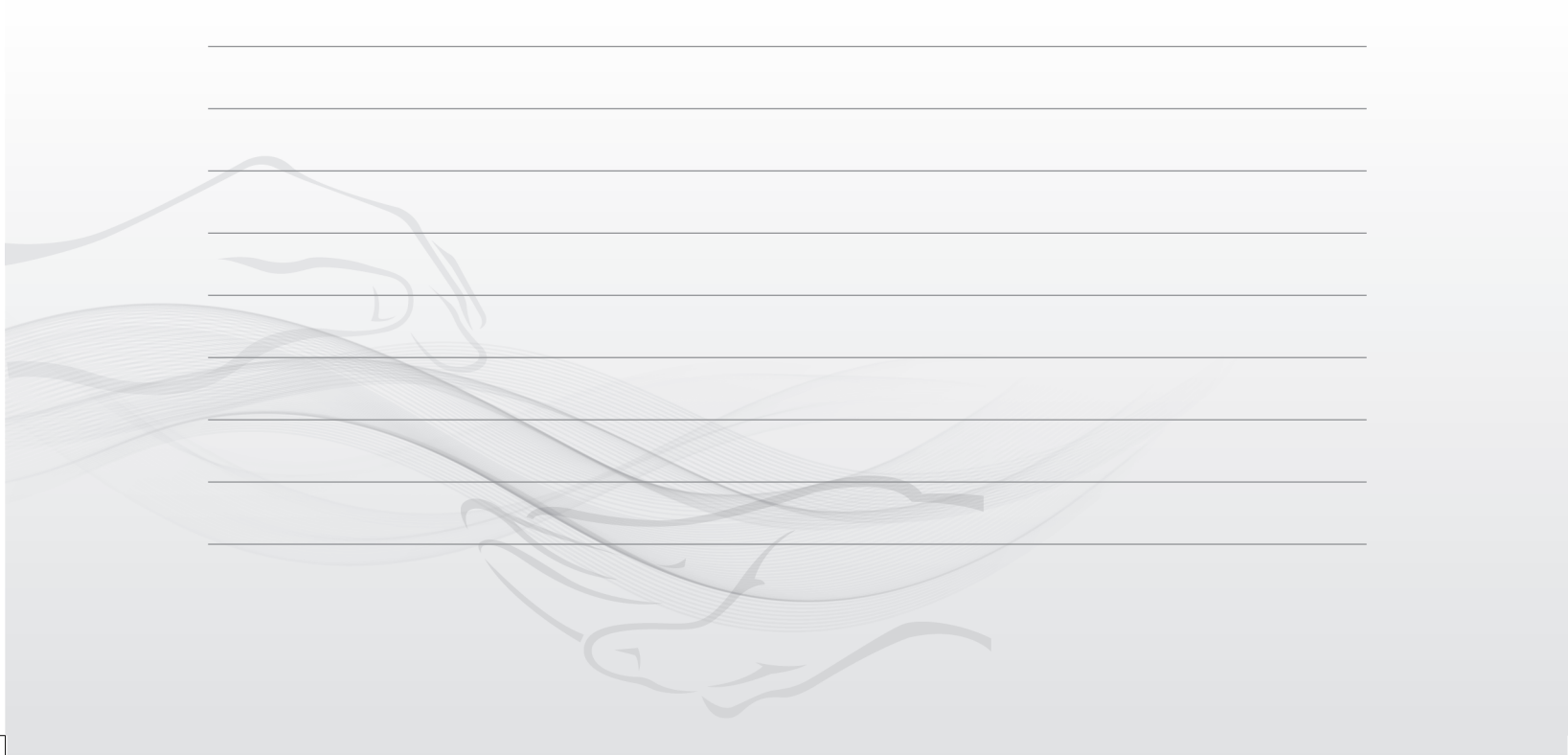
allowing apps to project computer-generated contents onto the real world images you see through special glasses. Of course we want the virtual museum guide in such an AR scenario not only being able to talk you through the exposition in speech, but also to sign you through the exposition! And then, as you do not understand a sign being used by the guide, you simply eyeblink to it in order to have the corresponding dictionary page opened, very much as you may be used to when reading a Kindle book in a foreign language today.

By the way: Did I say dictionary page? Dictionary pages somehow survived the transition from paper volumes to hypertext. Maybe we have to get rid of the concept when thinking about dictionaries for the day after tomorrow. In a virtual 3D world, a dictionary entry might actually look more like an apartment consisting of several rooms and lots of doors to the neighbours. Thank you very much for your attention!





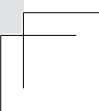
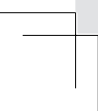
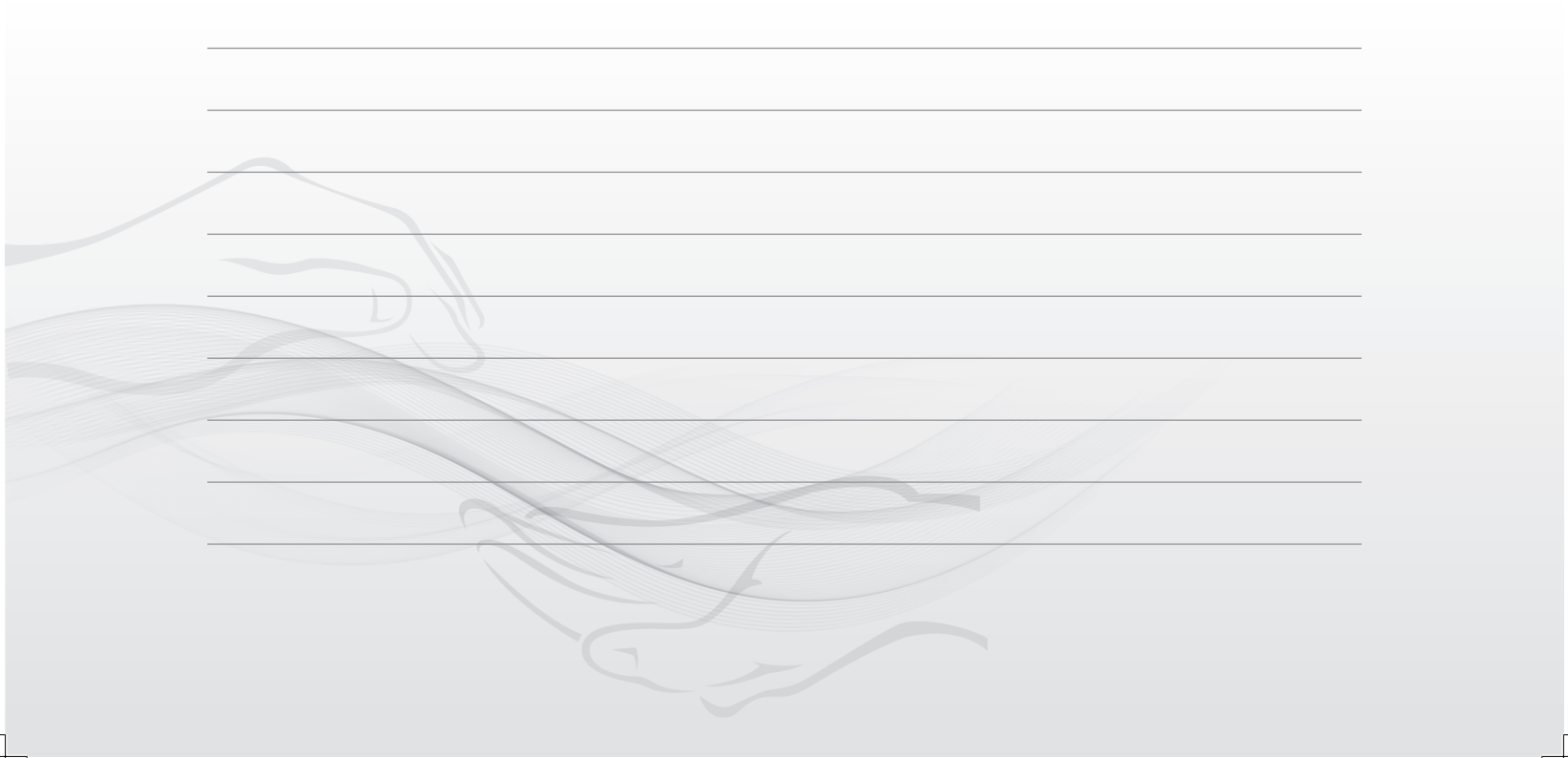
A series of horizontal lines for writing, spanning the width of the page.







Handwriting practice lines consisting of 20 horizontal lines.





2018 국제 학술 대회

보이는 언어의 기록, 수어사전

Records of Visible Language: Sign Language Dictionary

발 행 인 소강춘

기획·편집 정호성·서민경

발 행 일 2018. 11. 8.

발 행 처 국립국어원
서울시 강서구 금남화로 154
<http://www.korean.go.kr>

디 자 인 (주) 유니원커뮤니케이션즈

인 쇄 (주) 상상그래픽스

※ 저작권법 보호를 받는 간행물이므로
어떠한 형태로든 무단 전재 및 복제를 금합니다.