31 Managing Acquisition Data for Developing Large Sesotho, English, and French Corpora for CHILDES

- Class: 언어데이터과학
- Presenter: 이예은
- Date: 2024-10-29

목차

1. Introduction
2. The acquisition of morphosyntax: The Demuth Sesotho Corpus
3. The emergence of grammatical morphemes in American English: The Providence Corpus
4. The emergence of grammatical morphemes in French: The Lyon Corpus
5. Conclusions

Introduction

언어 습득 코퍼스 데이터

- 언어 습득 연구는 1800년대 찰스 다윈이 아들의 발달 과정을 기록한 일기 데이터에서 시작해, 아동 발달 연구의 일환으로 코퍼스가 사용되기 시작함
- 코퍼스는 특정 시기 아동의 인지 및 발화 양상을 포착하며, 특히 아동이 듣는 (child-directed speech) 발화 환경을 분석하는 데 유용함
- 언어 습득 코퍼스는 종단 연구(longitudinal study) 방식으로 수집되어 아동의 발달 궤적과 개인차를 장기적으로 추적하며, 실험 연구를 보완할 수 있음
- 코퍼스 설계는 연구 목적에 따라 대상 아동 수, 연령, 녹음 빈도 및 대화 상대 등이 달라질 수 있음
- 각기 다른 목적으로 수집된 코퍼스도 추후 실험 설계나 연구를 위한 파일럿 데이터로 활용하거나 새로운 연구 주제에 응용될 수 있음

Data

	Demuth Sesotho Corpus	Providence Corpus	Lyon Corpus
수집 시기	1980-1982	2002-2005	2002-2005
연구 목표	Sesotho 언어의 시제/상(tense-aspect)과 문법 구조(morphosyntax) 습득 연구	영어 음운 및 형태 초기 발달 연구	프랑스어 음운 및 형태 초기 발달 연구
자료 내용	4명의 Sesotho 모국어 아동이 가족과 상호작용하는 것을 녹음	6명의 영어 모국어 아동(1-3세)이 부모와 집에서 상호작용하는 모습을 녹음/녹화	5명의 프랑스어 모국어 아동(1-3세)이 어머니와 집에서 상호작용하는 모습을 녹음/녹화

Research Topic

- Sesotho, English, French의 언어습득 코퍼스 데이터가 어떻게 구축되고 활용되었는지

The acquisition of morphosyntax: The Demuth Sesotho Corpus

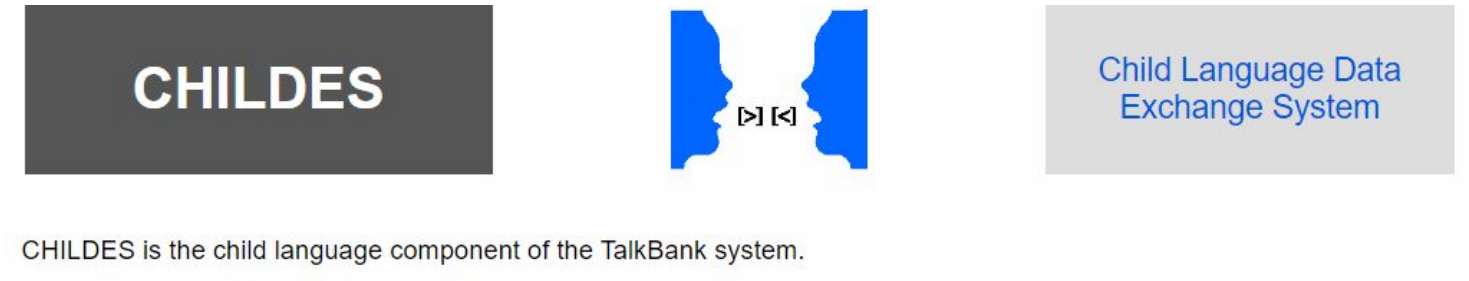
- **1980~1982년** 동안 남아프리카 **레소토(Lesotho)의 Mokhotlong 지역**에서 수집
- 반투어의 특징: 교착어, nominal modifier와 동사에서의 일치, 자유 어순, 성조 언어
- 550명이 거주하는 작은 마을에서 4명의 아동(2-4세)의 언어 발달을 종단 연구 방식으로 추적
  - **2-3세** 아동 3명과 **3-4세** 아동 1명이 가족 구성원들과 대화하는 모습을 매달 **3-4시간**씩 녹음해, 총 **98시간**의 코퍼스 데이터를 수집
- 연구의 초기 목적은 **시제와 상(tense/aspect)** 습득을 분석하는 것이었으나, 당시 의미 이론의 부족으로 이를 해결하지 못했고 이후 이론 발전과 함께 **형태통사 발달(morphosyntactic acquisition)** 연구로 확장

The acquisition of morphosyntax: The Demuth Sesotho Corpus

- 데이터의 **40%는 아동 발화**, 40%는 **성인 발화**, 20%는 **또래와 형제자매의 발화**로 구성되어, 아동 대상 발화(child-directed speech)의 특성을 파악하는 데 기여
- 1980년대에 수기로 작성된 전사본과 음성 파일은 나중에 **CHILDES 데이터베이스***에 기여하기 위해 Codes for the Human Analysis of Transcripts**를 사용해서 전산화(computerize) 및 형태론적 태깅(morphologically tag) 작업 진행
- CHILDES에 공개된 이 데이터는 반투어의 통사 구조 연구, 형태론적 구문 분석(morphological parsing), 아동의 성조 습득 등 다양한 주제를 탐구하는 데 활용 가능
- 처음에 수기로 진행된 transcription과 tagging의 정확도를 비교해 초기 연구의 신뢰성 확인 가능

cf. CHILDES(The Child Language Data Exchange System)

제1언어 습득 데이터를 저장하는 주요 레포지터리로, 1960년대부터 수집된 전사, 음성, 영상 파일이 공개되어 있음. 최근에는 더 광범위한 언어 데이터를 포함하는 TalkBank 시스템의 일부로 통합되었으며, 주로 아동의 언어와 성인의 아동 대상 발화(child-directed speech)를 연구하는 데 활용됨. 데이터베이스 내 모든 전사 파일은 CHAT 전사 매뉴얼을 따르며, 전사 및 분석 작업은 Leonid Spektor가 개발한 CLAN 프로그램을 통해 지원됨.



예시: https://sla.talkbank.org/TBB/phon/Eng-NA/Providence/Alex/010427.cha

The emergence of grammatical morphemes in American English: The Providence Corpus

- 아동이 문법 형태소를 습득하는 과정에서 형태소를 일관되지 않게(variably) 발화하는 시기가 있음이 관찰됨
- 이는 음운적 제약이나 불완전한 의미론/통사론적 지식 때문이라고 보았는데, (Brown 1973) 문법 형태소 습득 과정의 음성학적 증거가 부족했기 때문에 새로운 코퍼스 수집이 필요했음
- **6년** 동안 6명의 **1-3세 모노링구얼 (monolingual) 영어 아동**이 부모와 상호작용하는 모습을 주 1시간씩 녹음 및 녹화해 총 **364시간**의 오디오 및 비디오 데이터 수집

The emergence of grammatical morphemes in American English: The Providence Corpus

- 성인과 아동의 발화를 CHILDES transcription convention에 따라 전사
- **SAMPA(7비트 아스키 기반)*** 음성 기호체계를 사용해 데이터를 컴퓨터에서 처리 가능하도록 만들었으며, 나중에 **유니코드/IPA**로 변환
- 데이터의 **10%를 재전사**했을 때 정확도는 80~98%
- CHILDES 데이터베이스에 공개되어 다양한 연구(ex. phonological/prosodic effect on the acquisition of both inflectional morphemes (Song, Sundara & Demuth 2009), and articles (Demuth & McCullough 2009b), 모자 상호작용 연구 등) 및 언어 습득 수업 자료로 활용

The emergence of grammatical morphemes in French: The Lyon Corpus

- 단일 언어의 습득 과정에 대한 연구 결과를 일반화할 수 있는지에 대한 질문이 제기되면서, 교차 언어 연구(crosslinguistic research)의 필요성이 부각됨
- 영어와 운율적으로 매우 다른 프랑스어 코퍼스가 비교 대상으로 대두됨(영어: trochaic lexical stress, 프랑스어: phrase-final prominence)
- Providence Corpus와 최대한 유사한 방식으로 자료 수집
    - **1-3세 모노링구얼 프랑스어 아동** 5명이 가정에서 엄마와 상호작용하는 모습을 2주마다 한 시간씩 녹음해 총 **185시간**의 데이터를 확보

The emergence of grammatical morphemes in French: The Lyon Corpus

- 성인과 아동의 발화는 CHILDES 전사 규칙에 따라 전사됨
- 아동 발화 데이터는 SAMPA로 기록되었다가 유니코드 형식으로 변환
- 데이터의 10%를 재전사한 결과 정확도는 90~98%
- 프랑스어의 **자음 클러스터 습득** 등 다양한 음운론적 이슈를 탐구하는 데 활용

Conclusion

- 언어 습득 코퍼스를 수집하고 전사하며 처리하는 과정은 **노동 집약적**이지만 매우 가치 있는 작업임
- **데이터의 전산화**는 원본 데이터와 초기 분석을 비교해 연구의 **신뢰도와 정확성**을 검증할 수 있게 해주며, 추후 다른 research question을 다루는 데도 활용됨
- “the very act of transcription involves a reduction of information” (Ochs, 1979) 오디오 및 비디오 파일을 함께 첨부하면 **억양, 담화의 맥락 등 중요한 언어 정보를 보완**할 수 있음

Quiz Question

- IPA 기호에 기반하여 만들어졌고, 7비트 ASCII 문자를 사용하는 컴퓨터 처리용 음성 표기 체계는 무엇일까요? (약자로 입력해주세요)