

Morpheme Acquisition Analysis

Final Project Presentation

Sen Sub Laban

LING 2340

Dr. Na-Rae Han

Apr. 14, 2023

Contents

- Background (L1)
- Background (L2)
- Background (SOC)
- Order of acquisition
- Research questions
- Data curation
- CHAT & PyLangAcq
- Building Pandas dataframes
- Data samples
- Preliminary analysis
- Morpheme counts
- Next steps
- References
- Q&A

Background (L1)

- Morpheme order studies originated in the 1970's when researchers were looking into the "independent grammars assumption."
- There is likely a consistent order in which first/second language learners acquire proficiency in the use of **morphemes**, "minimal unit[s] of meaning," which may be lexical or grammatical.
 - "played" contains a lexical base form (play) and a grammatical morpheme (-ed) indicating past tense
- The field of natural order studies, and this project, is primarily concerned with the latter type, also known as **functors**.

Background (L2)

- The work of Brown (1973) and others was extended to second language acquisition to demonstrate that SLA is not just a matter of learned responses but that competence is actually developed according to a predictable series of benchmarks.
 - Studies in SLA supported researchers' expectations that the order of morpheme acquisition is largely consistent across language learners.
 - Also demonstrated that children and adult learners acquire morphemes in the same general order.
- Morpheme order studies became the basis for Krashen's (1985) Natural Order Hypothesis, hugely influential in the field of SLA.
- Determinants such as semantic complexity, input frequency, and native language transfer may play important roles.

How do researchers actually measure morpheme acquisition order?

Suppliance in obligatory context (SOC)

"Grammatical morphemes are obligatory in certain contexts, and so one can set an acquisition criterion not simply in terms of output, but in terms of output-where-required. Each obligatory context can be regarded as a kind of test item which the child passes by supplying the required morpheme or fails by supplying none or one that is not correct. This performance measure, the percentage of morphemes supplied in obligatory contexts, should not be dependent on the topic of conversation or the character of the interaction" (Brown, 1973, p. 255).

Order of acquisition of English morphemes in Major L1 and L2 Studies



L1 Studies		L2 Studies				
R. Brown (1973)	de Villiers and de Villiers (1973)	Dulay & Burt (1974b)	Bailey, Madden, and Krashen (1974)	Larsen-Freeman (1975)	Hakuta (1976)	Rosansky (1976)
		Children (Spanish and Chinese)	Adults (classified as Spanish and non-Spanish)	Adults (Arabic, Japanese, Persian, and Spanish)	Child (Japanese)	Children, Adolescents, Adults (Spanish)
N=3	N=21	N=60 Span. 55 Chin.	N=73	N=24	N=1	N=6
1 Pres. Prog.	2 Pres. Prog.	1 Art.	1 Pres. Prog.	1 Pres. Prog.	2 Pres. Prog.	1 Pres. Prog.
2.5 on	2 Plural	2 Copula	2 Plural	2 Copula	2 Copula	2
2.5 in	2 on	3 Prog.	3 Contr. Cop.	3 Art.	2 Aux.	3
4 Plural	4 in	4 Simple Plural	4 Art.	4 Aux.	4.5 in	4 Art.
5 Past Irreg.	5 Past Irreg.	5 Aux.	5 Past Irreg.	5 Short Plural	4.5 to	5 Copula
6 Poss.	6 Art.	6 Past Reg.	6 Poss.	6 Past Reg.	6 Past Aux.	6 Aux.
7 Uncon. Cop.	7 Poss.	7 Past Irreg.	7 Contr. Aux.	7 Sing.	7 on	7 Poss.
8 Art.	8.5 3 rd Pers. Irreg.	8 Long Plural	8 3 rd Pers. Pres.	8 Past Irreg.	8 Poss.	8 Past Irreg.
9 Past Reg.	8.5 Contr. Cop.	9 Poss.		9 Long Plural	9 Past Irreg.	9 Long Plural
10 3 rd Pers. Reg.	10.5 Contr. Cop.	10 3 rd Pers. Sing.		10 Poss.	10 Plural	10 Past Reg.
11 3 rd Pers. Irreg.	10.5 Past Reg.				11 Art.	11 3 rd Pers. Reg.
12 Uncontr. Aux.	12 Uncontr. Cop.				12 3 rd Pers. Reg.	
13 Contr. Cop.	13 Contr. Cop.				13 Past Reg.	
14 Contr. Aux.	14 Uncontr. Aux.				14 Gonna Aux.	

Adapted from Jeong, 2002.

Research questions



Does the sequence of morpheme acquisition in the data align with previously proposed acquisition orders?

How does the sequence potentially differ between the L1 and L2 corpora?

Data curation

TalkBank: CHILDES Frogs English Slobin Corpus

Native English speakers retell a wordless "frog story" picture book. Includes participants aged 3, 4, 5, 9 and 20.

TalkBank: SLABank Vercelloti Corpus

Adult learners entering an Intensive English Program (IEP) giving two-minute monologues on a given topic.

PELIC and COCA

I used these two corpora to supplement my work with learner metadata (Vercelloti is derived from PELIC) and morpheme frequency (COCA).

CHAT & PyLangAcq

- CHAT is a transcription format unique to TalkBank. Data stored hierarchically, similar to a JSON format.
- PyLangAcq is a Python library specifically developed to work with data in the CHAT format.



The CHAT Transcription Format

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

pylangacq.org

Lee, Jackson L., Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. 2016. Working with CHAT transcripts in Python. Technical report TR-2016-02, Department of Computer Science, University of Chicago.

Building Pandas dataframes

```
# initiating empty lists
file_path_list = []
participant_list = []
group_list = []
edu_list = []
tokens_list = []
pos_list = []
mor_list = []

# read entire corpus into a Reader object
Vercorpus = pylangacq.read_chat(path)
# compiling data into lists

for f in Vercorpus:
    file_path = f.file_paths()[0].split('/')[3]
    pos = []
    mor = []
    words = pylangacq.Reader.words(f)
    for token in pylangacq.Reader.tokens(f):
        pos.append(token.pos)
        mor.append(token.mor)
    for p in f.headers()[0]['Participants']:
        if re.match(r'^[0-9]{4}$', p):
            participant = p
            group = f.headers()[0]['Participants'][p]['group']
            edu = f.headers()[0]['Participants'][p]['education']
            file_path_list.append(file_path)
            participant_list.append(p)
            group_list.append(group)
            edu_list.append(edu)
            tokens_list.append(words)
            pos_list.append(pos)
            mor_list.append(mor)
```

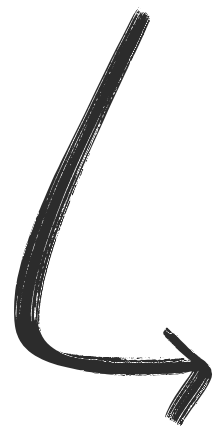
Python

```
# building the dataframe
Vercorpus = pd.DataFrame({'Filename':file_path_list,
                          'Participant':participant_list,
                          'Group':group_list,
                          'Education':edu_list,
                          'Tokens':tokens_list,
                          'POS':pos_list,
                          'Morphemes':mor_list})
```

Python

	Filename	Participant	Group	Education	Tokens	POS	Morphemes
0	Vercellotti\1060_3G1.cha	1060	1	level4	[my, topic, is, describe, your, favorite, meal...	[det:poss, n, cop, v, det:poss, adj, n, prep, ...	[my, topic, be&3S, describe, your, favorite, m...
1	Vercellotti\1060_3G2.cha	1060	2	level4	[the, topic, is, transportation, , in, this, ...	[det:art, n, cop, n, , prep, det:dem, n, qn, ...	[the, topic, be&3S, transport&dv-ATION, , in, ...
2	Vercellotti\1060_3G3.cha	1060	3	level4	[the, topic, is, someone, I, admire, , I'll, ...	[det:art, n, cop, pro:indef, pro:sub, v, , pr...	[the, topic, be&3S, someone, I, admire, , I, w...
3	Vercellotti\1060_4P1.cha	1060	1	level4	[the, topic, is, talking, about, a, problem, i...	[det:art, n, aux, part, prep, det:art, n, prep...	[the, topic, be&3S, talk-PRESP, about, a, prob...

Data samples



Data-Science-for-Linguists-2023/...

This is Sen's term project.



1

Contributor



0

Issues



0

Stars



0

Forks



Morpheme-Acquisition-Analysis/data_samples at main · Data-Science-for-Linguists-2023/Morpheme-Acquisition-Analysis

This is Sen's term project. . Contribute to Data-Science-for-Linguists-2023/Morpheme-Acquisition-Analysis development by creating an account on GitHub.

 GitHub

Preliminary analysis

```
# possessives
def get_poss(x):
    pattern = r'\w*-POSS\b'
    poss = re.findall(pattern, ' '.join(str(y) for y in x))
    return poss
```

```
# adding data to the data frames
Ncorpus['Poss_Count'] = Ncorpus.Morphemes.apply(get_poss).str.len()
Lcorpus['Poss_Count'] = Lcorpus.Morphemes.apply(get_poss).str.len()
```

✓ 0.1s


Python

```
# copula
def get_cop(x):
    pattern = r'cop'
    cops = re.findall(pattern, ' '.join(str(y) for y in x))
    return cops
```


```
# adding data to the data frames
Ncorpus['Cop_Count'] = Ncorpus.POS.apply(get_cop).str.len()
Lcorpus['Cop_Count'] = Lcorpus.POS.apply(get_cop).str.len()
```

✓ 0.0s


Python



I defined functions to compile and count the occurrence of 11 functors (selected from those investigated by Brown (1973)).



I appended the counts of each morpheme to the original dfs.



Finally, I normalized the morpheme counts by dividing the counts by text length.

Question: could I use tf-idf for normalization even though these are morphemes, not words?

Morpheme counts

Native speaker corpus

https://github.com/Data-Science-for-Linguists-2023/Morpheme-Acquisition-Analysis/blob/main/data_samples/Ncorp_counts.csv

L2 corpus

https://github.com/Data-Science-for-Linguists-2023/Morpheme-Acquisition-Analysis/blob/main/data_samples/Lcorp_counts.csv

Next Steps

① Create visualizations for deeper analysis.

Probably line graphs, with one line for each morpheme across the years/age range (x-axis). It should be easier to detect patterns this way.

② Compile COCA morpheme frequencies.

As an alternative to SOC, I can compare the frequency with which the learners use morphemes versus the frequency in a large-scale corpus.

③ Conduct statistical test of order ranking difference.

I need to research and decide what tests to conduct. This is necessary to conclude whether there is a difference between L1 and L2 orders.

References



- Brown, R. (1973). A first language. Cambridge, MA: Harvard University Press.
- Jeong, D. B. (2002). Second language acquisition in childhood. Seoul, Korea: Kyungjin Publishing Co.
- Juffs, A., Han, N-R., & Naismith, B. (2020). The University of Pittsburgh English Language Corpus (PELIC) [Data set]. <http://doi.org/10.5281/zenodo.3991977>
- Krashen, S. D. (1985). The Input Hypothesis: Issues and implications. New York: Longman.
- R. A. Berman & D. I. Slobin (1994). Relating events in narrative: A crosslinguistic developmental study. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 38(1), 90-111.

Q&A

Thank you for listening!