Simulating the developmental pattern of finiteness marking in English, Dutch, German, French and Spanish using MOSAIC

Julian Pine, Daniel Freudenthal
University of Liverpool
Fernand Gobet
Brunel University

Outline

- MOSAIC
- The OI phenomenon
- Simulating OI errors in Dutch, Spanish and German
- Simulating OI errors in Wh- Questions
- Comparing MOSAIC and the Variational Learning Model
- Conclusions

MOSAIC: Key Features

- Simple distributional learning mechanism
- Takes as input (orthographically transcribed) samples of Child-Directed Speech
- Produces output in the form of strings learned directly from the input or generated by substituting across generative links between items
- Learns to produce progressively longer utterances as a function of the amount of input it has seen
- Learns from the right edge of the utterance
  - Dolly's having a drink

    Drink

    A drink

    Having a drink

MOSAIC – Key Strengths

- Not intended as a realistic model of the language learning process
- Produces output that can be compared with that of real children
  - Can be used to simulate developmental data
- Learns from input with a realistic frequency distribution
  - Can be used to understand the role of the input in shaping the developmental data
- Uses exactly the same mechanism to simulate data from different languages
  - Can be used to build unified accounts of cross-linguistic phenomena

The OI phenomenon

- Children learning many languages go through a stage in which they produce non-finite verb forms in utterances in which a finite verb form is required
  - English: That go there
  - Dutch: Papa ijs eten (Daddy ice cream eat-INF)
  - German: Thorsten Ball haben (Thorsten ball have-INF)
- OI errors are rare in pro-drop languages such as Spanish and Italian
- Several generativist accounts that attempt to explain this cross-linguistic pattern (e.g. Rizzi, Hyams, Wexler)

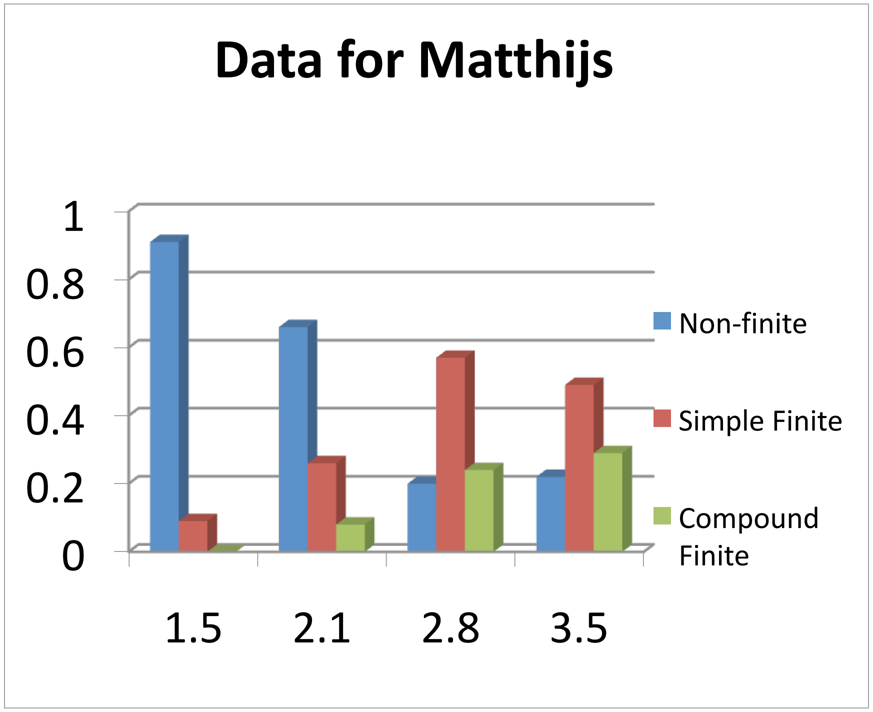
An alternative input-driven account

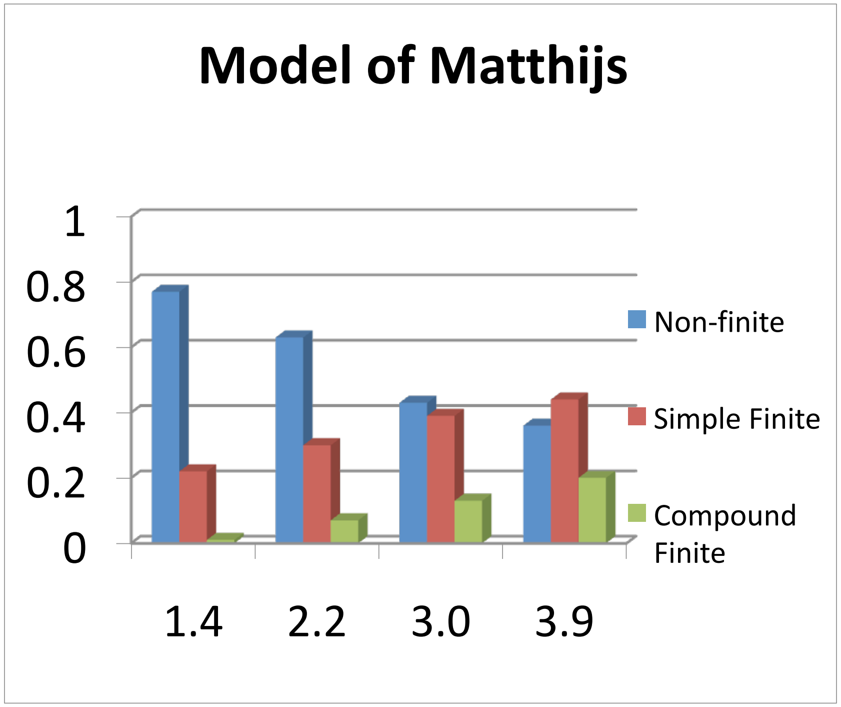
- OIs are reduced compound finites
  - Ijs eten from Hij kann ijs eten
  - Ice cream eat-INF (He can ice cream eat-INF)
- But rate of OIs in early child Dutch is much higher than rate of compound finites in Dutch CDS
- Compound finites occur at similar rates in OI and non-OI languages
- Is it possible to simulate this pattern in terms of the interaction between utterance-final learning and cross-linguistic variation in the input?

Study 1

- MOSAIC exposed repeatedly to speech addressed to a particular child
- Output generated after each run through input
- Output files selected on basis of MLU
- Compared with samples of child speech matched as closely as possible for MLU
- Child and model samples (automatically) coded into:
  - Non-finite (utterances with only a non-finite verb form)
  - Simple finite (utterances with only a finite verb form)
  - Compound finite (utterances with both a finite and a non-finite verb form)

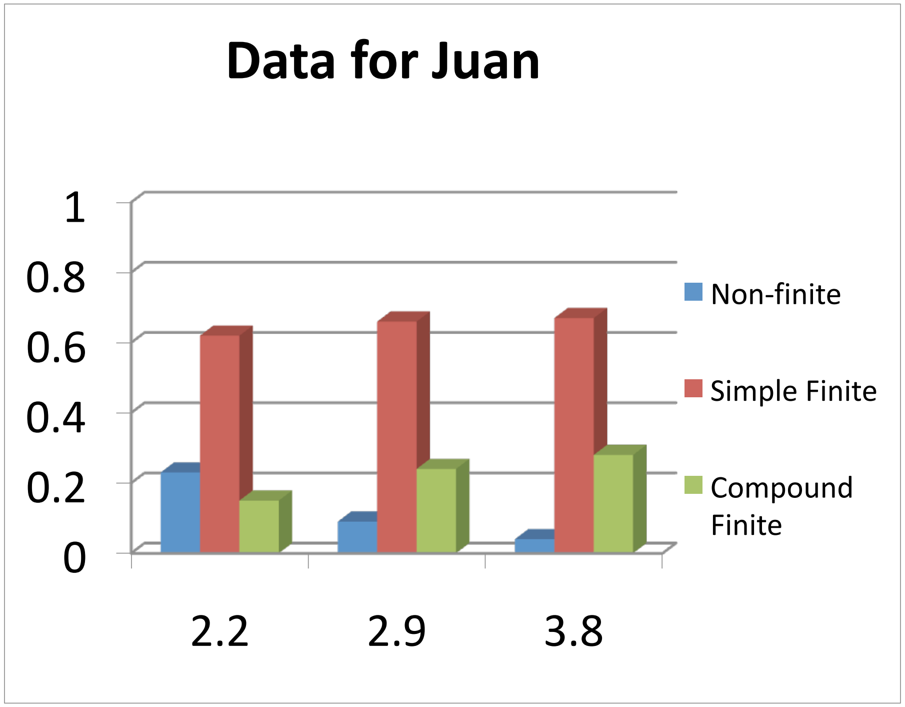
Pattern of finiteness marking as a function of MLU for Matthijs and MOSAIC-Matthijs (Dutch)



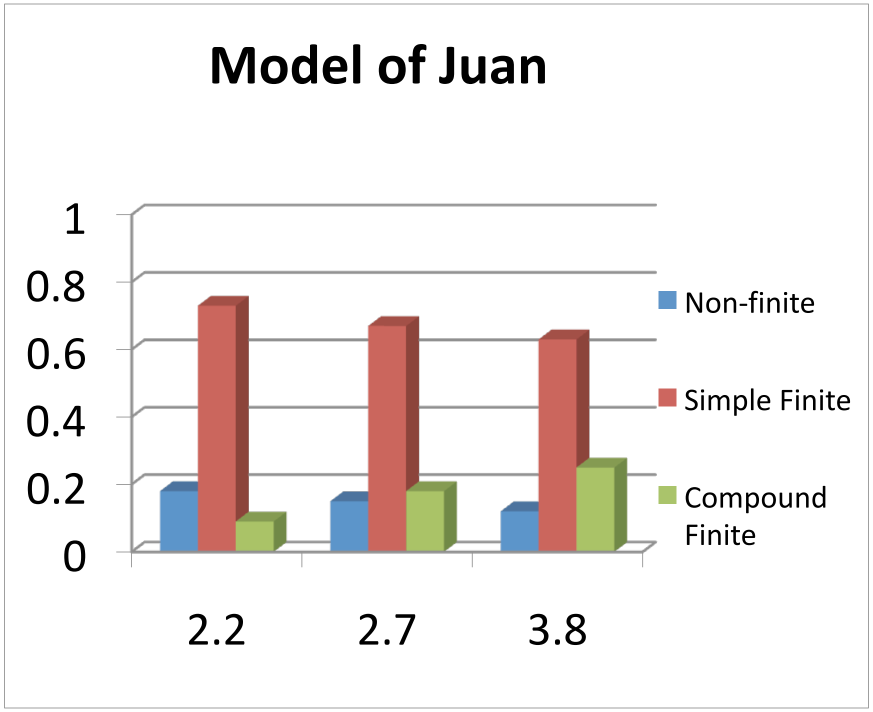


MOSAIC simulates high proportion of OI errors in Dutch

Pattern of finiteness marking as a function of MLU for Juan and MOSAIC-Juan (Spanish)



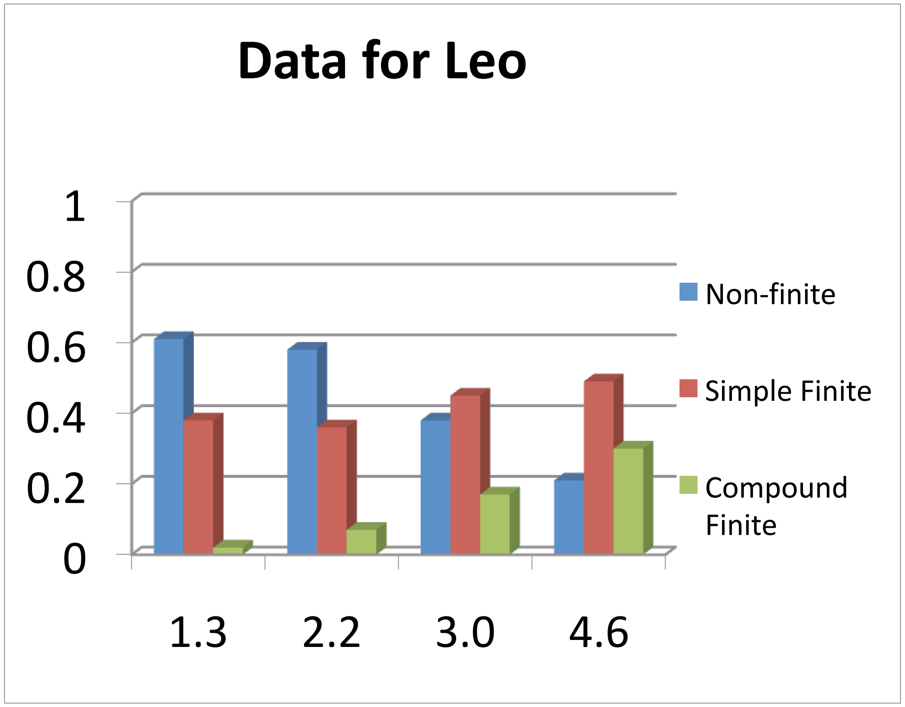
Data for Juan

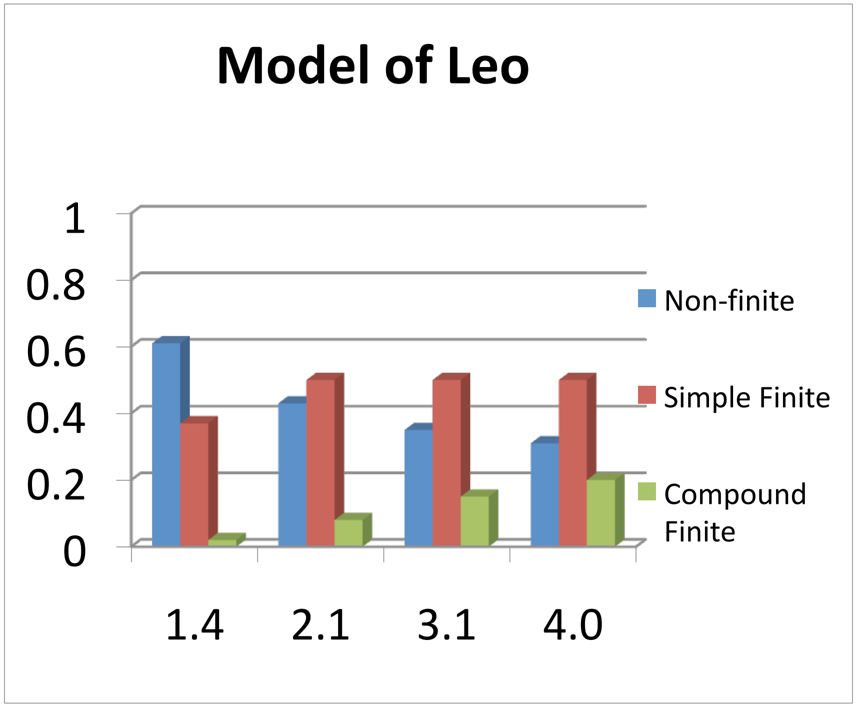


Model of Juan

MOSAIC simulates low proportion of OI errors in Spanish (and high proportion of simple finites)

Pattern of finiteness marking as a function of MLU for Leo and MOSAIC-Leo (German)





MOSAIC simulates (moderately) high proportion of OI errors in German

Best predictor of %OI errors at earliest stage is %Utterance-final verbs that are non-finite

	OI errors at lowest MLU point (%)	Compound finites in Input (%)	Utterance-final non-finites (%)
Dutch	91	34	87
German	61	29	66
Spanish	22	25	26

This variable interacts with utterance-final learning to explain qualitative differences (Dutch v Spanish) and quantitative differences (Dutch v German)

Study 2

- Freudenthal et al. (2007) show that MOSAIC can simulate cross-linguistic data in Dutch, German Spanish (and English) surprisingly well
- This version of the model has 3 important limitations
  - Utterance-final bias but no sensitivity to utterance-initial position
  - Does not distinguish between declaratives and questions in the input
  - Only simulates the pattern of errors in declaratives

OI errors in Wh- questions

- Cross-linguistic pattern of OI errors in Wh- questions is different from that in declaratives

	Declaratives	**Wh- questions**
English	Y	Y
Dutch/German	Y	N
Spanish	N	N

- Is it possible to simulate cross-linguistic pattern in declaratives AND Wh- questions in terms of edge first learning?

Modelling declaratives and questions

- Modified version of MOSAIC that distinguishes between interrogative and non-interrogative input
- Learns from left and right edge of the utterance
- Left edge words/phrases associated with (longer) right edge phrases
  - He (can) go home
  - Where (can) he go?
- Output a mixture of utterance-final phrases and concatenations of utterance-initial and -final phrases
  - Go home, He ____ go home
  - He go? Where ___ he go?

Proportion of OI errors in Wh- questions

	MLU = 2-2.5	MLU = 3-3.5	MLU = 4-4.5
English	.76	.26	.03
MOSAIC	.59	.35	.33
German	(.00 -.33)	.11	.00
MOSAIC	.20	.13	.11
Spanish	.00	.03	.00
MOSAIC	.04	.04	.04

MOSAIC simulates the pattern of OI errors in Wh-questions across English, German and Spanish

What's special about English?

- English does not have subject-main verb inversion
- As a result, all English object Wh- questions are potential models for OI errors
  - What does he want? What does he do? What can he see?
- German and Spanish do have subject-main verb inversion
- As a result, most German and Spanish Wh-questions are not potential models for OI errors
  - What wants he? What does he? What can he see?

Study 3

- One of the strengths of MOSAIC is that it makes quantitative predictions about variation in rates of OI errors across languages
- Most generativist accounts are designed to explain why some languages are OI languages and others are not, but have little to say about differences between OI languages
- Recent exception is Legate and Yang's (2007) Variational Learning Model (VLM)
- How does account implemented in MOSAIC compare with a generativist account that takes quantitative variation in the rate of OI errors more seriously?

Legate & Yang’s (2007) VLM

- A generativist parameter-setting model that focuses on variation in rates of OI errors across languages
- Child has to determine whether she is learning a tense-marking language (e.g. English) or a non-tense-marking language (e.g. Mandarin)
- Gradually rejects [-Tense] grammar on basis of exposure to clauses with tense- or tense-dependent morphology (e.g. kicked, kicks versus kick)
- Speed with which child rejects [-Tense] grammar depends on amount of evidence for tense-marking in input (Spanish > French > English)

Aims

- To test the VLM on a wider range of languages (including 3 languages with intermediate rates of OI errors: Dutch, German and French)
- To extend MOSAIC to French
- To compare MOSAIC and the VLM in terms of their ability to explain variation in rates of OI errors across languages (including the very high levels of OI errors in early child English)
- To differentiate between MOSAIC and the VLM by looking for lexical effects in the data

Lexical Effects

- According to the VLM, OI errors reflect the probabilistic use of a [-Tense] grammar
- VLM predicts similar levels of OI errors across lexical items (i.e. no lexical effects)
- According to MOSAIC, OI errors are compound finites with missing modals/auxiliaries
- MOSAIC predicts that verbs that tend to occur as the main verb in compound finites will be more likely to occur as OI errors in the child's output (i.e. lexical effects in all languages)

Method (Cross-linguistic fit)

- Analysed corpora in English (Theakston et al., 2001), Dutch (Bol, 1995), German (Behrens, 2006), French (Tremblay & Demuth, 2008) and Spanish (Aguado-Orea, 2004)
- Child data analysed for rate OI errors at MLU ≈ 2.0
- To test MOSAIC, model run on input corpora and output analysed for rate of OI errors at MLU ≈ 2.0
- For English, model run on input hand-coded for 3sg contexts (He can go-3SG v I can go) and only 3sg output analysed (Go-3SG v Go)
- To test VLM, input corpora analysed for proportion clauses with tense- or tense-dependent morphology

Proportion of OI errors in children and MOSAIC at MLU ≈ 2.0 and Proportion of clauses rewarding [+Tense] grammar

	Child	MOSAIC	[+Tense]
English	**.87**	**.63**	**.57**
Dutch	**.76**	**.65**	**.49**
German	.58	.49	.62
French	.32	.32	.67
Spanish	.20	.15	.81

- Both models predict rank order of Dutch > German > French > Spanish surprisingly well
- Both models fail to predict English > Dutch

Why the poor fit for English?

- MOSAIC
- %Utterance-final non-finites higher in Dutch (87%) than English (78%)
- Additional mechanism required that is sensitive to impoverished verb morphology in English
- Variational Learning Model
- Lots of evidence for [+Tense] grammar from copulas/auxiliaries which constitute 83% vs 56% of data in English vs Dutch
- Need to distinguish between evidence from copulas/auxiliaries and evidence from main verbs

Method (Lexical Effects)

- Identified all verbs used as correct simple finites or incorrect infinitives by the child (excluding copulas and auxiliaries)
- Calculated rate of OI errors for each verb in child's output
- Calculated proportion of times each verb occurred as the main verb in a compound finite (e.g. He can go) versus the main verb in a simple finite (e.g. That goes there) in child's input
- Calculated cross-item correlations between these two measures in each of the 5 languages

Correlations between rate of OI errors on individual verbs and proportion occurrences as infinitives v simple finites in the input

	Full Set	Restricted Set (N > 2)
English	.35*(43)	.55*(15)
Dutch	.71**(102)	.83**(59)
German	.48* (143)	.68**(69)
French	.45**(75)	.57**(37)
Spanish	.40**(69)	.29+(43)

- Evidence of lexical effects in all 5 languages
- Children’s use of OIs appears to reflect their origins in compound finites in the input

Conclusions

- Possible to simulate cross-linguistic pattern of OI errors in declaratives and Wh- questions on the assumption that OIs are reduced compound finites
- Pattern in declaratives can be explained in terms of interaction between utterance-final learning and variation in proportion of utterance-final non-finites
- Pattern in Wh- questions can be explained in terms of interaction between edge-first learning and differences in the way Wh- questions are formed
- This account can explain quantitative as well as qualitative variation at the OI stage
- It can also explain lexical effects in the data

Conclusions

- Cross-linguistic modelling is a powerful tool for investigating children's language that allows us to:
  – Identify weaknesses in arguments developed through armchair theorising
  – Explore potential interactions between particular processing strategies/constraints and variation in the distributional properties of the input language
  – Generate predictions about the relation between the child's language and the input that can be tested on developmental data
- Cross-linguistic modelling of OI errors provides strong evidence that OIs are learned from compound finites in the input