# Cross-linguistic developmental learning of Word Classes

Daniel Freudenthal, Julian Pine, Fernand Gobet.

University of Liverpool


# How do children learn word classes like noun or verb?

- Several sources of information:
  - Semantic – nouns refer to objects, verbs to actions
  - Prosody – initial stress more common in verbs (récord vs. record)
  - Distributional information
    - Nouns tend to be preceded by determiners, followed by verbs.
    - Verbs tend to be preceded by pronouns, followed by determiners and nouns.

# Redington et al. 1998 – distributional analysis

- For any number of target words, collect counts for the context words in preceding and following position(s).
  - Target words: 1000 most frequent words in a corpus
  - Context words: 150 most frequent words in a corpus
  - Each word has a vector of counts for contexts words.
  - Words with similar vectors are more likely to be of the same word class.
  - Cluster analysis of similarity matrix gives very good results.
  - Several variants developed.

- But:
- Mechanism is thought to be used by language-learning children, but not applied developmentally.
  - Typically used on large corpora, and complete utterances.
  - Ignores the fact that children's early utterances are just one or two words long.
  - Freudenthal et al. (2016) develop a more plausible version by embedding Redington's mechanism within an existing model of Language Acquisition (MOSAIC), that represents increasingly long utterances when shown more input.
  - MOSAIC has an utterance-final bias, learns the last words in the utterance first, and slowly works its way to the front.
  - MOSAIC thus gradually expands the contexts available to the mechanism, in a right-to-left manner.

- Freudenthal et al show that:
- The mechanism works developmentally, provided preceding and following similarity is expressed independently.
- The mechanism builds an early noun category. Nouns tend to occur in utterance-final position.
  - Children also show evidence of a productive noun category before a productive verb category (Olguin & Tomasello, '93; Tomasello & Olguin, '93)
  - Size of the early noun category is greatly enhanced if utterance endings are included as contextual elements.
- Freudenthal et al. (2016) introduce 'Noun-richness' – number of nouns produced over number of noun + (main) verbs.
- Noun richness in early (English) child speech is far greater (~ 80%) than in adults speech (~45%) – consistent with early noun class.
- Early noun class sufficiently large to explain this effect, through productive use of nouns in MOSAIC's output.

# Great, but what about other languages?

- Here we look at German and Dutch
  - English has SVO order, German and Dutch are SOV/V2. Dutch/German word order is less constrained. Fewer nouns in utterance-final position, lower child noun-richness?

1a. I eat a cookie (E)

1b. Ich esse ein Keks (G - I eat a cookie)

1c. Ik eet een koekje (D - I eat a cookie)

2a. I want to eat a cookie.

2b. Ich moechte ein Keks essen (G - I want a cookie eat)

2c. Ik wil een koekje eten (D - I want a cookie eat)

- English has neither Gender or Case,
- German has 3 Genders, and 4 cases, marked on definite and indefinite articles (as well as possessives, demonstrative, adjectives)

| | Nom. | Gen. | Dat. | Acc. |
|---|---|---|---|---|
| Masc. | der/ein | des/eines | dem/einem | den/einen |
| Fem. | die/eine | der/einer | der/einer | die/eine |
| Neut. | das/ein | des/eines | dem/einem | das/ein |
| Plural | die/eine | der/einer | den/einen | die/eine |

- Dutch has Gender (2) but not case.
- Gender is marked on the definite (but not indefinite article), and demonstratives but not possessives, and on (some) adjectives.

Indefinite article: *Een*

Definite articles: *de* (common) *het* (neuter)

- De/een boom (the/a tree)
- Het/een huis (the/a house)
- Plural De bomen, de huizen.

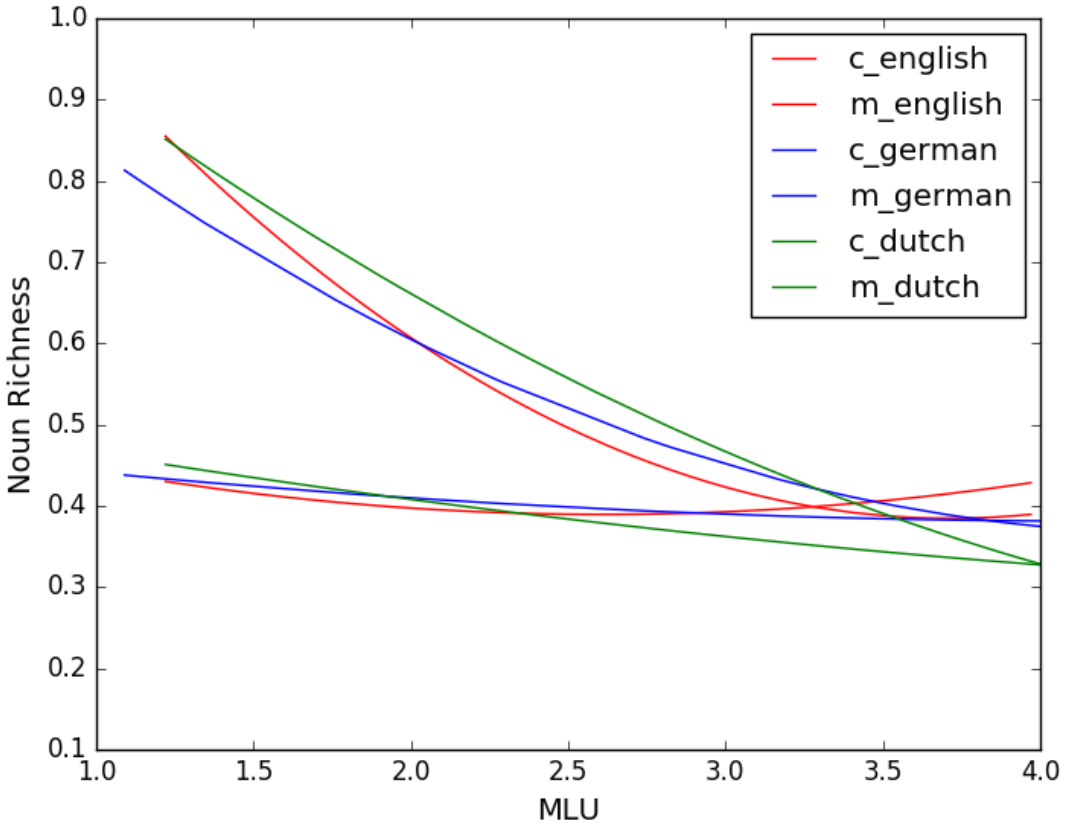
- Dutch and German word order differs from English – this may result in fewer nouns in utterance-final position.
- Gender/Case means more variation in determiners (in particular for German) – more difficult to learn the noun category? What if we treat the different forms of the determiner as one?
- But, children need to learn Gender, and distribution is the only(?) reliable cue?

1. Compare Child noun-richness in English, Dutch, German
2. Do distributional analysis across the 3 languages, compare the noun classes
3. Do distributional analysis with merged determiners.
4. Can we learn Gender through distributional analysis?

- Tried to use input corpora that are roughly equal in size:
- English: 6 largest corpora from Manchester corpus (~30-35k input utts)
- German: Rigol corpus – 4 children, ~ 45k input utts.
- Dutch: van Kampen – 2 children, 25k and 65k input utts.

- Noun richness computed on a tape-by-tape basis.
  - Trendlines for different languages.

Child and Adult noun richness in English, German and Dutch




German (initially) slightly lower, but very similar overall

- Distributional analysis in MOSAIC
- MOSAIC trained for 50 runs – distributional analysis performed at different points between run 36 (MLU 2) and 50 (MLU 5).
- Identical procedure: 1000 target words, 150 context words.
- Counts collected from the utterance-final phrases represented in MOSAIC. One position before and after target words.
- Two distance measures: rank order and cosine similarity on square root of counts.
- Two words considered the same if rank order > 0.4, OR cosine > 0.65.
  - Both measures give qualitatively similar results, but quantitatively best if combined.
  - No cluster analysis

| Run | Links | Overall accuracy | Noun-richness | Nouns | Verbs | Noun-accuracy | Verb-accuracy |
|---|---|---|---|---|---|---|---|
| English | | | | | | | |
| 36 | 1,641 | 0.8 | 0.94 | 1,218 | 70 | 0.83 | 0.42 |
| 38 | 2,215 | 0.8 | 0.91 | 1,553 | 153 | 0.83 | 0.52 |
| 40 | 3,037 | 0.83 | 0.89 | 2,230 | 237 | 0.85 | 0.63 |
| 44 | 4,144 | 0.90 | 0.86 | 3,164 | 437 | 0.91 | 0.81 |
| 50 | 4,576 | 0.91 | 0.83 | 3,375 | 615 | 0.92 | 0.87 |
| | | | | | | | |
| Dutch | | | | | | | |
| 36 | 1,140 | 0.73 | 0.95 | 774 | 34 | 0.77 | 0.23 |
| 38 | 2,030 | 0.78 | 0.96 | 1,467 | 62 | 0.80 | 0.38 |
| 40 | 2,995 | 0.81 | 0.96 | 2,260 | 90 | 0.82 | 0.43 |
| 44 | 3,496 | 0.85 | 0.91 | 2,582 | 256 | 0.85 | 0.75 |
| 50 | 3,310 | 0.84 | 0.80 | 2,122 | 502 | 0.84 | 0.86 |
| | | | | | | | |
| German | | | | | | | |
| 36 | 841 | 0.52 | 0.93 | 282 | 20 | 0.54 | 0.27 |
| 38 | 935 | 0.61 | 0.89 | 383 | 43 | 0.64 | 0.47 |
| 40 | 1227 | 0.71 | 0.87 | 581 | 86 | 0.71 | 0.61 |
| 44 | 1,985 | 0.78 | 0.78 | 905 | 253 | 0.80 | 0.84 |
| 50 | 2,563 | 0.79 | 0.52 | 754 | 697 | 0.83 | 0.89 |

- English gives overall good results. Large initial noun class, with verbs coming in later. Overall good accuracy.
- Dutch has lower accuracy overall, similar verb class, but smaller noun class.
- German is like Dutch – but much smaller noun class overall.

- Size of the noun class is inversely related to the complexity of the determiner system.
- Merging determiners:
- Der/des/dem/den/das/die -> det1; ein, eine, einer etc -> det2
- De/het -> det1; een -> det2

Results with merged determiners look remarkably similar across all three languages. Dutch Noun class increased by about 40%, German by factor 4 – but lower overall accuracy.

| Run | Links | Overall accuracy | Noun-richness | Nouns | Verbs | Noun-accuracy | Verb-accuracy |
|---|---|---|---|---|---|---|---|
| Dutch | | | | | | | |
| 36 | 1,515 | 0.7 | 0.96 | 997 | 37 | 0.73 | 0.17 |
| 38 | 2,749 | 0.76 | 0.96 | 1,955 | 70 | 0.78 | 0.3 |
| 40 | 4,140 | 0.8 | 0.96 | 3,134 | 104 | 0.81 | 0.36 |
| 44 | 5,151 | 0.84 | 0.93 | 3,940 | 292 | 0.86 | 0.65 |
| 50 | 4,788 | 0.84 | 0.84 | 3,290 | 573 | 0.85 | 0.8 |
| | | | | | | | |
| German | | | | | | | |
| 36 | 2,091 | 0.49 | 0.97 | 836 | 27 | 0.51 | 0.15 |
| 38 | 2,399 | 0.56 | 0.95 | 1,095 | 58 | 0.57 | 0.26 |
| 40 | 3,543 | 0.65 | 0.94 | 1,914 | 131 | 0.65 | 0.4 |
| 44 | 5,992 | 0.73 | 0.91 | 3,540 | 364 | 0.74 | 0.73 |
| 50 | 6,287 | 0.76 | 0.8 | 3,226 | 816 | 0.77 | 0.84 |

Can we learn Gender? Noun-confusion in noun-noun Links. No merging, run 50.

Dutch:

| | Com. | Neut | Plur. | | Com | Neut | Pl |
|---|---|---|---|---|---|---|---|
| Comm | 1415 | 187 | 102 | | 0.83 | 0.11 | 0.06 |
| Neut | 187 | 249 | 10 | | 0.42 | 0.56 | 0.02 |
| Pl | 102 | 10 | 17 | | 0.79 | 0.08 | 0.13 |

German:

| | Masc | Fem | Neut | Pl | | Masc | Fem | Neut | Pl |
|---|---|---|---|---|---|---|---|---|---|
| Masc | 216 | 15 | 39 | 5 | | 0.79 | 0.05 | 0.14 | 0.02 |
| Fem | 15 | 198 | 0 | 18 | | 0.06 | 0.86 | 0.0 | 0.08 |
| Neut | 39 | 0 | 203 | 2 | | 0.16 | 0.0 | 0.83 | 0.01 |
| Pl | 5 | 18 | 2 | 25 | | 0.1 | 0.36 | 0.03 | 0.51 |

Better separation in German, German Gender easier to learn.

# Conclusions

- Child Noun richness quite similar across languages, suggesting similar productivity around nouns.
- Size of noun class inversely related to complexity of determiner system.
- Merging determiners yields remarkably similar results across languages – despite differences in word order.
- Dutch/German accuracy lower, but the same developmental pattern – nouns first, verbs later.
- Dutch Gender more difficult to learn than German – no marking on the indefinite article – more confusable.

# Conclusions

- Why merge determiners when children don't produce them? They know more than they let on?
- Spanish (3yo) and French (2yo) can use gender to differentiate in looking-while-listening (Lew-Willams and Fernald (2007), van Heugten and Shi (2009). Determiner fully predictive of Gender.
- German (16mo) can use determiner to classify a novel word as noun (but not pronoun/verb) Hőhle et al., 2004.
- Dutch children appear delayed in the Lew-Williams paradigm (van Heugten and Johnson (2011), consistent with the poor separation of gender relative to German. Dutch really is more difficult to learn.

# Conclusions

- Children may represent word classes at multiple levels of abstraction that represent both the overall category, and the finer subclasses.

- We are not the first to merge determiners. Keibel (2005) does the same (with similar results), but he only looks at German, and doesn't consider child speech.

- Gender/case is an area where bilinguals struggle (at least in on-line measures), even for advanced learners. It seems you need to learn it early, and distributionally.

Thank You