Predicting Gender Assignment in Icelandic: A Longitudinal Corpus Case Study

Sigríður Mjöll Björnsdóttir and Sigríður Sigurjónsdóttir

1. Introduction

Grammatical gender has descriptively been defined as the sorting of nouns into classes, which is reflected in agreement patterns outside of the noun itself (Corbett 1991). Knowledge of a language that encodes gender involves being able to predict which nouns fall into which group, or, gender classes. Grammatical gender is a language-specific property that is subject to a great deal of cross-linguistic variation. This raises the question of how children are able to discover which patterns are predictive to gender assignment in the language they are acquiring. The age of acquisition and the learning trajectory of grammatical gender has been shown to be subject to cross-linguistic variation. But, hitherto there has been no account of gender acquisition that can explain children’s differing behaviors in the process of acquiring different gender systems.

In this paper, we show how the learning trajectory of a gender system can be predicted by modelling the productivity of mappings between noun endings and gender assignment, using Icelandic as a test case. Icelandic has a 3-gender system that distinguishes between masculine, feminine and neuter in both the singular and the plural. The three genders have been claimed to be represented at roughly equal frequency in Icelandic speech (Rögnvaldsson 1990). In addition, Icelandic has a 4-gender case system that distinguishes between nominative, accusative, dative and genitive. Neuter has been argued to be the gender default in Icelandic, on the basis of traditional syntactic criteria for the appearance of default morphological forms (see e.g. Thráinsson 2007, Rögnvaldsson 2013). Following Björnsdóttir (in progress), we argue that the domain of generalization to the child learner acquiring Icelandic gender is the case ending in the nominative singular. The current study found this form to be the most frequently attested nominal case form (47%) in a corpus of Icelandic child-directed speech (Sigurjónsdóttir 2007).

The learning trajectory of the Icelandic gender system was modeled using the Tolerance Principle (Yang 2016). The Tolerance Principle (TP) is an independently motivated quantitative measure that can predict with numerical precision, how many exceptions a linguistic pattern can tolerate in order to be...
perceived as a generalization by the child learner. We used the Tolerance Principle (TP) to measure the productivity of nominative singular endings to gender assignment in Icelandic on noun types that were extracted from a 500,000 word token corpus of Icelandic child-directed speech (Sigurjónsdóttir 2007). These measures of productivity were used to make predictions about the learning trajectory of grammatical gender, with respect to age of acquisition and target-consistency of gender assignment. In particular, we show how the TP can predict non-target consistent gender assignment in the acquisition of grammatical gender in Icelandic.

The predictions were tested on a longitudinal child production data (Sigurjónsdóttir 2007). The data consist of the spontaneous speech of a single child (7,000 word tokens) recorded by her caretaker in 2005-2007 between the ages of 1;6-4;3. The results can be summarized as follows: there were no non-target consistent forms attested for noun classes that have a nominative singular ending that is predictive of a specific gender class, as measured by the TP. Non-target consistent forms in the child production data exclusively targeted a subclass of nouns that the TP predicted to lack a gender default. This class of nouns constituted around 41% of all nouns produced by the child and the rate of non-target consistency within this class was quite high, or 18%. Furthermore, there was no systematic pattern detected to the non-target consistent forms produced. In other words, the child did not seem to resort to a gender default for nouns, which lacked a ‘gender-predictive’ nominative singular case ending. This is unexpected given the traditional theoretical assumption that neuter is the default gender in Icelandic.

The paper is organized as follows: section 2 gives a brief overview of key findings in the acquisition of grammatical gender. In section 3, we give both a descriptive overview of the gender system of Icelandic and present the hypothesis put forth in Björnsdóttir (in progress) about the domain of generalization to gender learning. Section 4 motivates the use of the TP in grammatical gender acquisition research and its application to a corpus of Icelandic child-directed speech. The predictions of the TP on the acquisition of grammatical gender in Icelandic are presented. In section 5, we discuss grammatical gender in the child-directed speech and the child production data. Section 6 concludes this paper with a summary of main findings and a discussion of some outstanding questions that need to be addressed with further data collection of children acquiring the Icelandic gender system. Finally, we discuss the implications of this work for further research on the theory of grammatical gender and the acquisition of gender systems cross-linguistically.

**2. Prior research on the acquisition of grammatical gender**

Young children have been shown to pay attention to formal properties of nouns when forming generalizations about gender assignment in their language (Karmiloff-Smith 1981). These formal properties manifest themselves on nominal
endings in the form of morphophonological cues to gender assignment. Whether other morphological or phonological correlations with gender assignment are useful to the learner is an open research question, although prior findings strongly indicate that this is in fact the domain of generalization for gender acquisition to the child learner (see e.g. Gvozdev 1961, Mills 1986). In spite of various theoretical arguments in favor of a semantic core to formal systems of gender assignment (see extensive discussion in Corbett 1991), children seem to be able to learn gender systems that are purely formal and detached from semantic motivation, as seems to be the case in Icelandic.

Gender systems differ with respect to how predictive noun endings are to gender assignment. Gender systems that have a set of predictive formal cues to gender assignment have been described as transparent, whereas gender systems that possess few such cues have been described as opaque. Transparent gender systems, like those of Italian and Spanish, are early acquired, yet non-transparent or opaque systems, as for example the Norwegian gender system, have been argued to be late acquired. These cross-linguistic differences are reflected in children’s differing learning trajectories, with transparency being the main predictor in children’s behavior (see e.g. Karmiloff-Smith 1981, Mills 1986, Rodina & Westergaard 2015). For instance, Spanish-learning children, aged 34-42 months, with an estimated vocabulary of only 500 words have been shown to use gender-marked articles in establishing reference (Lew-Williams & Fernald 2007). Children learning Spanish, therefore, seem to be able to form productive generalizations about gender assignment on the basis of a very small vocabulary. For other gender systems, children do seem to need larger vocabularies in order to learn the system. This is, for example, the case for the gender system in Norwegian, which has been shown not to be fully in place until around the age of 7 (Rodina & Westergaard 2015). However, whether the gender system is transparent or opaque, the learning task must consist of discovering which tendencies are predictive to gender assignment in the target language.

Even if children’s learning trajectories have been shown to differ cross-linguistically, children’s production of non-target consistent gender forms indicates that they undergo a parallel generalization-based process irrespective of the target language when acquiring gender. The process is characterized by a search for predictive tendencies to gender assignment that become generalizations. Children’s non-target consistent gender forms reflect the problem of overcoming syncretism between gender classes in the nominative singular.

In the case of syncretism, the learner must establish what the productive gender generalization is. For example, the noun ending -а is a very common noun ending for feminine nouns in Russian (1a). However, there are exceptions to this general pattern in the form of masculine nouns (1b) that also take this noun

---

1 Prior research on gender acquisition has not been explicit about whether the relevant generalizations to gender assignment are morphological or phonological in nature. The authors leave it up to future research to decide which is the case and refer to ‘noun endings’ in accordance to the cited references.
ending. The differences in gender assignment of these two nouns are realized on
the suffix of the agreeing adjective, as indicated by bold font in (1) below.

(1)  a. Mil’aja mam-a.
    Dear-f. mother-f.
    'Dear mother.'

       b. Mil’yj pap-a.
       Dear-m. father-m.
       'Dear father.'

Children acquiring Russian have been reported to treat masculine nouns, like
(1b), as feminine nouns before the age of 3. These kind of non-target consistent
forms have been attested both in spontaneous speech (Gvozdev 1961) and
experimental settings (Rodina 2014).

Thus, even in a transparent gender system, like Russian, children have been
argued to produce non-target consistent gender marking on nouns that are
syncretic as late as age 7;9 (Gvozdev 1961:442). The classes of nouns that pose
particular problems to Russian-learning children are feminine nouns ending in a
palatalized consonant (myš ‘mouse’, rož ‘rye’, krovat ‘bed’), that are syncretic
with masculine nouns in the nominative singular, on the one hand, and stem-
stressed neuter nouns, on the other hand, that can be syncretic with either
masculine or feminine. In case of the erroneous feminine nouns, the masculine is
generalized, which also has been argued to be the default gender for Russian. In
case of the erroneous neuter nouns, the child documented by Gvozdev (1961)
reanalyzed them in two different ways, both of which are consistent with the more
general principles of Russian grammar. If the final unstressed vowel of the noun
in question is retained, such a noun is treated as feminine, e.g. čudo ‘miracle’,
poleno ‘log’, ukrašenie ‘decoration’ (Gvozdev 1961: 442), and the agreeing form
reflects that. The other strategy consists of deleting the final vowel of stem-
stressed neuter nouns; the resulting form is then reinterpreted as masculine, e.g.
jabloko > jablok ‘apple’, doloto > dolot ‘chisel’ (Gvozdev 1961: 442).

These kind of non-target consistent forms, discussed above, are of a similar
nature to the non-target consistent forms that have been attested in the acquisition
of the English past tense. In the well-known case of English past-tense acquisition,
the child is faced with overcoming the problem of over-regularization at a certain
stage of learning. A longitudinal corpus study of Adam, a child acquiring English
(Brown 1973) showed that he produced his first recorded non-target consistent
past tense form (go-ed as opposed to went) at the age of 2;11. However, Adam’s
earlier transcripts, going back nearly a year, do not contain a single instance of an
incorrect irregular verbal past tense. This suggests that he needed to reach a
certain vocabulary threshold to make the distinction between productive and
unproductive patterns in the target language. We argue that the same holds for the
acquisition of grammatical gender.
To sum up, learning grammatical gender involves detecting regularities in the morphosyntactic behavior of nouns. Even for gender systems that have been described as transparent, like Russian, some ‘facts’ about gender assignment are learned later than others. These facts constitute exceptions to the otherwise productive patterns of gender assignment and need to be learned as such. Therefore, the non-target consistent forms are not at random, but reflect a process of the systematic grouping of nouns into classes on the basis of a shared formal property.

3. The gender system of Icelandic
3.1. A descriptive overview

Icelandic has a formal system of gender assignment that distinguishes between 3 gender classes; masculine, feminine and neuter. Gender is encoded in both the singular and the plural. In addition to gender marking, Icelandic also distinguishes between 4 cases: nominative, accusative, dative and genitive. Grammatical gender and case interact in systematic ways to generate patterns of noun inflection that have descriptively been referred to as inflection classes. From a learning perspective, this means that the child will encounter a noun of any gender with various nominal endings.

Gender agreement is realized on the definite suffix (2a), adjectives (2b), the verbal past participle (2c) and pronouns, even when the referent noun is inanimate (2d).

(2) a. Maður-inn, kon-an, barn-ið.
   Man.the-m., woman.the-f., child.the-n.
   'The man, the woman, the child.'

b. Falleg-ur maður, falleg-ø kona, falleg-t barn.
   Beautiful-m. man, beautiful-f. woman, beautiful-n. child
   'Beautiful man, beautiful woman, beautiful child.'

c. Maðurinn er fari-nn, konan er fari-n, barnið er fari-ð.
   Man.the is gone-m., woman.the is gone-f., child.the is gone-n
   'The man is gone, the woman is gone, the child is gone.'

d. Hann (stóll) er góður, hún (mynd) er góð, það (ljós) er gott.
   He (chair-m.) is good, she (picture-f.) is good, it (light-n.) is good
   'The chair is good, the picture is good, the light is good.'

A corpus study of modern Icelandic speech reported that nouns divide roughly evenly across gender classes (Rögnvaldsson 1990). Neuter has been argued to be

---

2 This paper will only present findings in the singular, but see Björnsdóttir (2018) for an analysis of noun pluralization in Icelandic from a learning perspective.
the default gender in Icelandic (Rögnvaldsson 2013). Evidence for this comes from agreement with oblique (3a) and sentential (3b) subjects:

(3) a. Mér er kalt.
    Me-dat. is cold-n.
    'I am cold.'

    b. Að hann sé farinn er hræðilegt.
    That he be gone is terrible-n.
    'That he is gone is terrible.'

While these are syntactic environments where default syntactic features are expected to appear, it is at present unclear whether this information is useful to the child acquiring Icelandic.

3.2. The present hypothesis

There is at present no theoretical account that explains explicitly what conditions the distribution of gender assignment in Icelandic. The gender system has been described as ‘arbitrary’ (Ingason 2016) or ‘idiosyncratic’ (Rögnvaldsson 2013) in theoretical terms. Furthermore, due to the interaction of gender, case and number, Icelandic inflectional morphology has been described as complex (Müller 2005). However, if the generalizations behind gender assignment in Icelandic are entirely idiosyncratic, it is hard to imagine how the system could be maintained between generations of speakers, who all had to go through the process of acquiring it. Therefore, there must be some detectable regularities beneath the surface cacophony that guide the Icelandic-learning child in the acquisition of grammatical gender.

We propose that the Icelandic-learning child generalizes on the basis of case endings in the nominative singular. Case ending is a word-final morpheme that participates in systematic alternations across different syntactic contexts. The nominative singular was found to be the most frequently attested nominal case form in a corpus of child-directed speech; roughly half (47%) of nouns were attested in the nominative singular (Sigurjónsdóttir 2007). Given its frequency in the input, the child learner may decide to generalize on the most informative nugget of information in the input, even when he/she encounters a noun in a different case form. Furthermore, there is diachronic evidence in the Germanic languages for the privilege of the nominative singular for driving inflection class shifts (Lahiri & Dresher 1983).

This restricts the domain of generalization to a set of only four inflectional markers that we hypothesize to be meaningful to the learner, as shown in table 1:
Table 1: Icelandic case endings in the nominative singular across gender classes

<table>
<thead>
<tr>
<th>Masculine (M)</th>
<th>Feminine (F)</th>
<th>Neuter (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-r</td>
<td>-a</td>
<td>-ø</td>
</tr>
<tr>
<td>-i</td>
<td>-ø</td>
<td>-a</td>
</tr>
<tr>
<td>-ø</td>
<td></td>
<td>-r</td>
</tr>
</tbody>
</table>

In (4), we show how these markers map on to actual words:

(4)  
  a.  Matu-r (M), lifu-r (F).  
       'Food, liver.'

  b.  Penn-i (M).  
       'Pen.'

  c.  Kann-a (F), aug-a (N).  
       'Mug, eye.'

  d.  ull-ø (F), stóll-ø (M), gull-ø (N).  
       'Wool, chair, gold.'

There is abundant syncretism between the three gender classes in the nominative singular (nom.sg.); all nominative singular case endings are promiscuous between 2 or more gender classes, apart from the masculine marker –i. However, not all of these case endings correlate equally strongly with a particular gender class. Informally speaking, on the basis of our native speaker intuitions we can describe the following correlations, listed in (5):

(5)  
  a. Nouns ending with the nom.sg. marker -r are almost always masc.  
  b. Nouns ending with the nom.sg. marker -r are rarely feminine  
  c. Nouns ending with the nom.sg. marker -i are almost always masc.  
  d. Nouns ending with the nom.sg. marker -a are almost always fem.  
  e. Nouns ending with the nom.sg. marker -a are rarely neuter  
  f. Nouns ending with the nom.sg. marker -ø are frequently neuter

In the next chapter, we will address directly what kind of mechanism is responsible for these native speaker intuitions and how they are formed in the developing grammar.

4. The acquisition of grammatical gender in Icelandic: a corpus case study

4.1. Research questions

On the basis of the preceding discussion, we ask how does the Icelandic-learning child form generalizations about gender assignment in the light of syncretism? In particular, we ask:
(6) a. Is it possible to quantify how predictive the correlations listed in (5) are to gender assignment in Icelandic?

b. Is the predictive power of correlations reflected in the learning trajectory of grammatical gender in Icelandic?

4.2. Methods

The Tolerance Principle (TP) is an independently motivated model of linguistic productivity that generalizes on the number of positive examples for a linguistic tendency. It has previously been applied to various morphosyntactic phenomena cross-linguistically, such as past tense formation in English, noun pluralization in German, dative substitution in Icelandic, to name a few examples (see Yang 2016 for case studies). The TP is a theory of how linguistic generalizations are formed. The TP can precisely quantify the number of exceptions a linguistic tendency can tolerate in order to become a linguistic generalization. It states that the exception threshold for a linguistic tendency cannot exceed $N/\ln N$ in order to be useful to the learner. The TP also allows for a non-productive linguistic correlation to emerge, if the number of exceptions to a correlation is higher than the measured threshold. The TP predicts such a correlation to be lexically arbitrary and learned on an item-to-item basis (see e.g. Gorman and Yang 2018 for predicting lexical gaps using the TP and Schuler et al. 2016 for experimental evidence for children making a categorical distinction between productive and unproductive patterns using an artificial language learning paradigm).

478 noun types in the nominative singular were extracted from 500,000 word token corpus of child-directed speech (Sigurjónsdóttir 2007). The noun types were tagged for gender and case ending in the nominative singular. We then used the Tolerance Principle (Yang 2016) to measure the productivity of nominative singular case endings to gender assignment in Icelandic. The application of the Tolerance Principle to noun types made it possible to quantify precisely the productivity of individual correlations between gender assignment and nominative singular case endings.

The predictions were tested on a longitudinal corpus of a child’s spontaneous speech (Sigurjónsdóttir 2007). The child was recorded between the ages of 1;6:12-4;3:07 and produced 7,000 word tokens in total. Nouns constitute about half of all the word tokens. Noun types from the child production data were extracted and coded for gender, case ending in the nominative singular and target-consistency. Since bare nouns in Icelandic have no overt gender marking, target-consistency was defined in terms of target-consistent use of the definite suffix and gender agreement dependencies.
5. Results
5.1. Productivity of gender assignment correlations in child-directed speech

The noun types extracted from the child-directed speech were distributed across the three genders in the following way:

Table 2: Distribution of noun types across gender classes in the corpus of Icelandic child-directed speech

<table>
<thead>
<tr>
<th>Masculine</th>
<th>Feminine</th>
<th>Neuter</th>
</tr>
</thead>
<tbody>
<tr>
<td>41% (196)</td>
<td>30% (143)</td>
<td>29% (139)</td>
</tr>
</tbody>
</table>

While the frequency of the three gender classes is not as evenly spread out in the child-directed speech as reported in e.g. Rögnvaldsson (2013), there is no gender class that is a statistical minority. Hence, frequency effects in gender acquisition of Icelandic should be at minimum. The numerical distribution of noun types across gender class and case ending in nominative singular is shown, below:

Table 3: Numerical distribution of noun types across gender class and nom. sg. case ending

<table>
<thead>
<tr>
<th>Masculine</th>
<th>Feminine</th>
<th>Neuter</th>
</tr>
</thead>
<tbody>
<tr>
<td>-r (63)</td>
<td>-a (133)</td>
<td>-ø (134)</td>
</tr>
<tr>
<td>-i (82)</td>
<td>-ø (35)</td>
<td>-a (2)</td>
</tr>
<tr>
<td>-ø (29)</td>
<td>-r (0)</td>
<td></td>
</tr>
</tbody>
</table>

It is obvious at the outset, just by looking at the numerical distribution alone, that some gender assignment correlations seem more robust than others. For example, 133 noun types that take the nom.sg. case ending -a are feminine, while only 2 noun types of the same ending are neuter in the corpus. The TP calculates the exception threshold for feminine nouns that take the nominative singular ending -a to be 133/ln133 = 27. Clearly, the two exceptions to this pattern attested in the corpus do not even come close to this threshold.

Another striking observation is that all noun types that take the nom.sg. case endings -r and -i are masculine, so there are no exceptions to this pattern attested in the child-directed data. There are feminine nouns that take the -r ending attested in the language, albeit infrequent, as their complete absence in the corpus reflects. For these noun classes, the exception threshold is trivially met: there are no exceptions to either the 63 positive examples of masculine nouns that take the nominative singular ending -r nor the 82 positive examples of masculine nouns taking the nominative singular ending -i.

The most interesting classes of nouns in the corpus are the ones that take -ø as a nom.sg. case ending. While neuter nouns taking the nom.sg. case ending -a may constitute a minor exception to a more general pattern, the case of the -ø ending nouns is less clear. This class of nouns is syncretic between all three gender classes and each gender class has numerous members within it, although neuter has a striking statistical dominance over both masculine and feminine. However,
the child-directed corpus counts show that the combined number of masculine and feminine nouns (29+35=64) exceeds the exception threshold (37) for the neuter to be a default for this class of nouns, as table 4 shows:

<table>
<thead>
<tr>
<th>Masculine</th>
<th>Feminine</th>
<th>Neuter</th>
<th>Total N</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>35</td>
<td>134</td>
<td>198</td>
<td>37</td>
</tr>
</tbody>
</table>

Given the frequency of neuter assignment within this noun class, one could expect that the child learner would overgeneralize neuter to masculine and feminine nouns within the class in the process of acquiring gender. Furthermore, neuter has been argued to be the default gender for Icelandic on the basis of syntactic agreement facts, as discussed in section 3.1. But the prediction of the TP for this class of nouns are clear: there is no default gender for the -ø nouns.

5.2. Predictions

On the basis of the findings presented in section 5.1. we make the following predictions for the learning trajectory of grammatical gender for the Icelandic child learner:

(7)  a. The child is target-consistent from very early on with the gender assignment of nouns that have gender-predictive nom.sg. case ending, -r, -i, -a, as predicted by the TP.

b. The child is non-target-consistent with the noun class that has a morphologically null (-ø) nom.sg. case ending. Furthermore, since this noun class is predicted by the TP to lack a gender default, the child will assign gender unsystematically, or at random.

5.3. Results: A longitudinal case study of an Icelandic-learning child

The results of the longitudinal corpus case study show that the child is target-consistent with gender assignment with nouns that have a nominative singular inflection ending (-r, -i, -a) that is productively associated with a particular gender class. In fact, there is literally no non-target consistent use of gender markers with these nouns in the corpus from the earliest recordings that date from before the first birthday. This suggests that the child only needs a small vocabulary to form generalizations about gender assignment about these classes of nouns.

The child’s non-target consistent gender forms exclusively target the class of nouns, the ø-nouns, that we have shown to lack a default gender feature. These nouns constitute 41% (43) of the 104 noun types produced by the child in the corpus. Of these, 18% (19) appear with non-target consistent gender marking between the ages 3;0-4;3. 37% of the non-target consistent gender agreement
occurs with masculine and neuter nouns, but 26% with feminine nouns. Figure 1 shows the numerical distribution of the non-target consistent forms as well as the gender that is assigned erroneously:

![Figure 1: Distribution of non-target consistent gender agreement across gender classes](image)

All three gender classes are subject to non-target consistent gender agreement within the ‘null’ subclass of nouns. Feminine is numerically least affected, but the effect is not statistically significant. What is interesting, however, is that the child does not seem to extend feminine agreement much to the masculine or neuter, as this takes place in only 3/14 cases with either a masculine or a neuter noun. However, this still constitutes over a fifth of the non-target consistent gender agreement for masculine and neuter nouns, but it seems that the child is more inclined to generalize either masculine or neuter. Some examples of this non-target consistent behavior are given in (8):

(8) a. Litla krúsíð mitt.
    Little-f/n. mug.def-n. my-n.
    'My little mug.'

    b. Niður úr skýnum.
    Down from sky.def-m.dat.
    'Down from the sky.'
c. **Falleg-o kjóll.**
   Beautiful-f. dress-m.
   'A beautiful dress.'

In (8a), we see a -o noun assigned neuter instead of the target consistent feminine, as evident on the definite suffix and the possessive pronoun. In (8b), a -o neuter noun is assigned a dative inflection marker that is only existent within the masculine paradigm. In (8c), the form of the adjective indicates that the child interprets this noun, that is masculine in the adult language, as feminine. Thus, there is no systematic direction in non-target consistent gender forms that the child produces with the -o nouns, which reflect the lack of a productive gender default.

6. Discussion

We have shown how the Tolerance Principle can model the learning trajectories of grammatical gender using Icelandic as a test case. Formal systems of gender assignment differ in terms of the productivity of individual gender assignment correlations, or, rules. Productive gender assignment rules are acquired early, like we demonstrated with the longitudinal corpus case study of Icelandic child speech.

The results of the child speech study also suggest that children acquiring Icelandic partition nouns in systematic ways and that this partitioning is guided by a search for productive cues to gender assignment. Icelandic does have productive gender assignment rules, but also lacks a productive gender default for a big subclass of nouns that are syncretic between all three gender classes. It seems that gender systems differ cross-linguistically with respect to the productivity of their individual gender assignment rules. Further work will apply the Tolerance Principle to languages with a full range of gender assignment productivity.

References


