Verb Innovations
In Spontaneous Speech of Children

Asya Achimova
Perm State University

Abstract
This paper is devoted to the acquisition of English verb morphology by English-speaking children aged 5-7. We also use data from Russian language acquisition. Child mistakes (or innovations as we call them) in formation of verb forms are studied. Special attention is paid to the factors that influence the choice of a verb paradigm in child speech. These factors are type and token frequency, cumulative frequency and phonological properties of a verb. The following claims are being tested: a) innovations and semantic errors are more frequent in narratives than in dialogues as narration requires longer planning. Therefore, less effort can be spent on grammar; b) occurrence of innovations in child speech is determined by token and type frequency as well as phonological properties of a verb.

1 Introduction

Acquisition of verb morphology is one of the main points in the development of a language system. Current interest in verb acquisition is supported by the existence of two controversial approaches: dual-system and single-system. They give different assumptions as to whether irregular and regular past forms of verbs are generated by one mechanism or by two distinct mechanisms (lexicon and grammar). Depending on the solution of the described issue two approaches draw different conclusions concerning the organization of mental language. That is why regular and irregular verb formation is an important issue to study.

In section 2, I will address theoretical questions raised in the research. Section 3 describes the data used for the study. Section 4 deals with the analysis of English and Russian data. These sections are followed by conclusions.

2 Theoretical background

2.1 Dual-system and single-system approaches to verb formation

The analysis of regular and irregular past forms has become an important tool used for understanding processes and mechanisms underlying cognition and language (Huang & Pinker: 2005). Marslen-Wilson and Tyler notice that regular and irregular past forms of verbs have been in focus of attention for more than 40 years already (Marslen-Wilson & Tyler 2006: 264). It probably happens so because the contrast between regular and irregular past forms is evident in English.

Modular or dual-system approach has been the leading one for several decades. Its main points are described in Pinker's work "Words and Rules" (Pinker 1991).
According to Pinker, irregular and regular past forms are processed by two distinct mechanisms. Regular verbs are generated in the module of grammar, while irregular forms occur due to the mechanisms of the lexicon. Forms of irregular verbs are stored in pairs. If a verb has a past form stored in the lexicon regular marking is blocked. Otherwise a rule applies. A rule is defined as an abstract operation of adding a suffix to a variable (root). Any time when access to a ready form is blocked for some reason the rule applies. The expansion of this rule on a wider variety of objects results in innovations, such as goed. Pinker notes that innovations built on irregular patterns are also possible yet very infrequent. Pinker shows that irregular patterns are phonologically dependent whereas regular suffix is employed disregard of phonological properties of a word. For example, it is usually added to newly created verbs and onomatopoeias.

Points discussed above are elaborated on in a later work by Steven Pinker and Michael Ullman “Past and Future of the Past Tense” (Pinker & Ullman 2002). First, clarifications are made concerning the term “rule” which was used to describe generation of regular past forms. It is noted that the theory does not proclaim the existence of an explicit rule “in order to make a past form it is necessary to ad –ed” (Pinker & Ullman 2002). The only thing that it does assume is that there is a past tense morpheme -ed, a variable (verb) and an operation of merger.

Second, Pinker and Ullman specify their position as for what types of verbs are stored in the lexicon and generated by the application of the symbolic rule. The initial theory viewed irregular forms as stored in the lexicon and regular past forms as generated by grammar. Now Pinker and Ullman qualify the assumption, saying that high frequency regular forms can also be stored in lexicon as well as all irregular forms.

Finally, Pinker and Ullman admit that lexicon in dual-system theories is characterized by some of the properties attributed to pattern associator which is a single-system model. The pattern associator model was first introduced by Rumelhart and McClelland (Rumelhart & McClelland 1986). Lexicon as well as pattern associator is able to generalize patterns. Input data are grouped according to various grounds. Then certain generalizations are made in the lexicon concerning morphological characteristics of units such as past tense building. Despite all common properties of lexicon and pattern associator model, they are not identical as lexicon in the understanding of the dual-system approach has certain morphological, phonological, semantic and syntactic features which are not taken into consideration in the connectionist model (a single-system model).

The single-system approach is represented in 2 types: connectionism (Rumelhart & McClelland 1986; MacWhinney & Leinbach 1991; Plunkett & Marchman 1991, 1993) and net model (Bybee 1995). Within the connectionist model interdependence of grammar and lexicon is underlined. The single associative mechanism processes incoming data and is able to generalize patterns. The stability and productivity, i.e. the ability to apply to new words are influenced by frequency effects and phonological properties of units.

It was noted above that connectionist models of pattern associator are limited as they do not take into consideration semantic and syntactic information. The argument about the limited character of single-system connectionist models is probably true only of early models of connectionism. Modern variants of connectionism take into consideration phonological features as well as semantics and argument structure. Language representation in modern connectionist theory is viewed as interconnection of three maps: phonological, semantic and argument. The key notion here is the
notion of competition which refers to relationship between language units activated during speech perception or production. On the phonological map compete auditory and articulate images; on the semantic map – meanings ascribed to sound signal (or articulatory sequence); and at last on the argument map compete arguments of a predicate and variants of word order. Competition of auditory images can be seen in self-correction, i.e. correction of slips of tongue including replacement of sounds similar in their articulation parameters or words that are phonologically similar (Ovchinnikova 2006).

Both modular and connectionist approaches correctly predict types and number of innovations. But the mechanisms underlying occurrence of innovations within modular and connectionist approaches are substantially different. The modular approach interprets innovations such as *comed* as a result of symbolic rule application. Innovations such as *brang* are viewed as a result of work of associative mechanisms in lexicon. On the contrary, single system approach regards all types of innovations within the same mechanism.

Within the connectionist approach the choice of a pattern occurs as a result of competition of various patterns. In this paper we will study what factors influence competition on the morphosyntactic arena.

### 2.2 Verb morphology

English verbs are generally divided into two groups: regular and irregular. Among irregular verbs there exist certain patterns or form building models. To the same pattern belong, for example, *ring-rang* and *sing-sang*. We can group verbs into a pattern disregard of their forms token frequency. Such method would be appropriate if we dealt with language as a system in general.

In our paper, we will adopt another approach. For us, it is important if a child is likely to hear a verb form. That is why we used frequency lists to determine token frequencies of verbs. We only took those verb forms which had a frequency not less than 1 item per million words. We will describe the process of grouping verbs into patterns in section 2 devoted to data description.

Russian verb morphology is more complicated. There are several productive and unproductive classes. Moreover, in Russian linguistics there are different classifications of verbs. In present research we adhere to approach presented in “Russian Grammar” (Shvedova 1980). There verb inflection types are defined as “groups of verbs belonging to the same conjugation type and having the same relation between present and past stems” (Shvedova 1980). Depending on the relation or changes between present and past stems 10 inflection types are discriminated. Among them we can see productive classes which present an active pattern in modern Russian and unproductive patterns which do not generate new verbs today.

Productive classes and subclasses are:

- class I:
  - subclass 1: *a-aj* alternation (igrala-igrayut ‘she played-they play’)
  - subclass 2: *e-ej* alternation (belela-belejut ‘she whitened-they whiten’),
- class II: *ova-uj* alternation (risovala-risujut ‘she drew-they draw’),
- class III: *nu-n* alternation (prignula-prignut ‘she jumped-they will jump’),
- class X:
  - subclass 1: *i-ø* alternation (milila-milyat ‘she washed with soap-they
2.3 Child language innovations

Verb acquisition is often described as U-shaped development. Dan Slobin (Slobin 1984) describes the following stages: 1) no marking - *break, drop*; 2) marking exists but in a limited number of cases - *broke, drop*; 3) overgeneralization of marking - *breaked, dropped*; 4) acquisition of standard marking - *broke; dropped*. Gradually, filters apply and a child masters a language.

The development of a morphological system is a gradual process. Unevenness of morphology acquisition results in innovations which tell us about the stages of system formation. Mastering a morphological category includes two stages: mastering its semantics and plane of expression (or marking). Semantics and form of expression have relative independence (Tseitlin 2000: 88). It is necessary to take into account this distinction as it allows classifying all mistakes into two types: those connected with mastering: 1) semantics - we will refer to them as errors; 2) form of expression - innovations. Under the term “innovation” (Tseitlin 1989) we understand forms which do not exist in standard language. We refer to forms like *goed* as innovations, not mistakes because a child does not yet know the whole system of a language and simply tries to fill in the gaps in the grammatical system creating new forms.

To the first group belong the cases of a wrong choice of tense or number in a certain context, for example, when a child does not take into consideration the sequence of tenses. The second group includes cases of building a past from according to a wrong paradigm, such as *maked*. We will regard in detail only the second group here.

The mechanism underlying innovations is generalization, i.e. the expansion of a pattern on a wider class of objects. It makes possible such forms as *goed* or *brang*. Dan Slobin (Slobin 1984) mentions the following rule governing generalization: “Avoid irregularities”. It means that a child will try to apply the same grammatical marking for one grammatical meaning. For example, if a child has noticed that –*ed* usually applies to past tense forms he or she will use such marking for other forms of verbs be they regular in standard language or irregular. Though the “knowledge” of the rule “avoid irregularities” is mostly subconscious, some children produce forms that seem to them more regular and refuse to admit exceptions that exist in standard language. Children even “explain” to adults that the invented forms sound better (Tseitlin 2000).

To sum up, innovations are the result of children’s creativity. These innovations are a valuable source of information as they show us the stages of language system development. In what follows we provide an analysis of factors determining innovations. Such analysis will give important material for discussion concerning dual and single-system approaches to verb formation.

3 Data

The data for our study was collected from several sources. First, we used a collection of monologues named “The Folk Stories of Children” (Sutton-Smith B. (ed.) 1981). This book contains narratives by children. To collect the narratives experimenters asked children to tell a story. Stories were recorded by an experimenter and later transcribed. From the whole set of stories we chose those
narrated by 5-7 year-old children. Then all mistakes were extracted and put in a list. The list included a mistake, the sentence where it was registered and information about a child (age, sex). Later we classified all types of mistakes into errors (semantic mistakes) and innovations (non-standard forms of expression of grammatical meaning).

Second, we used corpus data from the CHILDES database. Transcripts of speech of two English-speaking children were employed. The first one was a 7-year-old boy Christopher and the second was a 6-year-old boy First. The recordings contained their talks with parents. The procedure was the same as described above. As a result we got a list of errors and a list of innovations.

Evaluation of the influence of input on child speech is a complicated matter as it is raises the issue of adequacy of input modeling. For current purposes we need to determine the character of input for children in the age of 5-7.

For children aged 5-7 input is basically oral speech, both child-directed and non-directed one. For estimation of the impact of input we used two frequency lists based on oral corpora. List 1 was based on the oral part of the British National Corpus (BNC) (Leech, Rayson & Wilson 2003), while list 2 was based on parents’ speech from CHILDES (http://childes.psy.cmu.edu/topics/parentfreq.cdc). One of the purposes of the study is to compare various frequency lists.

We will use frequency lists to measure token frequency, type frequency and cumulative frequency. Token frequency is the frequency of a form in input per million words. Type frequency is a number of verbs having the same pattern. For instance a pattern including such pairs as bind-bound, find-found has a type frequency of 2. Cumulative frequency is counted by addition of token frequencies of past forms of one pattern.

Let us now describe briefly the frequency lists. BNC includes a collection of written and spoken texts of various genres. It comprises 100 million words. The list contains items with frequency over 1 item per million words (ipm). Out of the big list we made a small list containing verbs in the simple past form. Then the list was divided into 2 groups: regular and irregular verbs. Irregular verbs were then grouped according to patterns. Patterns have different type frequency and cumulative frequency. For example, the pattern get-got includes only 2 pairs: get-got, forget-forgot but the cumulative frequency of forms got and forgot is high. Pairs throw-threw, blow-blew, grow-grew make up a pattern with a higher type frequency.

The CHILDES list is based on parents’ speech extracted from the CHILDES database. The corpus basically contains talks during meals, telling stories, talks accompanying games etc (MacWhinney 2000, Li & Shirai 2000). From the frequency list we collected verbs in the past simple form. Later we grouped verbs according to classes (regular and irregular) and patterns as it was described above.

As Russian data we used a list of innovations collected by S.Tseitlin (Tseitlin 1989) and a “Dictionary of Child Language” edited by V.Kharchenko (Kharchenko 2005). Tseitlin’s list contained all types of innovations so we extracted from it innovations in verb formation. Kharchenko’s dictionary included both lexical and grammatical innovations of all parts of speech. We made a list of grammatical verb innovations.
4 Analysis of form building innovations

4.1 Narratives vs. Dialogues

We had two types of English data: monologues and dialogues. It has been noticed that in dialogues innovations were considerably less frequent. We have registered only one form *holded* that was built in the regular type and no innovations built on irregular patterns. It confirmed our initial assumption that the number of innovations is connected with the complexity of a cognitive task. For young children, monologues were more challenging than dialogues because monologues required longer planning. That is why innovations were more numerous in monologues.

4.2 Innovations connected with the choice of a paradigm and factors which determine them in English.

The question in focus is: what factors can determine the choice of a certain paradigm? It is also necessary to find out how these factors interact. Among possible factors we regard token and type frequency, cumulative frequency of past tense forms of one pattern and phonological features of a verb.

The most widespread innovation was building of a past form of an irregular verb within a paradigm of regular verbs (*stealed, knowed*). Children chose a model with a higher index of regularity, i.e. type frequency.

Our BNC list contains 224 past tense forms of regular verbs. They occur in total 7699 times so it is their cumulative frequency. Let us compare this figure with cumulative frequency of some irregular groups. For example, *got* and *forgot* occur together 8097 times, while *was* – 5068 times. Regular verbs class has a very high type frequency. The cumulative frequency is also relatively high.

Type frequency of regular verbs is even higher in CHILDES: it contains 500 different verbs. Their cumulative frequency is 7244. The distribution of regular verbs in two frequency lists is shown in table 1:

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>CHILDES</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of verbs</td>
<td>224</td>
<td>500</td>
</tr>
<tr>
<td>cumulative frequency of Simple Past forms</td>
<td>7699</td>
<td>7244</td>
</tr>
</tbody>
</table>

High type frequency of the regular class is probably the reason explaining why the majority of innovations are built according to this model. According to the regular type the following forms were constructed: *flied, breaked, holded* etc. This model may be considered the basic model.

Type frequency becomes a leading factor when a child deals with low frequency past forms. Our data confirm that such forms as *stole, fed, ran* (that might become *stealed, feeded, and runned* in child speech) have very low token frequency in adult oral speech (from 0 to 40 ipm, i.e. items per million words). We suppose that high type frequency determines the type of innovations, while low token frequency shows which forms are prone to innovations.

On the other hand, high token frequency can not guarantee successful resistance of an item to the influence of the regular pattern (Tseitlin 2000). For example, *knew*...
has a frequency of 168 ipm in adult speech, still there occurs a form *knowed* in child speech. Other studies report innovations like *comed* (Pinker 1991), cf. *came* – 473 ipm.

However, it is impossible to interpret all kinds of innovations as the result of type frequency effect. Several children used the form *brang* instead of *brought*. It seems to be very important to analyze this form as it occurred in speech of several children of different age. We can probably see here the influence of the model *ring-rang* (Pinker 1991).

The analysis of type frequency on the basis of the BNC list shows that the class *bring-brought* has 3 items (past forms: *brought, taught, and thought*). Other possible members are infrequent (less than 1 ipm). The class *ring-rang* has only 1 item (others infrequent). Type frequency does not work here.

According to the CHILDES list classes *bring-brought* and *ring-rang* have almost equal type frequency. The pattern *ring-rang* has higher regularity in parents’ speech than in the BNC. It includes 4 verbs: rang, sang, drank and sank. The pattern bring-brought includes 3 items.

All past forms of *ring-rang* pattern have low token frequency: 4-20 ipm (see table 2). The token frequency of *brought* is 175 in the CHILDES list (see table 3) and 143 in the BNC list. This difference is considered to be unimportant.

| Table 2. Token frequency of verbs of *rang-rang* pattern according to the CHILDES list |
|---|---|
| verb | ipm |
| drank | 20 |
| sang | 17 |
| rang | 12 |
| sank | 4 |
| total | 53 |

| Table 3. Token frequency of verbs of *bring-brought* pattern according to the CHILDES list |
|---|---|
| verb | ipm |
| brought | 175 |
| taught | 32 |
| thought | 583 |
| total | 790 |

The cumulative frequency of *ring-rang* pattern is 53 ipm. The cumulative frequency of “brought” group is 790 ipm (see tables 2, 3).

The factor of regularity does not seem to be important in the competition between two paradigms. It is not token frequency that plays the role either. In the BNC frequency list *brought* occurs 143 times, whereas *rang* – 40 times. In the CHILDES list token frequency of *brought* outnumbers even the cumulative frequency of “*ring-rang*” class (see tables 2, 3). It generally agrees with the data obtained from the BNC list.

The degree of regularity of classes “*ring-rang*” and “*bring-brought*” is practically the same though “*bring-brought*” has a higher type frequency. Nevertheless, even high type frequency cannot provide greater stability of *brought* in child speech. Children still make innovations such as *brang*.

A possible explanation can be presented in terms of phonology. Words *ring* and
bring are very close phonetically. Input from the phonological map serves as a resonance factor and makes the child choose a certain paradigm. It can also be a version that ring and bring form the same analogy like come and become, stand and understand.

A certain possibility is still left that a child could have heard a form like brang in adult’s speech. Elena Lieven (personal communication) notes that though the form brang does not belong to standard English it might be heard in Southern English dialects. Yet, even in these dialects it is rare. The form brang is possible in the Glaswegian dialect (Michael Kerins: personal communication). Nevertheless, it has never been registered in CHILDES or BNC. So the possibility of its occurrence could not affect the results of our research. Elena Lieven mentions, that in adult’s speech brang is possible as a linguistic joke. This probably is an indirect sign of the productivity of a pattern.

Preliminary results show that high token frequency is not a guarantee of a successful acquisition of an item. Type frequency does not always determine mastering of a pattern as well. Token and type frequency are important factors for discrimination of regular and irregular patterns, i.e. the patterns where the difference in type frequency is substantial. But these factors are not influential when a child deals with various irregular patterns. In other words, if a child feels that some irregular pattern should be employed, he/she needs to make choice between irregular models. But this choice will not be governed by frequency factors. Our data show that phonological features of verbs would be the most influential.

4.3 Evidence from Russian

English has always been chosen to test single- and dual-system approaches (see Pinker & Ullman 2002 for discussion) to language acquisition. It has given us evidence for understanding the nature of cognitive processes. Yet, some typological characteristics of English prevent us from getting certain type of data. First, English morphology is not as rich as of some other languages. Second, as mentions Eva Dabrowska (Dabrowska 2001), the mechanisms underlying formation of regular and irregular forms are different. Regular verbs form their past forms by adding a suffix while derivation of irregular verbs involves vowel changes which are cognitively more complex than adding an affix. Dabrowska supposes that it explains to us the fact that English verb innovations are mostly built on the regular pattern. Third, the existence of only one productive class does not allow investigating relations between several productive models.

Studying acquisition of morphology of languages such as Russian might give us new data. Several attempts have been made in research devoted to acquisition of verb morphology in Russian, e.g. Gor & Chernigovskaya 2003. As it has been described above, Russian has several productive and unproductive classes. Productive are the following classes and subclasses:

- class I:
  - subclass 1: a-aj alternation (igrala-igrayut ‘she played- they play’)
  - subclass 2: e-ej alternation (belela-belejut ‘she whitened-they whiten’),
- class II: ova-uj alternation (risovala-risujut ‘she drew-they draw’),
- class III: nu-n alternation (prignula-prignut ‘she jumped-they will jump’),
- class X:
  - subclass 1: i-ø alternation (milila-milyat ‘she washed with soap-they wash
with soap').

Other subclasses and classes are unproductive. Besides, the difference in type frequency between productive and unproductive classes as well as within different productive classes is not as evident as in English. Is it possible therefore to use dual or single-system approaches to account for Russian data? To answer this question we need to see if innovations are influenced by more than one productive class.

The data suggest that most innovations exhibit the influence of productive class 1: risovaju instead of risuju ‘I draw’. Alongside with innovations of class 1 there exist innovations influenced by other productive and unproductive classes. Such forms as priviknul instead of ‘privik’ ‘got used to’ are the result of productive class 3 influence. Traces of unproductive class 4 paradigm (ø-n) are seen in forms otdohnul ‘had a rest’; str’oh instead of stryahnuł ‘shook of’.

We have found out that there exist innovations influenced by more than one productive class. Classes are not distinguished on semantic basis but are the result of tradition in language. There are no direct rules how to learn which class a verb belongs to. It means that dual-system approach cannot fully account for Russian data as it does not regulate the relations between several productive (or regular) classes.

The existence of several productive classes in Russian also implies that for a child the choice of a proper paradigm will probably be governed by various factors including token frequency, type frequency, cumulative frequency of verb forms and phonological features of a verb. If this is the case, it is an argument in favor of single system approach to morphology acquisition, where various factors cooperate.

To estimate if the enumerated factors act for Russian as well as for English we need to use data from a frequency list based on the oral part of some Russian corpus. Unfortunately, today there are no such lists as existing frequency dictionaries are based on written corpora of Russian. As a future goal of the research we see the creation of such a dictionary.

5 Conclusion

Dual-system and single system approaches give right predictions concerning the number and the proportion of innovations on the regular and irregular types in English. Yet, they have different interpretation of the mechanisms producing innovations. The dual system approach is not fully applicable to languages with several productive classes.

We have found out that there are certain factors that influence the choice of verb paradigm. These are type and token frequency, cumulative frequency and phonological properties. Our analysis gives right predictions concerning the types of innovations and the ratio of innovations based on regular and irregular types. Properties of lexical units such as frequency and phonological shape allow us to account for innovations built on the regular and irregular type not using the notion of rule. To measure frequencies we used two independent dictionaries. We did not register differences in their data that could affect the results of our study.

Most innovations of English children belong to the regular type. According to our data it is explained by high type frequency of regular class. Other factors may become more influential for the choice of a paradigm within irregular patterns. Thus, in the case of brang, phonological similarity between ring and bring were the strongest factor.

English data have shown that innovations are more frequent in monologues. It
happens because more effort is spent on planning and less effort is left for language control.

We have confirmed the assumption that a child tries to find logic and regularity in a language both in regular and irregular classes of verbs and, therefore, he/she builds his/her own language system.

References


Data sources


CHILDES http://childes.psy.cmu.edu/

Parental Frequency Count http://childes.psy.cmu.edu/topics/parentfreq.cdc
