OVERGENERALIZATION OF VERBS
-
THE CHANGE OF THE GERMAN VERB SYSTEM

MARISA DELZ, BENJAMIN LAYER, SARAH SCHULZ, JOHANNES WAHLE

Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen, Wilhelmstrasse 19,
Tübingen, 72072, Germany
{marisa.delz; sarah.schulz; johannes.wahle}@student.uni-tuebingen.de
b-s-layer@gmx.de

In this paper we want to show, that the change of the German verb system from Middle High German to New High German can be simulated and explained by an adoption of the Iterated Learning Model. We claim, that the change of the German verb system is due to the frequency of usage and the process of overgeneralization. This is the first application of an Iterated Learning Model simulating the change of "real" German language, so we can also prove the quality of the Iterated Learning Model as an explanatory model for language change.

1. Introduction

The computational simulation of processes in natural language is a frequently discussed topic in current linguistics. The phenomenon debated in this paper, is the simplification of the German verb inflexion system in the change of time. In Middle High German the past tense form was formed either strong, weak, or irregular. The strong verbs are divided into seven classes. Behind every of the seven ablaut classes stands a regularity, which allows us to predict which verb belongs to which class.

In Modern High German the seven classes disappeared and two new main verb classes remained. This two categories are the regular and the irregular verbs. The regular verbs form the major category, where as for the irregular verbs only a small category is left. We claim, that the loss of the ablaut classes was not a "spontaneous" idea of the language but a regular evolutionary process. This is shown by the use of an Iterated Learning Model, which simulates the change in the German language. Our model could easily be applied to other languages as well.

2. The Middle High German verb system

The past tense of Middle High German verbs was mainly formed by two different morphological processes together with a very small class, the irregular verbs. One of the main processes to build the past form, was the *ablaut*, the so called *strong* form (see Table 1) and the other was the "*-(e)te*" suffix, the so called *weak* form. For our purpose all verbs, which have another past tense form were not taken into account.

The abluat is a qualitative and/or quantitative change of the vowel. It brings a regularity with it, which allows us to assign different verbs to seven different classes. All of these classes have in common that they refer to the indo-european e/o-ablaut. A sound change phenomena, which happened on the way from the indo-european language to Middle High German, caused the differentiated sound cluster as it appears in Middle High German. The regularities characterizing the ablaut classes refer to the phonological environment of the stem vowel. The verbs, for example, which belong to the first class, are divided into two subclasses. The first one is characterized by an /î/ as the stem vowel, which is followed by any consonant except /h/, /r/ or /w/. This combination of sounds causes an /ei/ as stem vowel in the past form (see Table 1). The second subclass is characterized by the same stem vowel - /î/ -, but if this vowel is followed by /h/, /r/ or /w/ the stem vowel changes in the past tense to /ê/ (see Table 1); Paul (1989) uses for every class another characteristic feature to build up the ablaut classes but it ends up the same.

Table 1. ablaut class I

	present		past			
class	*infinitive*	*1. Pers. Sing. Ind.*	*1. Pers. Sing. Ind.*	*1. Pers. Pl. Ind.*	*participle*	*translation*
I a	rîten	rîte	reit	riten	geriten	to ride
I b	zîhen	zîhe	zêch	zigen	gezigen	to blame

The past tense form of the weak Middle High German verbs was formed by adding only the dental suffix *-(e)te* to the verb stem. An example for a Middle High German weak verb is legen - leg*ete* (to lay). The verbs which belong to this group are formed secondarily, i.e. they were derived from strong verbs, adjectives or nouns. There are additionally some verbs which are primarily in the group of the weak verbs (c.f. Paul, 1989). The *-(e)te* suffix was probably the result of the cliticalization of the past tense form of *tuon* (to do) which was used before for a periphrastic past form (c.f. Hennings, 2003).

3. The Modern High German verb system

In Modern High German, there are two main categories of verbs left, the irregular and the regular verbs. The category of the regular verbs is the major category. The regular verbs form their past tense with the suffix *-(e)te* (c.f. Eisenberg, 2009, p.449). This rule facilitates the German grammar and makes it easier for learners to learn the language. The irregular verbs are the smaller category. This category is actually divided into two subcategories, the strong verbs and the irregular verbs. The irregular verbs have an ablaut and the same suffix as the regular verbs in the past tense, namely the suffix *-(e)te*, e.g. denken (to think)- d*a*ch*te*. These verbs are also called mix conjugations (c.f. Eisenberg, 2009, p. 454). The strong verbs, are verbs with a completely different past tense form then the others, e.g. reiten (to ride)- *ritt*. For our purpose this difference is not needed. Both subcategories belong to the category of irregular verbs. Compared to the regular verbs, the past tense form of the irregular verbs need to be saved in the mental lexicon.

The category of the regular verbs increased over time and the category of the irregular verbs decreased. One reason for this change is the frequency of usage. The irregular verbs have a higher frequency of usage, than the regular verbs (c.f. Nübling, 2008, p. 57). Most of the irregular verbs, mainly the strong verbs, belong to the basic vocabulary of the language (c.f. Eisenberg, 2009, p. 440). Although they need to be saved in the mental lexicon, it is profitable for a learner to memorize these verbs. The irregular verbs have a lower frequency of usage. Therefore it is profitable for a learner to form them via a rule. The other reason is the frequency of types. The regular verbs have a higher frequency of types, than the irregular verbs. New verbs in a language are mainly derivations of nouns, adjectives or names and are inflected regular, e.g. voll (full)- füllen, füll*te*. Adopted verbs from French or English are inflected regular, e.g. to jog- joggen, jogg*te*. Irregular verbs, which are not used frequently, are now formed regular, too (c.f. Eisenberg, 2009, p. 456).

These reasons causes the increase of the regular verbs and also the decrease of the irregular ones. Although the irregular verbs decrease, they will never disappear, because their belonging to the basic vocabulary and their high frequency of usage saves them from disappearance.

4. The Iterated Learning Model

Our solution, for this phenomenon, is based on the idea of *iterated learning*. The Iterated Learning Model (ILM in the following) is based on the hypothesis, that not the shape of human language capability forms the language, but that languages are pushed to certain shapes due to the way they are passed from generation to generation. The speakers learn the language the preceding generation uses and pass this learned language on to the next generation and so on.

The crucial term in this context is that of the *bottleneck*. The bottleneck describes the narrow connection between the languages of the speaker and the learner, i.e. the language internalized by a certain speaker and the language as it is used by a population. The connection is bidirectional. The language of the teacher influences the learners language and the learners language will become a teacher language. As stated above, neither reception nor production is complete. No speaker produces all possible phrases and none perceives all phrases he will possibly produce. For example, a speaker of a language is capable to inflect several thousand verbs in the correct way, even if he has never heard all of the possible forms. He will abduct rules from forms he knows and will try to build the required form respecting this rules. During the process of passing over the langeuage from one to the other generation, the bottleneck regularities may develop (c.f. Dowman, Kirby, & Griffiths, 2006) and it is even possible to show implementations of the ILM, how compositionality (c.f. Smith, Kirby, & Brighton, 2003) and recursion (c.f. Kirby & Hurford, 2002) develop over time.

The central element of every ILM is a learning algorithm allowing generalization over learned forms. There are two competing mechanisms: the holistic one, just looking at the data and saving it, and the generalizing one, comparing the data and calculating a probability for each inflexion form. To provide such learning, different algorithms have been used. The most frequent ones are based on neuronal networks (c.f. Kirby & Hurford, 2002; Smith et al., 2003) and grammar induction (c.f. Kirby & Hurford, 2002).

5. Statistical Database

To prove the ILM approach and to simulate the development of the German verb system we tried to construct a realistic database. It is built up by an analysis of the digitized Lexer (Burch, 2002) and Ruoff (1981). The search engine provides a mechanism, which allows a traditional classification of the verbs. The query was done by choosing the box for the ablaut class consciously. The regular verbs were searched the same way. The values for the verb classes were summed up, like in Table 2, and the relations of the ablaut classes to each other were taken over to an input file for the model.

Then, the frequency values of Ruoff (1981) were also taken over into the input file. The list was completed, by adding the historically correct ablaut classes. With this method it was possible to create a realistic input file.

6. The model

We developed a simulation of the process of overgeneralization. The model simulates the change of the verb classes from Middle High German to Modern High German. Since this process took place over many generations of German speakers the approach of an Iterated Learning Model seems to be an appropriate framework for the implementation. The implementation should take three main factors into account: the frequency of the verbs, the size of the bottleneck and the time of

Table 2. Statistic of Middle High German verbs

		absolute	relative
complete		10879	100%
regular		9727	89.41%
irregular (strong)		1152	10.59%
from them	"Ablautreihe"		
	1	238	2.19%
	2	171	1.58%
	3	288	2.65%
	4	99	0.91%
	5	115	1.06%
	6	96	0.88%
	7	145	1.33%

iteration, i.e. the number of generations. In addition to this, there are some other assumptions which characterize the simple model.

The population at any one time consists of just one speaker teaching the language and one hearer learning the language. Because of this assumption, the frequency of a verb has a strong influence on the evolution of a verb. Therefore, we keep the frequency of the verbs fixed over the generations and the verb database does not change. The model does not take reinvention of verbs into account. We start off with a database containing 327 verbs. Any database needs to fulfill the criteria of the input file.[a] Every learner has to assign an inflexion class to every verb. Therefore, a final verb inflexion distribution for the whole database has to be found in every generation. The whole vocabulary must be learned before a learner becomes a teacher. That means that the teacher's vocabulary consists of two parts: a part that was learned by hearing a verb and saving it into the memory and another part compiled by overgeneralization.

The inflexion classes are the only dynamic structure in this model. We start with a realistic distribution of verb inflexion classes and update it every generation. Every speaker can just teach what he has learned including the inflexion that was learned by overgeneralization. This is maybe the most crucial point of this program. The whole overgeneralization mechanism relies on the frequency of the inflexion classes of the learned words. As every ILM this model is first running through the process of one generation and iterates over it. The procedure of one iteration step is shown in Figure 1. This model provides the results of every iteration step as its output. The output pictures the distribution of the verb inflexion class of every learner generation after the overgeneralization.

Once a word was heard it is put into the learners memory, which is the basis for

[a] An input file should contain a verb, its frequency in percent and a inflexion class number per line.

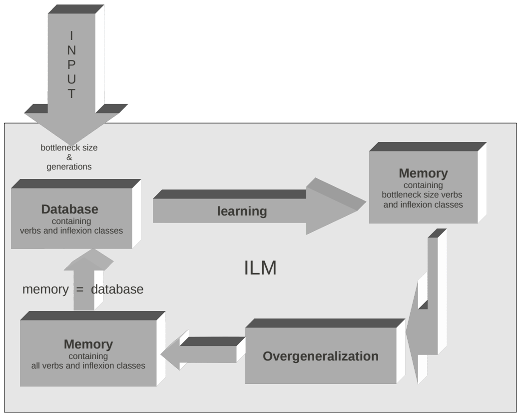


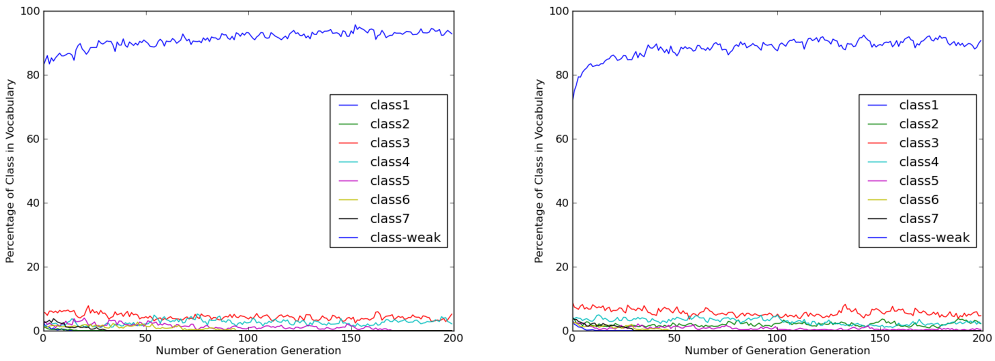
Figure 1. Architecture of the ILM of verb overgeneralization.

the calculation of the probability of an inflexion class for the overgeneralization step. After the learning step a distribution of the inflexion classes is calculated dependent on the already learned vocabulary. This distribution builds the foundation for the overgeneralization probabilities. That means, the vocabulary is learned by just saving heard verbs into the memory and afterwards the missing verbs and their inflexion classes are learned on the basis of vocabulary already saved in the memory. The generation number determines the termination case for the iteration loop over the whole learning process including memorizing heard words and overgeneralization.

At this point a reference to the structure of the database and the implementation of inflexion classes should be made. The memory is structured in that way that it does not contain inflected verbs, but the information about the inflexion class a verb belongs to. This inflexion classes are in turn related to the verb database via parallelism. The database is just used as a reference list. The computation itself is abstracted from the language.

7. Analysis

Now running the algorithm over our data set of 327 different verbs, the output graph (see figure 2(a)) shows the tendency that one can expect considering the development in the actual German language. The graphs were generated by averaging over ten different outputs. While several of the strong inflection classes die out, others remain stable after some time of decrease, and the class of weak inflected verbs grows and approaches 100%.



(a) Example output with a bottleneck size of 50%

(b) Example output with a bottleneck size of 50% for changed database

Figure 2. output graphs

The figure plots the percentage of every inflection class in the vocabulary after both learning and generalization over generations.

Due to the statistical database (see Table 2) the size of the seven ablaut classes is not that big, that clear boundaries can be seen at the starting point. But nonetheless the graph shows the general development of the German verb system, the increase of the class of the weak verbs and the decrease of the other classes. Trusting the graph there should be no more verbs belonging to class 1,2 and 5-7. But there are indeed such verbs, which belong to this classes. This result may be due to the fact, that our model uses only one learner per generation and other language "conserving" factors (e.g. written documents, school grammars, etc.) were not taken into account. Our database starts with a very high percentage of weak inflected verbs, the effect is only small but nevertheless noticeable. But the effect stays stable even when decreasing the initial percentage of weak verbs significantly. Figure 2(b) shows an output starting with a fraction of weak forms about only 70%. The result is an even stronger increase in the weak form whilst the first ten or twenty generations.
The size of the bottleneck provides a way to influence the resulting graph. Augmenting the bottleneck size results in more stable plots, particularly when raising it above 100%, this means that 327 verbs are drawn from the pool. In general this comes up to only slightly above 100 different verbs. Lowering the size of the bottleneck leads to a fast disappearance of classes. Outputs that produced realistic

graphs were created for bottleneck sizes between 50% and 100%. Taking a look at the actual verbs in the last generations it stands out that mainly the frequent verbs persist in the strong inflection classes over the generations while the rare ones tend to change their classes.

8. Conclusion

The ILM seems to offer an adequate framework to replicate the development in the German inflection system over the last 500 years. The learning and generalization mechanism we use is based on the simple assumption that there is a tendency to inflect unknown verbs in the same way as most of the others. The pressure of learning through a bottleneck, creates some kind of magnet effect, attracting forms from smaller classes to the bigger ones. The complex inflexion system based on verb classes disappeared and changed to a system with only weak and strong forms, whereby the strong forms are remnants of the middle high German inflection classes. With the model, we can simulate the change in the German language from Middle High German to Modern High German and it proves that the change is due to the frequency of usage and the process of overgeneralization.

References

Burch, T. H. (Ed.). (2002). *Mittelhochdeutsche wörterbücher im verbund.* Stuttgart: Hirzel [u.a.].

Dowman, M., Kirby, S., & Griffiths, T. L. (2006). Innateness and culture in the evolution of language. In (p. 83-90).

Eisenberg, K. R., Peter ; Kunkel-Razum (Ed.). (2009). *Der duden in zwölf bänden* (Vol. 4: Duden - Die Grammatik : unentbehrlich für richtiges Deutsch;). Mannheim: Dudenverl.

Hennings, T. (2003). *Einführung in das mittelhochdeutsche.* Walter de Gruyter.

Kirby, S., & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In (p. 121-148). London.

Nübling, D. (Ed.). (2008). *Historische sprachwissenschaft des deutschen : eine einführung in die prinzipien des sprachwandels* (2. überarb. Aufl. ed.). Tübingen: Narr.

Paul, H. (1989). *Mittelhochdeutsche grammatik.* Niemeyer.

Ruoff, A. (1981). *Häufigkeitswörterbuch gesprochener sprache: gesondert nach wortarten, alphabetisch, rückläufig alphabetisch und nach häufigkeit geordnet.* Max Niemeyer Verlag.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*.