An agent-based model of a historical word order change

Jelke Bloem **Arjen Versloot** **Fred Weerman**

Amsterdam Center for Language and Communication
University of Amsterdam
1012 VB Amsterdam, Netherlands
{j.bloem, a.p.versloot, f.p.weerman}@uva.nl

Abstract

We aim to demonstrate that agent-based models can be a useful tool for historical linguists, by modeling the historical development of verbal cluster word order in Germanic languages. Our results show that the current order in German may have developed due to increased use of subordinate clauses, while the English order is predicted to be influenced by the grammaticalization of the verb *to have*. The methodology we use makes few assumptions, making it broadly applicable to other phenomena of language change.

1 Introduction

Agent-based modeling is a method for simulating the behaviour of individual agents (i.e. a speaker of a language) in a larger community of agents (i.e. all speakers of the language). While agent-based models have been successfully used as tools in the field of evolutionary linguistics to study how linguistic structures may have emerged, they have not yet spread to the field of historical linguistics, which is more interested in describing and modeling change in existing natural languages. Both fields are concerned with changing language models, although the starting assumptions and context are different. In historical linguistics there is data available about structures in earlier and more modern states of the language, while in evolutionary linguistics the structures have to emerge from the implemented mechanisms. Nevertheless, the mechanisms described, such as grammaticalization, are often similar and lend themselves to study using similar methodology.

In the field of evolutionary linguistics, agent-based models are used to model language as a complex dynamic system, whose structure depends on the interactions of its speakers. An early overview of such work is provided by Steels (1997), who emphasizes the possibilities of modeling various aspects of language in this way. Among this work is a study by Briscoe (1997) on the default word order of languages, though it assumes a framework of universal grammar in which learning consists of setting parameters. Subsequent work included the application of this method to specific domains of linguistics, such as the emergence of vowel systems (De Boer, 2000) and the development of agent-based models specific to language, such as the iterated learning model of Kirby and Hurford (2002). Language change was often only discussed in terms of the emergence of new structures, and lacked comparisons to historical data (de Boer and Zuidema, 2009), or used artificial languages, as noted by Choudhury et al. (2007), whose own work is an exception. A few other studies that relate to historical linguistics can be found. Daland et al. (2007) and Van Trijp (2012) model some apparent idiosyncrasies in inflectional paradigms of natural languages, Daland et al. (2007) doing so with a model that includes social structure, and Van Trijp (2012) using the Fluid Construction Grammar framework. A further example is Landsbergen et al. (2010)'s study that models some mechanisms of language change from the perspective of cultural evolution. Overall, agent-based language studies informed by historical data are not widespread, and often involve many assumptions or dependence on a framework. A recent exception to this is a study by Pijpops and Beuls (2015) on Dutch regular and irregular verbs.

Our emphasis in this work is on creating an agent-based model that makes minimal assumptions, in order for the presented methodology to be useful for any theory of language that allows for functionalism in language change. Our case study, the historical development of verbal cluster order in Germanic languages, involves a word order variation in which multiple constructions are grammat-

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, pages 22–27,
Lisbon, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

ical. This kind of phenomenon has not been investigated with an agent-based model before. Besides syntactic analyses (Evers, 1975), recent work on verb clusters has also discussed non-syntactic factors influencing word order, using frequency-based methods (De Sutter, 2005; Arfs, 2007; Bloem et al., 2014) and historical data (Coussé, 2008). We follow up on this line of work with our agent-based model, in which a functional bias induces language change. Using this model, we will show how the current orders of verb clusters in modern West-Germanic languages might have developed and diverged from the proto-Germanic cluster orders.

In the next section, we briefly outline the phenomenon of verbal cluster order variation. We then describe the methodology of the simulation and its initial state, followed by the results and a discussion of those results

2 Verbal clusters

Many verbal cluster word orders are attested in different Germanic languages (Wurmbrand, 2006). We will illustrate this with a Dutch example, a language where the ordering of these verbs is relatively free. In two-verb clusters, the finite verb can be positioned before or after the infinitive:

(1) Ik denk dat ik het **heb begrepen**.
I think that I it have understood
'I think that I have understood it'

(2) Ik denk dat ik het **begrepen heb**.
I think that I it understood have
'I think that I have understood it'

In the literature, construction 1 is called the 1-2 order (ascending order or red order), and construction 2 is called the 2-1 order (descending order or green order). Both orders are grammatical in Dutch, and express the same meaning, though there are differences in usage. German and Frisian only allow order (2) for two-verb clusters, while English and Scandinavian languages only allow order (1) [1]. Despite these differences, all of these languages evolved from Proto-West-Germanic.

This raises the question of why some of the West Germanic languages ended up with verbal clusters in 2-1 order, and others with the 1-2 order. To study this, we need to select some factors that may have influenced the change, and the best place to look for this is the Dutch language, in which both orders are possible. Language variation often indicates language change, with the variation being a state of transition from one structure to another, in which both structures can be used. Factors that correlate with different word order preferences in modern Dutch may therefore be involved in the change as well.

% 1-2	mod+inf	have+PP	cop+PP
main	97%	50%	10%
sub	80%	50%	5%

Table 1: Reconstructed proto-Germanic probabilities for the 1-2 order.

The order variation in Dutch has been claimed to be an instance of language change in progress. In the 15th century, the 2-1 order was used almost exclusively. After this, the 1-2 order starts appearing in texts, and becomes increasingly frequent, moving towards the current state of the language (Coussé, 2008). This was not the first time the 1-2 order had been attested though, it also appears in some of the oldest Dutch texts.

3 Methodology

Our simulation consists of a group of agents that can function as speakers and recipients of verbal cluster utterances. Each agent has its own instance of a probabilistic language model that stores and produces such utterances. We will first describe the language model and the linguistic features of verbal clusters that it stores, and then we will explain what happens when the simulation is ran and the agents interact.

To find linguistic features that may be associated more with one order than with the other, we rely on synchronic corpus studies of Dutch, the language in which both orders are possible. Associations have been found with a variety of factors, including contextual factors such as regional differences between speakers (Coussé et al., 2008). When creating a language model for an agent, we are only interested in factors that may cause a particular speaker (or agent) to choose a particular word order. A recent study found that verbal cluster order variation correlates with both constructional factors (the use of a particular linguistic form) and processing factors (such as sentence length) (Bloem et al., 2014). We will examine only the construc-

[1] English and Scandinavian verb groups are generally not called verb clusters in the literature because they can be interrupted by nonverbal material, but for the purposes of this study the distinction is not important.

tional factors, because those are likely to be stored in the lexicon with their own associated word order preferences. The most important of these are the main clause / subordinate clause distinction (there are more 2-1 orders in main clauses), and the type of auxiliary verb (there are more 2-1 orders when a copula verb is used in a cluster). These two factors not only have different order preferences in modern Dutch, but have also undergone historical changes that may have triggered our word order change: subordinate clauses have become more prevalent, and one type of auxiliary verb, *to have*, grammaticalized during the time period we are interested in.

We will assume that the two factors, clause type and auxiliary type, are stored as features, each with their own word order preferences. This way of storing features is based on the bidirectional model in Versloot (2008), though our models learn by interacting rather than iterating.

Table 1 shows all of the possible combinations of feature values a verbal cluster can have in our model. Our model assumes two clause types (main and subordinate) and three different types of auxiliary verbs, reflecting the historical sources of verb clusters:

1. Clause type feature
   (a) Main clause context
   (b) Subordinate clause context
2. Auxiliary type feature
   (a) modal + infinitive: the origin of verb clusters in Germanic
   (b) 'to have' + participial main verb (PP): arose only later in history to extend the possibilities of expressing temporal and aspectual features
   (c) copula + PP: originally a passive, predicative, construction — not purely verbal, rather adjectival.

A cluster can have either of two word orders: the 1-2 and the 2-1 order.

The simulation consists of a language agents, each starting out with n exemplars of verbal clusters, stored in the agent's language model. An agent's language model contains the type of information shown in Table 1: for each possible combination of feature values, exemplars are stored. In addition to their features they have the property of being either in the 2-1 or 1-2 order (from which a percentage can be calculated, as in the table). The agents' language models do not contain any other structures. We did not use an existing framework in order to have as few parameters as possible. The simulation was implemented in the Python programming language.

When the model is run, each run consists of $a * n * i$ interactions. In an interaction i, a random agent is picked as the speaker and another random agent as the recipient. The speaker agent generates a verbal cluster based on its language model, and the recipient agent stores it as an exemplar. When a speaker agent generates a verbal cluster, it picks the features of a random exemplar from its language model, and then assigns word order based on the word order probabilities of both of its features individually. A 1-2 (ascending) realization of a modal subordinate clause cluster may be produced according to the following:

$$P(asc|x) = P(asc|x_{sub}) + P(asc|x_{modinf}) \quad (1)$$

where x is a set of feature values. $P(asc|x_{sub})$ is the probability of a subordinate clause being in 1-2 order, and $P(asc|x_{modinf})$ for the modal+infinitive construction type. These probabilities are calculated from the stored frequency of the features in 1-2 contexts:

$$P(asc|x_{sub}) = \frac{F(sub, asc)}{F(sub)} \quad (2)$$

So, the probability of a modal subordinate clause cluster being expressed in the 1-2 order depends on how many exemplars the agent has stored in which a subordinate clause cluster was in the 1-2 order (relative to 2-1), as well as exemplars in which a modal cluster was in the 1-2 order (relative to 2-1). Example (3) is an example of a modal subordinate clause cluster in 1-2 order, though our language model is more abstract and does not use actual words, only the features.

(3) Ik denk dat ik het **wil** **horen**
I think that I it want hear
'I think that I want to hear it'

After producing this exemplar, the agent normally deletes it from its own storage, because we do not want the relative frequencies of the various feature values (i.e. the number of copular verbs) to vary randomly. We are only interested in the word order. Furthermore, this avoids an endless growth of the agents' language models. Only when a growth factor applies, this deletion does not happen. The simulation includes two growth factors g_have and g_sub to simulate two relevant historical changes: the grammaticalization of 'to have' as an auxiliary verb, and an increase in the use of subordinate clauses. When these growth factors are set to 1, after every a interactions, an exemplar with the relevant feature is kept where it otherwise would have been deleted from the language model. A growth factor of 2 doubles the rate. g_have applies while there are fewer *have*-clusters than clusters of either of the other types, and g_sub while there are fewer subordinate clause clusters than main clause clusters.

When an agent is the recipient of a verbal cluster exemplar, it simply stores it in its language model, including the word order. So, when example (3) is perceived, the 1-2 order production probability of subordinate clause clusters and that of modal clusters will go up (separately) in the language model of the recipient agent. A critical learning bias is simulated here: the tendency to decompose an utterance into features and storing information about the features, rather than storing it as a whole. This is the only assumption we make about the language faculty in this model, and it is a functional one. It simulates the fact that people do not perfectly copy a language from each other.

We initialize each experiment with 30 agents ($a = 30$), and $i = 5000$ to simulate a long time course in which simulations will almost always stabilize in the end. With fewer agents, some agents lose all of their exemplars during the simulation. Each agent starts with a language model of 73 exemplars ($n = 73$) that follows frequency patterns as reconstructed for 6th century Germanic, based on a comparison of verb cluster frequencies in Old English, Old High German and Old Frisian texts. These figures are also summarized in table 1. For any unattested combination of features and word order a single exemplar is included to simulate noise.

4 Results

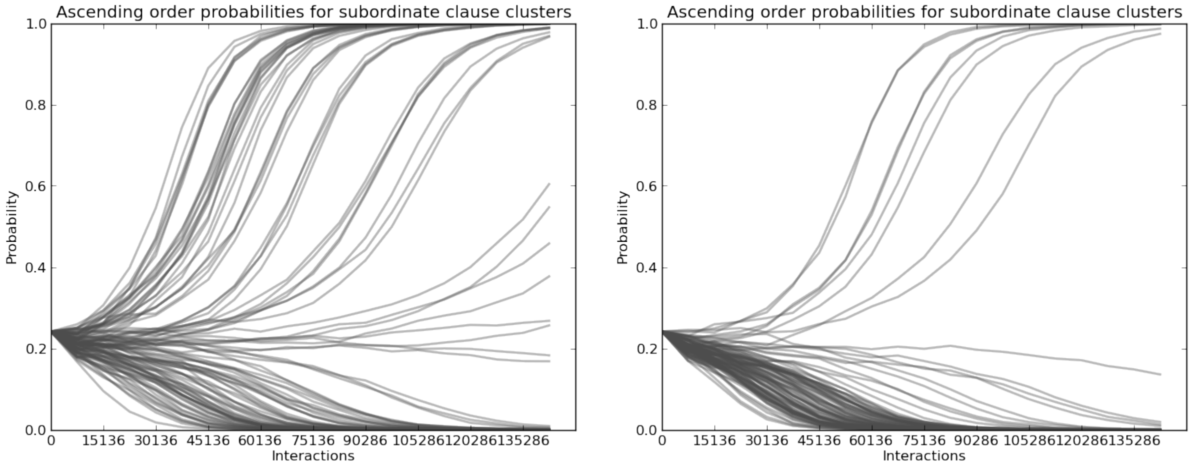
Figures (a) and (b) show example results of the agent-based model simulation, with different parameter settings. The graphs show the results of 50 different simulation runs overlaid, each run being a possible language. The X-axis represents time (in number of interactions) and the Y-axis represents the proportion of 1-2 orders, a value between 0 and 1. The proportions are calculated over all of the agents in the simulation. When the simulation is ran for long enough, it will always stabilize into a situation where a language either has only 1-2 or only 2-1 orders, though some feature combinations stabilize faster than others. Due to space constraints, we only show results for subordinate clause clusters (with any auxiliary type), but the general patterns are similar for all of the features, though some change sooner than others. We can observe that the model correctly predicts both languages with dominant 1-2 orders such as English, and dominant 2-1 orders as in German.

However, a model that predicts everything is not very interesting. We would like to know when a language in the model becomes English-like or German-like. We can do this by changing the growth factors: the rise of subordinate clauses (g_sub) and of *to have* (g_have). Figure (a) shows simulations in which *to have* grammaticalizes faster, while in Figure (b), subordinate clauses catch on more quickly. A clear difference can be observed — Figure (a) shows more languages gaining English-like 1-2 orders (35% 1-2, 56% 2-1 and the rest had not stabilized yet), while Figure (b) shows more German-like 2-1 orders (92% 2-1, 7% 1-2). Different speeds of grammaticalization of *to have* and growth of subordinate clauses result in different dominant word orders.

5 Discussion

With this study, we hope to have shown that an agent-based model with just a single learning bias can be used to gain insight into processes of change in natural languages, and generate new hypotheses. Specifically, the model makes two predictions: that *to have* grammaticalized faster in English, and that subordinate clauses gained use more quickly in German. These predictions can be tested using historical corpora of these languages in future work.

In the model, the 2-1 order is supported by subordinate clauses. Due to verb-second (V2) movement in these languages, the finite verb (the 1) precedes the other verb in main clauses (the 2). This 1-2 order differentiates main clauses from subordinate clauses, motivating the preservation of a 2-1 order in the subordinate clauses. Increased use of subordinate clauses may then have supported the 2-1 order as the default order. However, if *to have* grammaticalizes earlier, the 1-2 order is sup-



(a) 50 runs with faster growth of have+pp constructions ($g\_have = 2$, $g\_sub = 1$)

(b) 50 runs with faster growth of subordinate clauses ($g\_have = 1$, $g\_sub = 2$)

ported. This new grammatical verb becomes associated with the most prevalent word order at the time, and pushes the language further in the direction of that word order. In the beginning this is the 1-2 order, more associated with main clauses in proto-West-Germanic due to V2 movement, but later on the 2-1 order is more prevalent, due to its association with subordinate clauses.

Our model cannot yet account for the current state of the Dutch language, which first moved towards mainly 2-1 orders like German, and then shifted towards 1-2 orders again (Coussé, 2008), a change that is still in progress. There is evidence that the 1-2 order has become the default order (Meyer and Weerman, submitted), and this second change was likely caused by a factor outside the scope of our model, such as language contact.

Nevertheless, we believe that agent-based modelling can be a useful tool for historical linguists, particularly those working with frequency-based explanations. The present work and the study of Pijpops and Beuls (2015) show that testing of different mechanisms and parameters in a simulation, informed by historical data, can provide additional evidence for theories on what may or may not have been possible in a case of language change, given the assumptions built into the model. We believe it is particularly interesting to test how few assumptions are necessary to explain the observed historical data, which previous work has not focused on.

We would like to emphasize that this method is applicable to other cases of language change in which the use of structures changed over time. Any processes of historical change that can be captured in terms of frequencies and features may be used as factors to be investigated, and the fact that the model makes few assumptions also means that no particular social or cultural phenomena need to have happened for the model to be applicable. However, these simplifications also limit the extend of what can be modeled. In future work, contact phenomena could be simulated by including non-learning agents, or influxes of agents with different language models. Subsequent work on other cases of historical change may need to include such additional assumptions, if they are known to have been historically relevant.

References

Mona Arfs. *Rood of groen? De interne woordvolgorde in tweeledige werkwoordelijke eindgroepen met een voltooid deelwoord en een hulpwerkwoord in bijzinnen*. Göteborg University, 2007.

Jelke Bloem, Arjen Versloot, and Fred Weerman. Applying automatically parsed corpora to the study of language variation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1974–1984, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/C14-1186`.

Ted Briscoe. Co-evolution of language and of the language acquisition device. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 418–427. Association for Computational Linguistics, 1997.

Monojit Choudhury, Vaibhav Jalan, Sudeshna Sarkar, and Anupam Basu. Evolution, optimization, and language change: The case of bengali verb inflections. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 65–74. Association for Computational Linguistics, 2007.

Evie Coussé. *Motivaties voor volgordevariatie. Een diachrone studie van werkwoordvolgorde in het Nederlands.* Universiteit Gent, 2008.

Evie Coussé, Mona Arfs, and Gert De Sutter. Variabele werkwoordsvolgorde in de Nederlandse werkwoordelijke eindgroep. Een taalgebruiksgebaseerd perspectief op de synchronie en diachronie van de zgn. rode en groene woordvolgorde. *Taal aan den lijve. Het gebruik van corpora in taalkundig onderzoek en taalonderwijs*, pages 29–47, 2008.

Robert Daland, Andrea D Sims, and Janet Pierrehumbert. Much ado about nothing: A social network model of russian paradigmatic gaps. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 936. Citeseer, 2007.

Bart De Boer. Self-organization in vowel systems. *Journal of phonetics*, 28(4):441–465, 2000.

Bart de Boer and Willem Zuidema. Models of language evolution: Does the math add up. *ILLC Preprint Series PP-2009-49, University of Amsterdam*, 2009.

Gert De Sutter. *Rood, groen, corpus! Een taalgebruiksgebaseerde analyse van woordvolgordevariatie in tweeledige werkwoordelijke eindgroepen.* University of Leuven: PhD thesis, 2005.

Arnold Evers. *The transformational cycle in Dutch and German*, volume 75. Indiana University Linguistics Club Bloomington, 1975.

Simon Kirby and James R Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language*, pages 121–147. Springer, 2002.

Frank Landsbergen, Robert Lachlan, Carel ten Cate, and Arie Verhagen. A cultural evolutionary model of patterns in semantic change. *Linguistics*, 48(2):363–390, 2010.

Caitlin Meyer and Fred Weerman. Cracking the cluster: The acquisition of verb raising in Dutch. *Manuscript in preparation*, submitted.

Dirk Pijpops and Katrien Beuls. Strong "island of resilience" in the weak flood. Dutch strategies for past tense formation implemented in an agent-based model. Presented at Computational Linguistics in the Netherlands (CLIN) 25, Antwerp, 2015.

Luc Steels. The synthetic modeling of language origins. *Evolution of communication*, 1(1):1–34, 1997.

Remi Van Trijp. Self-assessing agents for explaining language change: A case study in german. In *ECAI*, pages 798–803, 2012.

Arjen Pieter Versloot. Mechanisms of language change: vowel reduction in 15th century West Frisian. 2008.

Susi Wurmbrand. Verb clusters, verb raising, and restructuring. *The Blackwell companion to syntax*, pages 229–343, 2006.