OXFORD

GERMANIC PHYLOGENY

FREDERIK HARTMANN

OXFORD STUDIES IN DIACHRONIC & HISTORICAL LINGUISTICS

Germanic Phylogeny

OXFORD STUDIES IN DIACHRONIC AND HISTORICAL LINGUISTICS

General editors
Adam Ledgeway and Ian Roberts, University of Cambridge

Advisory editors

Cynthia L. Allen, *Australian National University*; Ricardo Bermúdez-Otero, *University of Manchester*; Theresa Biberauer, *University of Cambridge*; Charlotte Galves, *University of Campinas*; Geoff Horrocks, *University of Cambridge*; Paul Kiparsky, *Stanford University*; David Lightfoot, *Georgetown University*; Giuseppe Longobardi, *University of York*; George Walkden, *University of Konstanz*; David Willis, *University of Oxford*

RECENTLY PUBLISHED IN THE SERIES

44
Romance Object Clitics
Microvariation and Linguistic Change
Diego Pescarini

45
The Diachrony of Differential Object Marking in Romanian
Virginia Hill and Alexandru Mardale

46
Noun-Based Constructions in the History of Portuguese and Spanish
Patrícia Amaral and Manuel Delicado Cantero

47
Syntactic Change in French
Sam Wolfe

48
Periphrasis and Inflexion in Diachrony
A View from Romance
Edited by Adam Ledgeway, John Charles Smith, and Nigel Vincent

49
Functional Heads Across Time
Syntactic Reanalysis and Change
Edited by Barbara Egedi and Veronika Hegedűs

50
Alignment and Alignment Change in the Indo-European Family
Edited by Eystein Dahl

51
Germanic Phylogeny
Frederik Hartmann

For a complete list of titles published and in preparation for the series, see pp. 276–80

Germanic Phylogeny

FREDERIK HARTMANN

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Frederik Hartmann 2023

The moral rights of the author have been asserted

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2023930750

ISBN 978–0–19–887273–3

DOI: 10.1093/oso/9780198872733.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Contents

Series preface

Modern diachronic linguistics has important contacts with other subdisciplines, notably first-language acquisition, learnability theory, computational linguistics, sociolinguistics, and the traditional philological study of texts. It is now recognized in the wider field that diachronic linguistics can make a novel contribution to linguistic theory, to historical linguistics, and arguably to cognitive science more widely.

This series provides a forum for work in both diachronic and historical linguistics, including work on change in grammar, sound, and meaning within and across languages; synchronic studies of languages in the past; and descriptive histories of one or more languages. It is intended to reflect and encourage the links between these subjects and fields such as those mentioned above.

The goal of the series is to publish high-quality monographs and collections of papers in diachronic linguistics generally, i.e. studies focussing on change in linguistic structure, and/or change in grammars, which are also intended to make a contribution to linguistic theory, by developing and adopting a current theoretical model, by raising wider questions concerning the nature of language change or by developing theoretical connections with other areas of linguistics and cognitive science as listed above. There is no bias towards a particular language or language family, or towards a particular theoretical framework; work in all theoretical frameworks, and work based on the descriptive tradition of language typology, as well as quantitatively based work using theoretical ideas, also feature in the series.

Adam Ledgeway and Ian Roberts
University of Cambridge

Acknowledgements

A myriad of people have contributed in one way or another to this book through advice and support such that I cannot possibly acknowledge every single contribution in the way it deserves. I thus want to, at least in a few sentences, mention all of you, who helped in making this book happen. To those who I may have inadvertently forgotten in this list, I sincerely apologize.

As this book grew out of my PhD work, it was supported financially by the Studienstiftung des Deutschen Volkes and Deutsche Forschungsgemeinschaft (grant no. 429663384, 'Germanic dispersion beyond trees and waves') who funded me during this period.

I am greatly indebted to my supervisor George Walkden whose support throughout and beyond my PhD cannot be overstated. This book has benefited greatly from his advice and his input, but he has been even more influential on my work in general by fostering my academic personality and by helping me pursue my linguistic enthusiasm in productive ways. I am so grateful that he has been my mentor ever since the first time I walked into his office in June 2018.

Likewise, I want to thank my second PhD supervisor Gerhard Jäger who encouraged and supported my academic endeavours since I was a master's student in Tübingen at the Department of Linguistics. In matters of computational linguistics, it was often ultimately him who pointed me in the right direction whether or not he realized it at the time. He incited me to pursue Bayesian modelling and agent-based models which today make up a large proportion of my methodological toolbox.

Many thanks are due to Karsten Donnay who helped me grasp the topic of simulation models during multiple long discussions early on. He was a strong influence on my views on agent-based models and my first approaches to implement these models in a linguistic setting.

I would like to thank all of those whose expertise helped me with this project. Specifically, I want to thank Patrick Stiles and Johanna Nichols for giving me feedback on individual aspects of Germanic phylogeny. Further, Andy Wedel,

Henri Kauhanen, and Katerina Kalouli have given me helpful advice especially at the early stages of the project which has enabled me to navigate the difficulties of starting a big computational project.

In particular I want to thank Hans-Jörg Karlsen who helped me put my findings in the archaeological context. Without him, the sections on historical influences on the linguistic diversifications in Germanic would not be in this book.

The project benefited from many discussions with colleagues about the ins-and-outs of Germanic cladistics and modelling complex systems, most notably Richard Blythe, Jiaming Luo, Rolf Bremmer, and Nelson Goering. In particular, I want to thank Anne Popkema and Johanneke Sytsema who were mainly responsible for organizing the Old Frisian Summer School in Oxford and Groningen which I had the great fortune of attending in 2019 and to be invited to as a speaker in 2021. In both years, we had great discussions about earlier Germanic languages in general and the role of Old Frisian in the diversification of West Germanic. In this vein, I want to thank the other speakers and also the attendees who truly made the summer school a hotspot for lively debate and a place-to-be for early Germanic linguistics during those weeks.

I am grateful to Miriam Butt and her group at the University of Konstanz who took me in as a fellow computational linguist, enabled me to go to conferences with them, and gave me access to computing power of which I was in dire need.

Special thanks are due to my student assistant Chiara Riegger for her help with checking, organizing, and formatting the data and bibliography. Moreover, this book greatly benefited from our joint work on Burgundian which helped in fleshing out the Burgundian data.

I want to thank my father Jörg Hartmann whose graphic design skills helped me with the implementation of the wave model graphs.

Lastly, I want to thank all the unsung heroes who have supported (and continue to support) me on my academic way, namely my family and my friends who, in more ways than they might realize, made it possible for me to write this book. Most of all, I am incredibly grateful to my lovely wife Sigrid for always believing in me and for giving me strength, support, and distraction when I most needed it.

Preface

The general field of Indo-European cladistics is one of the most well researched fields in the study of genealogical relationships between languages. We have relatively precise information on how individual Indo-European subclades split up from their most recent common ancestor. However, for some Indo-European subfamilies, such as Germanic, we still have open questions as to the nature and the detailed structure of the diversification of this family.

The aim of this book is twofold: firstly, it aims to examine the Germanic language family with computational methods while building on the rich pool of previous research. The goal is, ideally, to be able to tell the most accurate story of the linguistic diversification of Germanic from the break-up of Proto-Germanic to the individual daughter languages. Secondly, this book introduces a novel method for a computational implementation of the wave model that can be used to investigate similar problems concerning wave-like diversification processes in language families.

The reader might find that this book involves a high level of intricate cladistical aspects of Germanic. I attempt to convey the computational aspects in a way that is accessible to all readers, computational and non-computational alike. Although parts of this book can be used as an introduction to phylogenetic algorithms and simulation-based models of language, it is, at its core, a study on Germanic, phylogenetics, and computational wave-model implementations. The structure of this book is such that the first chapters focus on introducing, justifying, and applying the models whereas the chapter on Germanic phylogeny then pools the insights gained from the computational analyses together with previous research to describe the process of Germanic diversification. That is, this chapter seeks to unify all computational and non-computational studies on Germanic phylogeny to paint the most complete picture of this genealogy to date.

This book is the outcome of my PhD work at the University of Konstanz, the topic of which came to me during my work on Vandalic during my graduate studies in Tübingen. Thus, my 2020 book on the Vandalic language is in some ways the spiritual prequel to this book at hand. After I had worked on the Vandalic relationships with other Germanic languages, I felt the need to re-examine the Germanic family more in detail and with methods that have not yet been applied to Germanic.

List of figures

List of tables

1
Introduction

The linguistic history of the Germanic languages is among the best understood areas in the field of historical linguistics. Since the early days of linguistic investigations, generations of researchers have provided the foundation of what we know today. We have reconstructed the Germanic protolanguage *Proto-Germanic* (PGmc) to a degree where gaps in our knowledge are only found on minor or peripheral issues, at least as regards the phonology, morphology, and lexicon of the language.

But despite this detailed coverage, Germanic phylogeny (i.e. the linguistic relatedness within the family) is comparably unclear. The reasons for this is that early Germanic linguistics faces what can be described as a black-box problem. We can reconstruct Proto-Germanic in detail but the earliest extensive textual evidence, except minor text fragments in runic inscriptions, is not attested for at least 800 years (even more than 1,000 for some languages) after the demise of Germanic linguistic unity. This leaves a gap of many hundred years in the records which can only be filled by investigating the later attested languages and reconstructing their possible diversification from a somewhat coherent unity into the individual daughters.

Germanic linguistic research has also yielded insights into the further subgrouping of the family. We have a fairly good understanding that there are North and West Germanic subgroups that are themselves descended from a Northwest Germanic clade (cf. Grønvik 1998: 134–135; Seebold 2013). Unfortunately, this grouping is rather coarse given that the language family comprises at least six well-attested and diverse daughters from which the modern Germanic languages descend. Not only is this subgrouping in itself coarse, but solely assigning languages to these subfamilies does not yield insights into how these subfamilies evolved out of the common ancestor of Proto-Germanic. Those endeavours that aim at shedding light on the questions in detail are often very much debated. Some of the current issues can be listed in the following comprehensive overview.

The earliest definable—now commonly accepted—subgroup is that of Northwest Germanic, yet the language that is excluded from Northwest Germanic, Gothic, is often assigned to a coarsely defined 'East Germanic'.

Germanic Phylogeny. Frederik Hartmann, Oxford University Press. © Frederik Hartmann (2023).
DOI: 10.1093/oso/9780198872733.003.0001

However, the notion of East Germanic as a Germanic subgroup next to Northwest Germanic has been called into question (e.g. Hartmann 2020; Hartmann and Riegger 2022). Moreover, what was the situation that yielded the split between these groups? Some have argued that the split was brought forth by Northwest Germanic undergoing certain subgroup-defining changes that left behind a conservative East Germanic (e.g. Grønvik 1998: 148). Yet it is still unclear whether the data warrant such conclusions as such a notion requires finding clear innovations indicative of a common development among the East Germanic languages which is not the case (see e.g. Hartmann 2020: 115–124). Furthermore, various smaller Germanic languages such as Burgundian and Vandalic have rarely been scrutinized regarding their position in the family and are often assumed to be 'East Germanic' without a clear definition of what this subgroup constitutes.

Having established that Northwest Germanic is its own subgroup, the question arises of why we find so few Northwest Germanic innovations and what this implies about the earliest diversification of Germanic.

Even within Northwest Germanic, especially as pertaining to West Germanic, we find a long-standing debate about whether or not West Germanic constitutes a protolanguage which would in turn suggest the West Germanic languages to either descend from a fairly homogeneous subgroup or from a loosely connected dialect continuum. This issue is connected to two somewhat linked debates about the validity of further subgroups such as Anglo-Frisian or Ingvaeonic (for a comprehensive overview see Stiles 1995). It becomes increasingly clear that we have to see Germanic, and especially West Germanic, as a highly connected area where contact and horizontal transmission of changes frequently occurs and for which it is difficult, if not impossible, to draw clear family trees.

Given these problems, I consider research into the Germanic languages to be in need of a thorough investigation using methods that go beyond, but complement, the traditional methods. Computational research has received much attention in historical linguistics in recent decades as the field of linguistics in its entirety moves towards increased use of quantitative and computational models. The advancement of widely available computational resources and methods calls for a detailed examination of early Germanic linguistics.

The study at hand attempts to be such an investigation. The goal, as will be more precisely defined in later sections, is to apply both computational tree-based phylogenetic and wave-model oriented approaches to the Germanic family to gain novel insights in long-standing debates. In this, I will apply both previously used Bayesian phylogenetic models to the problem and create

a novel algorithm which represents a computational implementation of the wave model by means of agent-based simulations modelling linguistic spread, geographical factors, and diachrony. In some sense, the study itself therefore pursues two aims: modelling Germanic linguistic diversification up to the earliest attested languages *and* presenting and evaluating a novel approach that can be used as a method to model linguistic diversification based on wave-like transmission.

The investigation is therefore chiefly computational, drawing heavily on previous research in historical Germanic linguistics. Without the thorough work of generations of researchers, it would not be possible for the models to build on this knowledge. This, however, entails that the study predominantly discusses linguistic issues and issues pertaining to Germanic phylogeny on a meta-level. In other words, examining the intricate details of certain linguistic changes as arguments for a specific subgrouping is beyond the scope of this endeavour. This is not to say that detailed analysis is irrelevant, but rather that the methods and the viewpoint of the present investigation build on these previous studies rather than re-examining the evidence in detail. The conclusions about Germanic phylogeny therefore stem from a confluence of previous research and novel methods building on this research to obtain a clearer picture of certain issues.

It needs to be stressed that this work seeks not to replace previous research by computational models but attempts to thoroughly investigate the problem at hand using quantitative and computational methods based on traditional research to enrich the picture with powerful tools in order to improve our understanding of these processes.

1.1 A note on the definition of the term *cladistics*

The term 'cladistics' does not, as of yet, have a fixed definition in diachronic linguistics and is sometimes used interchangeably with 'phylogenetics'. For this reason, I henceforth adopt the following definitions: 'Linguistic cladistics' as used in this book describes the linguistic inquiry into language relationships based on the (commonly accepted) assumption that languages descend from one another and linguistic families diversify from a common ancestor.[1] 'Linguistic phylogenetics' is a way of studying cladistical relationships

[1] Note that in this definition, the process of the descent is unspecified, meaning that not just tree-like diversification models can be used in cladistical investigations. Hence, cladistics contains investigations of genetic relationships between languages without assumptions about the shape of the descent process.

by employing methods that model linguistic traits across time to analyse phylogenies, chiefly in the form of evolutionary tree models.

1.2 Summary of cladistical theories concerning Germanic subgroupings

Over the decades, there have been a number of theories regarding potential subgroupings of the Germanic languages. The most prominent of the discussed theories are summarized here to outline the basic proposals and their research history. Note that these proposals are reviewed in detail in sections 5.1 to 5.6.

1.2.1 North Germanic, West Germanic, East Germanic

The first and earliest grouping of the Germanic languages was a tripartite split of the Germanic languages in North, West, and East Germanic languages. This notion can be found in the earliest linguistic research, for example in Krahe (1948); Prokosch (1939); Schleicher (1860); Wrede (1886). This idea was based both on linguistic considerations but also on Roman and Greek historiographic work where, for example, we find the proposal of a common origin of the 'Gothic peoples', among which the East Germanic languages were counted (cf. Braune and Heidermanns 2004: 4). Very early, the tripartite division of Germanic was challenged from multiple angles with researchers proposing two potential further subgroupings of the three languages: Gotho-Nordic and Northwest Germanic.[2] To this day, the tripartite division is still found as the basic assumption of Germanic subgrouping in many books and studies, including introductory works.

1.2.2 Gotho-Nordic

A close relationship between Gothic (or East Germanic in general) and North Germanic first was appealing to many early researchers who based their investigations partly on historiographic work (cf. Grønvik 1998: 70). In his *Getica*, Jordanes uses the foundational myth of Gothic origins in Scandinavia (cf. Miller 2019: 1–2) which, were this to be believed, would warrant closer

[2] There is also the notion that East and West Germanic were more closely related (see e.g. Kortlandt 2001), yet since this theory has never in the past had a strong following, I omitted the proposal at this point. I do not consider it further here.

inspection of Northeast Germanic relations. Further, some supposed linguistic changes common to Gothic and Old Norse brought the Gotho-Nordic hypothesis some adherents (e.g. Schwarz 1951; Krahe and Meid 1969: 37–38), yet it was ultimately abandoned in the common consensus in favour of Northwest Germanic.

1.2.3 Northwest Germanic

Northwest Germanic is the commonly accepted second-order subgrouping of Germanic at least starting with Kuhn (1955), which proposes a closer relationship of the North and West Germanic languages to the exclusion of Gothic. Examples for such changes are, for instance, lowering of earlier *ē to *ā or the loss of several inflectional categories (see Ringe and Taylor 2014: 10–24). Although this theory is accepted in most contemporary research, criticisms of the concept are found in earlier research (chiefly pre-1980) suggesting alternative groupings such as Gotho-Nordic.

1.2.4 Ingvaeonic and Anglo-Frisian

Further subgroupings have been proposed predominantly in the context of West Germanic with *Ingvaeonic* being a subgroup consisting of Old Saxon, Old English, and Old Frisian (e.g. Schwarz 1951), and *Anglo-Frisian* which is proposed by some as a linguistic ancestor to Old English and Old Frisian (for an extensive survey see Nielsen 1981). However, the research history into these subgroupings is intricate as Old Saxon is suggested to be a hybrid language which does not fit perfectly into an Ingvaeonic subgroup (e.g. Nielsen 1989: 79). Moreover, some have cast doubt on whether or not the languages can in fact be regarded as related via their own linguistic ancestors or whether their similarities are due to geographical proximity and membership of a larger dialect continuum (e.g. Stiles 1995, 2013).

1.3 Computational modelling of the Germanic languages

There have been computational studies in the past investigating Germanic phylogeny at least as a by-product of their analyses.

Among the first quantitative attempts to model early Indo-European language relatedness was Ringe, Warnow, and Taylor (2002) who base their

investigation on a lexical dataset. The results of their findings do not contain Germanic interrelationships but cast light on some of the difficulties of placing Germanic in a larger Indo-European family tree.

In the early 2000s, Gray and Atkinson (2003) published a Bayesian phylogenetic study that received much attention in the following years with many researchers heavily criticizing the approach for a variety of reasons (e.g. Chang et al. 2015; Pereltsvaig and Lewis 2015). In this study, they attempt to date the break-up of the Indo-European languages in order to investigate the question of the most likely Indo-European homeland. They eventually estimate an early date for the Indo-European disintegration, 8,700 years before present (Gray and Atkinson 2003: 437), thus concluding the Anatolian homeland theory to be correct. Their analysis also includes the Germanic languages, albeit only in the form of modern variants, inferring a split between North and West Germanic languages 1,750 years before present.

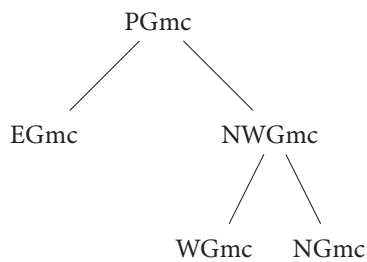
As a Bayesian phylogenetic reevaluation of the Gray–Atkinson model, Chang et al. (2015) presented a study which provided evidence against their claim using a model incorporating fixed ancestral states. In particular, the authors constrained certain extinct languages in their dataset to be treated as ancestor nodes of extant languages (e.g. Latin as an ancestor node of the Romance languages). In the Germanic branch of their model results, they arrive at the traditionally assumed Gothic–Northwest Germanic split with a further division of Northwest Germanic into a branch containing English, Dutch, and German and a lineage comprised of, among others, Norwegian and Swedish.

More recent studies, such as Verkerk (2019), aim at a compromise between strictly tree-like structures common in Bayesian phylogenetics and notions of other forms of diversification, namely horizontal contact and linguistic spread. The ideas to display a language family—in this case Germanic—as a horizontal diversification process ('wave-like') is not unique to computational modelling. Traditional linguistics has debated the *wave theory* since it was proposed by scholars such as Schmidt (1872) and applied to Germanic language relatedness (e.g. Kufner 1972). Today, the notion is that Germanic can be understood as a chain of dialects occupying a defined geographical area (Roberge 2020). This assumption of Germanic as a dialect continuum raises the question whether or not it would be beneficial to further the understanding of this language family by computationally modelling the development of the Germanic languages as a gradual diversification process, starting from a geographically influenced dialect continuum.

1.4 Wave model, tree model, and Germanic phylogeny

The process of linguistic development from a common ancestor language has often been framed as running along two different models of linguistic descent: the tree and wave models.

The origins of the viewpoints of tree and wave models can be traced back to research into Indo-European linguistic relationships in the mid-nineteenth century. One of the researchers, to whom an early version of the tree model idea is attributed, is August Schleicher who was among the first to describe Indo-European cladistics using a family tree (Schleicher 1860). A short time later, Schmidt (1872) put forth the theory of a wave-like diversification of the Indo-European languages, suggesting that languages emerge through overlaying isoglosses rather than through splitting from an earlier ancestor. The diagram below shows the most widely accepted family tree of Germanic with the second-order grouping Northwest Germanic.



Traditional family tree

The wave model itself has never received a commonly accepted definition as the tree model had but most current research encompassing aspects of wave-like relationships define the wave model as a model which uses intersecting isoglosses to define linguistic subgroups. These concepts are closely related to the notions behind *dialect geography* which investigates the geographical distribution of languages, variants, and linguistic features in a given area or for a given linguistic family. For Germanic, Nielsen (1989: 116–133) summarizes the earlier research into dialect-geographical aspects of early Germanic.

The wave theory describes a diversification process in which innovations occur in a linguistic community and spread through the area either encompassing all members (or sub-units) of the speech community or stopping earlier, thus only affecting a subset of members. When repeated multiple times

with multiple innovations, this process yields a linguistic area that is characterized by overlaying innovations. As a result, areas will arise that tend to share more innovations with their nearest neighbours than with communities farther away by virtue of more intensive contact and exchange.

While in earlier research, both models were seen as mutually exclusive, more recent overviews point out that both capture different aspects of linguistic diversification (e.g. Hock 1991: 454).

The emergence of linguistic subgroups through innovation spreading has strong ties to the geographical space they occur in, as shown in recent Labovian sociolinguistic studies (summarized in Labov 2001: 35–73). Although linguistic spread is not (always) congruent with geographical distance, the spread of an innovation permeating through a speech community which eventually dies out is less likely to affect communities at the other end of the dialectal region. Knowledge of the geographical position therefore complements and aids the modelling of the diversification process in question.

Some approaches forego the geographical component and rely solely on linguistic data, such as historical glottometry (as presented in François 2015) (see section 6.2).

While the wave models come closer to how certain languages diversify into subgroups, especially in high-contact and close proximity situations, they are considered to be less easily visualized and harder to summarize with a small number of parameters. Tree models, on the other hand, can be regarded as easier to interpret. This is especially true for the dimensionality of the display. Whereas trees are by definition two dimensional, exhibiting unidirectional branches which can have a certain length and a determinable split time and ordering. Wave-like diversification processes, at least in the most simple definition, operate in three dimensions: two dimensions for the geographical spread of the waves with one temporal dimension for the development of the spread of innovations over time, whereas by definition, wave model diagrams are necessarily two-dimensional (see for discussion Anttila 1989: 300–310). This makes them inherently more complex and less well interpretable. Moreover, from a modelling perspective, a linguistic stemma is more clear and less complicated to devise for a given family, as they mostly only require approximate estimates of similarity and linguistic history of each branch to be collected. Wave-model displays rely on either certain distance measures or measures of group coherence, or they require the researcher to plot a large number of isoglosses on a geographical map. A study that previously used geographical information incorporated in a phylogenetic model is Bouckaert et al. (2012).

With the advent of computational methods in linguistics and large computational resources being readily available, more complex problems that could not be analysed with earlier methods are now in reach.

Germanic is, in some ways, a model case for this issue as the diversification of certain Germanic subgroups is increasingly seen as a diversification of dialect continua in more recent literature (e.g. Seebold 2013; Stiles 2013). Moreover, the family is reasonably well-understood and recent such that we have large datasets and a rich research history which makes it ideally suited to being analysed quantitatively. The present study therefore aims to present a computational wave-model approach that has previously not been applied in cladistics. It is a computational agent-based implementation of the wave theory taking into account temporal and diachronic aspects. In this, it is distinct from previous implementations such as historical glottometry insofar as it operates on computational simulations, statistical principles, and specifically aims at modelling the diversification process rather than displaying single numerical relationships between languages in the form of subgroups.

In short, the approach rests on multiple individual simulations of the Germanic diversification process under the assumption of wave-like innovation spreading. Those simulations that show an isogloss pattern that comes close to the observed linguistic data can then be further analysed to see if there are common patterns of diversification under these best-fit simulations. These simulations take in the factors of time, geography, and linguistic features to approximate the spreading process in order to reconstruct the possible pathways of how the disintegration unfolded.

2
Data

The data for this study were drawn from previous work by Agee (2018) to a great extent. In this study, Agee applies the glottometric framework developed in François (2015) to Germanic. The basis of this glottometric approach is to use a binary innovation dataset to estimate association strength between members of subgroups of a given language family. For his investigation, Agee devised a large database for the Germanic languages containing a large set of innovations from Proto-Germanic to the earliest attested daughter languages. Included in this database are phonological, morphological, syntactic, and lexical innovations. Here, the definition of *innovation* is a change pattern in the structure of a language. According to this definition, a phonological innovation is an identifiable sound change, whereas a syntactic innovation is a change to the syntactic structure of the language. Not that this type of data, with regard to syntactic innovations, is different from the parametric approach taken in some other phylogenetic analyses using syntactic data (such as Longobardi, Guardiano, et al. 2013). Parametric data involve a binary specification of particular structural properties, whereas innovation data involve a binary specification of whether a particular change has occurred or not.

The survey strategy in that study was to extract information of post-PGmc innovations from secondary literature (Agee 2018: 19–20). Although other sources were used, especially for non-NWGmc languages, Agee reports that he relied mostly on the outlines and analyses by Ringe (2017) and Ringe and Taylor (2014). The decision of which innovations were included was based on whether they occur in the 'core vocabulary', a notion the study draws from François (2015) and defined as the top 200 words (Agee 2018: 23). This means that, 'a lexically specific sound change, lexical replacement, lexically specific levelling, etc. is only considered if it affects a word within the core vocabulary' (Agee 2018: 23). This limits the number of innovations to be considered to a certain occurrence frequency. In other words, as the database aims at reflecting the major decisive innovations, Agee decided to omit smaller changes that are mainly word-specific.

For the languages Old Frisian and Old Saxon, the database does not include innovations that are regarded by Agee as exclusively belonging to these

Germanic Phylogeny. Frederik Hartmann, Oxford University Press. © Frederik Hartmann (2023).
DOI: 10.1093/oso/9780198872733.003.0002

languages as the author regards them as too under-researched (Agee 2018: 21–22). Therefore, these innovations were added to the database for the study at hand (see Appendix) by drawing on Bremmer (2009) and Rauch (1992). Moreover, the languages Burgundian and Vandalic are not included in the database and were therefore added to the final dataset. Wherever it is unknown whether a specific innovation is found in Vandalic or Burgundian due to the scarce availability of data, the innovation was given a '?' in the respective field to indicate uncertainty. It is important to stress that this part (Old Frisian and Old Saxon individual innovations and Vandalic and Burgundian in general) is based on original research as these innovations were added to extend the database for the purposes of the study at hand.[1]

To illustrate how the innovations were coded and fed into the database, I extracted four innovations from the dataset to be discussed here (see Table 2.1). The first two entries are present in the original database in Agee (2018) whereas the latter two were added to the database as part of the present investigation.

The first innovation represents the Northwest Germanic lowering of earlier *ē to Proto-Northwest Germanic *ā which can be found in Old Norse, Old English, Old Frisian, Old Saxon, and Old High German (see Ringe and Taylor 2014: 10–13). This is clearly preserved in ON *ráð* and OHG *rāt* < PGmc *rēdaz ('advice'). Fulk (2018: 60–61) notes that this change started in the south of the Germanic-speaking area spreading northward whereas at the same time in the northern parts of the later West Germanic area, onomastic material suggests an un-lowered state. This innovation excludes Gothic and both Vandalic (see Hartmann 2020: 99–107) and Burgundian (see Hartmann and Riegger 2022).

Table 2.1 Sample innovations from the dataset

Innovation	GO	ON	OE	OF	OS	OHG	VAND	BURG
*ē > *ā/ [+stress]	0	1	1	1	1	1	0	0
dual > emptyset	0	1	1	1	1	1	?	?
*ō > *ū/ [+stress]	0	0	0	0	0	0	1	0
*e, *i > i	1	0	0	0	0	0	0	1

[1] After further inspection, it was found that most innovations of Old Saxon that are considered uncontroversial are already in the original database as they are either parallel or common innovations with other languages. However, it was confirmed in preliminary model runs that adding dummy innovations does not change the model estimates in meaningful ways. This means that should independent innovations need to be added to Old Saxon due to different coding decisions or future research, they would not significantly change the model results presented here.

Gothic, for example, shows a continuation of this phoneme in the words Goth. *slepan* < PGmc *slēpaną ('to sleep') and Goth. *ufblesan* < PGmc *blēsaną ('to blow') (Ringe and Taylor 2014: 11). In Vandalic, we find that PGmc *ē is preserved in its quality as observed in words such as Vand. *rēþ *rēða- < PGmc *rēdaz ('advice') (cf. Goth *garēdaba*, OHG *rāt*) or Vand. *mērV- < PGmc *mērijaz ('famous') (cf. OHG *māri*) (see Hartmann 2020: 132–133). It has to be noted that although it is difficult to ascertain vowel quantity in Vandalic due to the attestation situation, the continuation of the vowel quality is apparent. In Burgundian, the preservation of earlier *ē is likewise present: thus we find Burg. *rēða- < PGmc *rēdaz ('advice') attested in, for example, *Leubaredus* and Burg. *mērja- < PGmc *mērijaz ('famous') attested in, for example, *Sigismerem*. For this reason, the innovations for all three languages need to be coded as **0**, i.e. the innovation *ē > *ā is not present in these languages.

The Proto-Germanic dual in verbs was inherited from Proto-Indo-European (see Ringe 2017: 260) and subsequently lost in the daughters Old English, Old High German, Old Norse, Old Frisian, and Old Saxon (Ringe and Taylor 2014: 20–21). In these languages, the number category was replaced by the plural. Thus we find that in Gothic, the Proto-Germanic dual forms are retained while Old High German lost them in favour of the plural. Table 2.2 shows the contrast between Gothic and Old High German in the paradigm of the verb 'to take' (see also Fulk 2018: 273).

Both in Burgundian and in Vandalic, the scarcity of the data does not permit determining whether the dual was present or absent in both languages (Hartmann 2020; Hartmann and Riegger 2022). Thus, Gothic is coded as **0** in this dataset, since it retains the dual forms; both Vandalic and Burgundian receive **?**, since presence or absence of dual cannot be determined and all other languages that exhibit no dual forms are coded as **1**.

The Vandalic raising of stressed *ō > *ū can be determined to be a Late Vandalic change (see Hartmann 2020: 105–106). It can be observed in the forms Vand. *blūma < PGmc *blōmô 'flower' (cf. Goth *blōma*, OE *blōma*) and Vand. *mūþ *mūða- < PGmc *mōdaz 'anger, mind' (cf. Goth *mōþs*, OE

Table 2.2 Sample innovations from the dataset

	PGmc	Gothic	Old High German	
Singular	*nemō	*nima*	*nimu*	'I take'
Dual	*nemōz	*nimōs*	–	'We (both) take'
Plural	*nemamaz	*nimam*	*nemamēs*	'We take'