The Invisible Hand and Sound Change

By J. Sulik (SLKJUS001)

Submitted in part fulfillment of the requirements for the degree of MA (Linguistics).
Supervisor: Prof Roger Lass

I wish to thank Prof Roger Lass for his supervision of this dissertation, and Prof Nigel Love for his comments on earlier drafts. I am also grateful to Milena Sulik for her proofreading of the final draft.

Declaration

1) This is my own work.
2) The contents of this dissertation have not been submitted in full or in part for any other degree.
3) All substantial contributions to or quotations in this dissertation have been cited and referenced.

Signed by candidate

J. Sulik

The Invisible Hand and Sound Change

Dissertation submitted in part fulfilment of the requirements for the degree of Master of Arts, University of Cape Town, December 2003.

I hereby declare that this dissertation is my own work.

Justin William Bernard Sulik
SLKJUS001

Acknowledgements

I wish to thank professor Roger Lass for his supervision of this thesis, for sharing his incredible knowledge of Germanic history and historical linguistics, and for his ability to get me to say what I mean. Thanks, also, to professor Nigel Love for his comments on this thesis, and for various courses he has taught me over the last few years, which first stretched my attentions beyond Generativism.

Contents

Abstract

Rudi Keller's publication of *On Language Change: the invisible hand in language* was very well received. His book presents a suitable model for explaining language change. It argues that language is fundamentally different from the objects of study of the physical sciences, and that it therefore requires a form of explanation distinct from those found in other fields. Nevertheless, its influence has not been as significant as it deserves. A factor contributing to this state of affairs is its inability to deal with sound change. Lass (1997: 363ff.) discusses how Keller's book is thus a partial answer to an important question. In that sound change has always been among the most common concerns of the average historical linguist, it seems that Keller's argument would be significantly strengthened if sound change were explicable by his model.

This thesis suggests a possible solution to the problem, by showing that Germanic *i*-umlaut can be interpreted as a consequence of the invisible hand, and that it is therefore an example of change as envisaged by Keller. Other examples of change presented here include the history of the Greek sigmatic future, snowballing (Aitchison, 1987), and tonogenesis.

The discussion is based on Eco's semiotics, and it highlights the distinction between (linguistic) information and the intersubjective communication of that information. In focusing on the intersubjectivity of communication, reference must be made to structuration theory (from Deumert, 2003), and to the meta-theory of Lass (1997). The approach here is contrasted with other common forms of explanation, such as the hermeneutic mode, and functionalism.

0) Introduction

Weinreich, Labov & Herzog (1968) represented a new descriptive or methodological departure for historical linguistics, based on the Labovian variation studies of the 60's. Such work represented the first serious attempts to account for language change while characterising language as a heterogeneous structure, i.e. to allow structured or orderly variation over time. Most work up until that point, from the Neo-grammarians to Saussure to Chomsky, had assumed that language can only coherently be dealt with as a homogeneous structure, as is argued throughout Wienreich *et al.* (1968). Such an assumption underlies Saussure's distinction between synchronous *langue* on one hand, and diachrony and *parole* on the other, or Chomsky's use of an idealised speaker and community to investigate competence as opposed to performance.

Weinreich *et al.* suggest that, if we equate structure with homogeneity, our 'theory of language change becomes removed from empirical foundations almost entirely' (Weinreich *et al.*, 1968: 129). If structure and homogeneity are equated, then it follows that either change is not structured, or it is homogeneous. On the basis of empirical study, the authors conclude that neither option is capable of providing a detailed and realistic description of change itself.

These claims, however, can be characterised by another important set of oppositions: that between macro- and microscopic levels. The latter is concerned with speakers, their actions, their psychological states and their environments, while the macroscopic level is speaker-independent (cf. the discussion throughout Deumert, 2003). Many pre-Saussureans, for example, focused largely on the macro-scopic (i.e. on a proto-language), as did Saussure himself (i.e. on *langue*). Weinreich *et al.* (1968), on the other hand, argue that reference needs to be made to both levels in explaining change.

Though the Labovian model does make reference to a micro-level, such a framework has 'shown little interest in formulating a micro-theory based on the actions and behaviours of individuals to back up their collective accounts (i.e. to show how the collective patterns arise as the result of a multitude of individual actions)' (Deumert, 2003: 49). Before language change can be fully explained, then, this descriptive gap had to be filled.

This challenge was taken up by Keller (1994). Basing his discussion on the work of various philosophers of the Scottish Enlightenment (among others), he shows how language cannot be subsumed under either natural or artificial entities, though it might share characteristics with both. In particular, it is characterised as a phenomenon of the third kind, which include those 'establishments, which are indeed the results of human action, but not the execution of human design' (Ferguson, 1767; quoted in Keller, 1994: 57). As an entity of this so-called third type, language will obviously need a mode of explanation different from that used for the physical sciences, and this mode is referred to as an explanation from the invisible hand. More particularly, what is being explained here is how micro-variation ends up being macro-structure – the connection with Deumert's criticisms of Labov should thus be clear.

Keller's model in its essentials can have relatively little to say about the initial source of variation in any categorical sense. The model principally shows how this variation might become macro-structure over time. Though Keller generally does posit maxims (causes of action) as the source of variation, his model will be shown to be compatible with a less agentive account, presenting a framework within which a variety of possible sources of variation can be examined.[1] Keller's model is thus not incompatible with approaches from sociolinguistics or language processing, and it is remarkably anti-hegemonic compared with other theories of explanation. For example, it is not as universalist as generativist or typical semiotic accounts sometimes are (Keller, 1994: 154).

In that this discussion thus incorporates (or has the potential to incorporate) several extra-linguistic considerations, and in that it recognises the unique nature of entities like language as deserving their own particular form of explanation, and in that it begins to allow us some way of deriving macro-structure from micro-variation, this model has obvious importance for modern historical linguistics. Nyman (1994: 231) claims that 'few books have as much to contribute as Keller's to our understanding of the very basic issues of human language and its change'.

Though various criticisms of the theory are possible (cf. Deumert, 2003: § 13), one which Keller himself has acknowledged is that it seems incapable of dealing with sound change (cf. Adamska-Sałaciak, 1991: 175; Deumert, 2003: 62). Given that sound change has been one of the major concerns of historical linguistics, it seems

that it would be useful, if possible, to plug that hole. Lass (1997: 363) highlights the importance of connecting the micro- and macro-levels in modern historical linguistics, but stresses that Keller's model has only partial success here, due to this gap. My thesis thus seeks to suggest a way of incorporating certain sound changes into Keller's model. If successful, then, the result will be a model of language change (or more correctly, a model for explaining language change) that achieves explanatory adequacy across a wider range of phenomena, no longer arbitrarily excluding sound change. Other benefits of doing this will be outlined throughout the final chapters.

I will assume a fairly neutral theory of phonology, which incorporates elements of autosegmentalism and particle phonology, without making reference to something as specific as government phonology. This is, after all, an essay on historical linguistics and not phonology. The lack of a formalist credo here is in part to due to a warning that (in general) generative linguistics incorrectly 'equates the move from informal to formal notation with the move from description to explanation' (Berg, 1998: 2). Since I am engaged in finding a suitable explanation, steering clear of excessive appeal to formalism may help ensure that I avoid this pitfall. In general, I will create my own formalisms as illustration of various points, and as needed, and not as an appeal to mental processes.

Chapter one outlines the development and essential features of explanations from the invisible hand, and in doing so gives a clearer idea of just what aspect of language can be considered a phenomenon of the third type. Chapter two suggests why it has been difficult to include sound change in this model, and suggests in what direction we might look for a solution. Chapter three discusses in great detail my central example: Germanic *i*-umlaut. It shows why a change such as this is different from a change like lenition, and why we need to invoke all the extra baggage of the third kind in order to explain it, giving its historical context.

Chapter four outlines the notion of language habits, while chapter five provides examples of such habits. Chapter six synthesises the contents of chapters three through five into an invisible-hand explanation of sound change.

---

[1] An overview of various such possibilities for the source of variation is presented in Berg (1998: ch. 2).

1) Tertium Datur

1.1 The development of theories of phenomena of the third kind

For much of Western intellectual history, a binary distinction has been made among the things that are: we find on one hand, natural things, physical things, things that are independent of human action; on the other, artificial things, historical things, things that are the product of human design. This distinction is at least as old as the fifth century BC:

> The two terms *nomos* (pl. *nomoi*) and *physis* are key-words ... of Greek thought. In earlier writers they do not necessarily appear incompatible or antithetical, but in the intellectual climate of the fifth century they came to be commonly regarded as opposed and mutually exclusive. (Guthrie, 1978 vol. III: 55)

As early as Plato, however, we see dissatisfaction with this dichotomy. In *The Cratylus*, which examines the naturalness of names, we see Plato 'set out opposing theories[2], to show that neither is wholly right and conclude only that the matter needs more thought' (Guthrie, 1978 vol. IV: 16). The nature of names or signs, then, remained unresolved until the Scottish school of philosophy applied its methods to their analysis. In response to what is known as Mandeville's paradox (Keller, 1994: § 2.2), a new description of certain phenomena was outlined. The paradox, basically put, is that in a community of bees, each bee driven by his own avarice, the hive as a whole should prosper. The set of consequences for a community (or the macro-level), then, are not predictable from the goals of the individuals within that community (or the causes of the micro-level). Further, the final consequence, the prosperity of the hive, is not the goal of any individual member, and it is therefore not a concern motivating their action. The objects of such a macro-level are nevertheless constituted out of the actions of the individuals comprising it, hence Ferguson's 'establishments, which are indeed the results of human action, but not the execution of human design' (Ferguson, 1767; quoted in Keller, 1994: 57). Language, like the economy, is one of

[2] Either that names are natural or that they are conventional.

these macro-objects. Language exists only as far as people use it; nevertheless, people do not have control over the development of their language.[3]

First, however, a suitable example (from Keller, 1994: 63-64) will demonstrate the unique nature of phenomena of the third kind. If driver A were to spill his coffee and thus hit the brakes, we can hypothesise about his goals: possibly he got a shock and reacted out of instinct (i.e. had no conscious goal); possibly he wished to slow down to give himself time to decide what to do about the scalding coffee. Driver B, directly behind him, would see him braking, and would brake too. In fact, operating on the maxim 'rather brake too much than too little to avoid a crash', he will end up driving even slower than driver A. The same goes for driver C, driving slower than B, etc. until the traffic eventually comes to a complete standstill. Causing a traffic jam was the goal of no individual in the entire process. Yet it was created as a consequence of the collectivity of individuals acting on their own goals.

A few points can thus be made about such phenomena. A central feature is that we will expect them to have two levels of causation. There is the intermediate level of results, which coincide with the goals of individual actions *if* the action is successful and goal-driven (just as braking is the goal of each driver in the above example, with the result that each driver becomes slower than the preceding one). These results *en masse* form a set of consequences (the unintended consequence of a lot of people braking is that there will be a traffic jam). The apparent teleology of the consequences, then, merely stems from a large enough group of people behaving in the same way, or reacting in a similar way to their environment. Environment will thus turn out to be an important notion throughout my discussion.

1.2 Language as a phenomenon of the third kind

Keller argues that language is a phenomenon of the third kind (Keller, 1994: § 4.1). While this is informative, it may be inaccurate. Language consists of a highly complex network of sub-systems: 'it is a neurological, physical, sociological,

[3] This model for describing a distinct class of phenomena has been applied by Adam Smith to economics, and to language by Carl Menger (Keller, 1994: 66). It was Keller, however, who first applied it to language change, and examined what would constitute a suitable method of explaining such a phenomenon.

psychological reality among others' (Berg, 1998: 6). A similar point is made in Lass (1997: 353), and Keller himself states quite clearly that

> [i]t is not my intention to make a bid for theoretical hegemony. I do not deny that language is (also) a system of symbols, a code, an object of Popper's world 3, Humboldt's *energeia*, or that it can be meaningfully studied as a Chomskian I-grammar. (Keller, 1994: 154)

It is by no means given that if one such system is of the third kind, then the others will also be. Though Keller admits this, he nevertheless maintains that language change is only possible as a phenomenon of the third kind. However, Adamska-Sałaciak (1991: 177) criticises this as an unwarranted assumption. Phenomena of the third kind must be created by interpersonal interaction. The value of their units is thus inherently social.[4] The mainstream in linguistics, however, argues that much of language is biological.[5] I think there is room on the podium for both views, but it would be best to decide which systems belong in which camp.

Chomsky's approach to language has been criticised by those linguists of a more philosophical bent, particularly for its unfalsifiability (Berg, 1998: 5). Such criticism aside, I think that for the purposes of a thesis such as this, I can take it as assumed that the core of the generativist hypothesis still holds, or at least that language 'can be meaningfully studied as a Chomskian I-grammar' (Keller, 1994: 154), basic syntactic structures being prime examples here. In trying to form a question by reversing the order of the words of a statement, then, I am not breaking social, but neurological or processing rules. Evidence generally quoted to support these kinds of claims are the lack of such processes in the languages of the world, suggesting that they are a product of our make-up as humans, rather than products of our societies.

I do something very different when I break the rules governing the use of words or sounds, though. To spell things out very clearly to a hearer, I can use the word [junɪvɜ:sɪtɪ] to mean 'university'. A great many people do not speak this precisely, however, and a more likely form would be [junɵvɜ:ʂtɪ] or even [juɳvɜ:ʂtɪ]. An

[4] Though not necessarily in the Saussurean sense, as I shall outline towards the end of this chapter.
[5] Or pseudo-biological. Berg suggests that the generativist use of biology is at best half-hearted (Berg,1998: 5). I mean "biological" here in whatever sense the word is implied by Chomsky's description of I-grammar.

extreme but entirely possible form (at least, in South African English) might be [jəɱvə:stɪ̯], and I would probably still be understood if I were to use this form in context. There thus exists a wide range of possible pronunciations for a given word, which may include more or less phonemic information, depending on the context.[6]

It may be worth stressing that I am not claiming that there is a wide range of possibilities for differing phonemic information here; rather there is a certain linguistic type which contains a maximal amount of information, and which may be instantiated by more or less accurate tokens. In this, I will have to distinguish two kinds of variation. There is Labov's structured variation, variation that "means" something in a social context. There is also the kind of variation seen here, that does not mean anything in a comparable way. I will assume that the former is logically prior to the latter in production, i.e. that a speaker will (unconsciously) pick a structured variant of a word before this variant can be pronounced with more or less accuracy.

The above example involves sound, but similar examples can illustrate semantics. I can say 'please pass me the spatula', or simply 'pass me that'. In a certain context, I might even wave my hand agitatedly in a vague direction and say 'thingy', and still be understood. This raises two important issues. Firstly, it seems we are dealing with guidelines more than rules in the use of these sounds or meanings (again, I am discussing their use, not their formal representation); secondly, the rules or guidelines are imposed by the interaction of speaker and hearer and by the context of this interaction, and not necessarily by their respective biological make-ups. All of this needs a bit of discussion, though.

It is widely agreed that communication involves a toss-up between getting a message across clearly (which takes a lot of energy but is more likely to be successful) and getting it done easily (which saves energy, but is also less likely to be successful). Such a suggestion has been made variously by Keller and by Berg (and a host of others). In using more energy, the speaker has a certain goal in mind: effective communication. There exists, though, a whole host of background reasons as to why he wishes to communicate effectively. Keller posits the maxim 'talk in such a way that you are socially successful, at the lowest possible cost' (Keller, 1994: 107).

[6] The discussion in this paragraph is based on Keller (1994: 108-109).

Communication is obviously central to the whole business of speaking, though this is a truism. It nevertheless has bearing on the subject at hand.

There is no biological imperative driving me to use a certain set of sounds to mean a certain thing, action or state of affairs. However, if I wish to communicate with someone, I will be far more successful if I use some method that he or she knows. Indicating something by Crusoe-like gestures might be possible, but it is doubtless not very effective. In our experience, people use certain sounds to mean certain things. If we imitate their behaviour, we figure that they will understand that we mean certain things if we use the same or similar sounds[7]. There is a certain maximal energy I can expend on following this tradition; alternatively, I can use less energy, and be less precise in my sounds or meaning, but nevertheless profit by it for various reasons.

This pairing of sound and meaning is the essential element of a sign (ignoring for the moment the differences between Saussure's, Peirce's and Eco's versions of the sign). A sign is a tradition (though it may be other things as well), and if we wish to communicate effectively with a member of a community, we have little choice but to use the same tradition. A sign is thus not subject to our control. Borges has stated this more imaginatively in saying that 'words are symbols that assume a shared memory' (Borges, 1979: 33), where this shared memory is our mutual experience of communicating within a speech community, and not subject to conscious control.

Various arguments, however, have been made to suggest that it is at best an imprecise tradition. Taylor (1995) discusses the possibility of treating the sign as a fuzzy category centred on a prototype. The central meaning or sound of the sign is firmly established, but its borders remain undefined. Tokens can be more central or more peripheral members of such a sign-set. This is compatible with Eco's version of the sign, though, which is based on experience. Eco (1984) discusses a model for (part of) the sign that he calls "Model-Q". This is a web of previously heard, related meanings, structured rather like the internet (see fig. 1). As each individual's experiences differ, their signs will differ, though often minutely. A pre-industrial speaker of English, for example, will not have the sense of 'plant' as 'factory' that a post-industrial speaker might.

[7] Though I am by no means suggesting that language boils down to such imitation. Notice, however, the importance of desiring to communicate as a condition here.

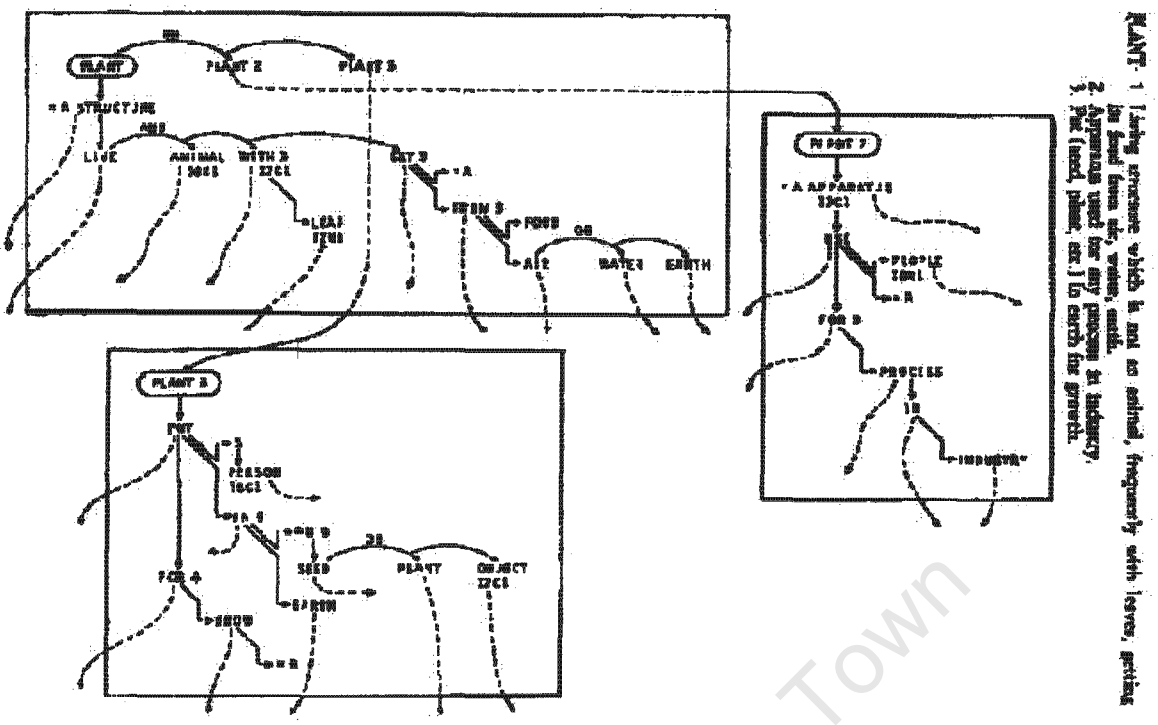


Figure 1. (Eco, 1976: 123)

Nevertheless, we can assume that most people's experiences of communicating within a community are comparable. In that communication is embedded within a social network, and in that each person in that network will potentially have experience of communicating with each other person, the type remains fairly uniform across the network. This is to say, speakers A and C might never talk to each other. If they both regularly speak to B, however, (assuming that this is a monolingual community) they will have to speak in a similar fashion, by virtue of their indirect connection through B, or through the social network. Though many speakers in a community will never speak to each other, then, there will nevertheless be a large deal of coherence among their Model-Qs.

It is thus possible to grant the type a degree of (virtual) social reality. Given that new speakers will not be able to deviate far from the established type if they wish to communicate, the fact of communication will ensure that this virtual social unit is comparable with all the individual memories of which it is made up (within a given community). I will thus use "type" to refer indiscriminately to both the speaker's mental representation and the social extrapolation of this representation.

Both these models, Taylor's and Eco's, allow a certain amount of free range for speakers and hearers, and they are in their essentials compatible. Hence my use of "guidelines" above, indicating that not all speakers slavishly reproduce exactly the same tokens of a sign. From the above discussion, the third-kind nature of signs should be apparent. Signs are created by human interaction and allow for individual variation. Nevertheless, an individual cannot invent signs at will, being forced by his desire to communicate effectively to use signs that his experience has taught him are normal.

It is thus more accurate to claim that the sign (and not language) is a phenomenon of the third kind. It is in this sense that we best interpret Keller's statement:

> What do I follow when I construct an English sentence correctly or shrink back from eating dog meat, or if I like to wear trousers[8] rather than skirts, or if I prefer to sit on a chair rather than on the floor when I eat – my reason or my instinct? Neither! I follow traditions which have emerged in my country; I follow social rules. (Keller, 1994: 40)

It would read better if it were 'use an English sign' rather than 'construct an English sentence'. Change in a sign will thus be a phenomenon of the third kind, but there exist types of language change that do not make reference to the sign. Naturally, these will not be relevant to the discussion here.

It is important, however, to establish how people behave with regard to phenomena of the third kind. Keller argues that

> [t]he dichotomy 'nature versus art' is related to a parallel but no less misleading dichotomy: 'instinct versus reason' or 'emotion versus reason'. Just as one distinguishes on the level of objects between artefacts and natural phenomena, one distinguishes on the level of behaviour between behaviour guided by reason and behaviour guided by instinct or emotion. Thus the arguably most fantastic and certainly most decisive of

[8] This is not to say that English is a mere matter of fashion. It seems that there is no clearly definable boundaries in such cases, though. Clearly a man wearing a skirt "means" something that a man wearing trousers doesn't. A man wearing tight trousers, though, doesn't "mean" quite as much, in comparison with a man wearing non-tight trousers. The former example involves clear social *mores*; the latter might, too, but to a significantly smaller degree. The notion of ground will be discussed as being of use here. Here, though, the sense of "meaning" is not directly comparable with linguistic meaning. Cf. Lass' criticism's of Eco's universalist semiotics, for the meaning of tight jeans (Lass, 1997: 340).

human abilities goes by the wayside: the ability to establish customs or traditions and to behave according to rules. (Keller, 1994: 40)

Having established that the sign is a phenomenon of the third kind, we can thus suppose that people behave in a corresponding fashion when using signs. Habitual behaviour (the kind Keller refers to in the establishing of customs) shares features with both instinct and reason, but is nevertheless a distinct kind. I do not reason with myself every time I use a particular sign; nor are my signs hardwired into me. Rather, I am compelled to follow custom in using signs, because that is what people around me have always done, and because my experience suggests that I will be able to communicate effectively if I use the same signs. Alternatively, it is necessary to follow such habits in order to be speaking a certain language at all. It is only English social habit that associates /dɒg/ and 'dog', and if I use the same sounds to mean 'elephant' or 'blunderbuss', I am no longer speaking English in any meaningful sense of the word "English". Such a notion is thus anti-universalist: the habits of English speakers are constructed out of their experience of English of a particular period and place, and not necessarily out of their experiences as human beings.

The above is not intended to claim that language is no more than a system of signs, though, and I fully admit the possibility that even signs are constituted of at least some non-social aspects. A sign consists of expression and content, and it seems possible to argue that linguistic expression is constrained by phonological processing rules. We can thus make a distinction between those elements of the sign that are purely linguistic (the phonological structure of the prototype or lexical form) and those that are partially social (the sign itself).

1.3 Explanations of phenomena of the third kind

Phenomena of the third kind are not amenable to the same form of explanation as natural objects. As these structures are created out of the chance results of free will in a social context, we must consider them significantly different from the effects of the

blind and exceptionless operation of physical laws above the quantum level.[9] Physical explanations are typically of the D-N type, and they are used to make falsifiable predictions; neither criterion applies to phenomena of the third kind.

If we are not entirely concerned with prediction, then what is the use of explaining an historical phenomenon such as language change? What does such explanation add to description? From different viewpoints, both Eco and Keller answer that such non-D-N explanations serve us by diagnosis (Keller, 1994: 74; Eco, 1984: § 0.7). That is to say, instead of collecting a set of valid premises and deducing a conclusion from them, we observe phenomena, treat them as conclusions, and abduce backwards to hypothesis the state of affairs that may have caused them. Explanations of such a kind, then, are not so much "true" as "useful and coherent". The central distinction between physical explanations and those of phenomena of the third kind is in the former's use of deductive logic[10], as opposed to the latter's use of abduction. In explaining sound change abductively, then, we may gain insight into the nature of language, the nature of the discipline that studies it, and into historical contexts surrounding the relevant events.[11]

Keller gives the following as elements of an explanation from the invisible hand:

> 1 the depiction of the motives, intentions, goals, convictions (and such like) on which the actions of the individuals who participate in the generation of the phenomenon in question are based, including the general conditions of their actions;
> 2 the depiction of the process that explains the generation of structure by the multitude of individual actions;
> 3 the depiction of the structure generated by these actions. (Keller, 1994: 70)

> *An invisible-hand explanation explains its explanandum, a phenomenon of the third kind, as the causal consequence of individual intentional actions which are based on at least partially similar intentions.* (Keller, 1994: 71; emphasis Keller's)

---

[9] Obviously, though, there are physical systems so complex that "laws" are not possible, as in meteorology. The above thus refers to systems *such* as Newtonian *physics*, *while* ack*now*ledg*ing* the limitations of this description.

[10] Though settling on an hypothesis initially involves abduction, the final picture will be deductive.

[11] I think, though, that the distinction between abductive diagnosis and deductive prediction is not as clear as either camp suggests (Eco (1976) and Keller (1994) speaking in favour of abduction, and Berg (1998) in favour of deduction). It makes more sense to suggest that in inquiring into something like sound change, we start with some small idea of cause and of effect, and repeatedly work between them, abductively and deductively, clarifying both with each stage.

Put simplistically, (1) describes the factors determining the micro-processes, (3) describes the macro-process, and (2) describes how to get from the former to the latter. Point (1) is central to the diagnosis: we take (3) as given and abduce back to its causes. In phenomena of this kind, these (indirect) causes are the micro-processes of (1). It is partially to discover these influences on the actions of individuals that we undertake an explanation of this nature.

Deumert has criticised Keller's assumptions about a theory of action on these grounds, however. Much of Keller's discussion focuses on the goals of individuals, and how these drive their actions. The phrasing of point (1) above, for example, is suggestive of conscious agency on the part of the individual, and the italicised quote speaks only of intentions as the deciding factor in action. The sense of "action" found in his examples is thus very similar to the following:

> Our ability to perform apparently random, motiveless acts is undeniable – but only if they are done with some purpose or other in mind do they count as *acts* in the first place. Someone who kicks the cat while having an epileptic fit is not acting freely: their kicks were not actions at all. (Margaret Boden, *The Creative Mind*, quoted in Lass, 1997: 325)

Deumert (2003) responds, however, that much of human action is not determined by such intentions. She discusses how, instead, this intention is often a *post hoc* illusion, making reference to Giddens' structuration theory. I think, though, that this is more a criticism of Keller's presentation of such explanations, than of the nature of these explanations themselves. Keller's model handles the consequences of actions, and this does not necessitate any particular account of action theory. Point (1) above ends with "the general conditions of their actions". These general conditions can be stated in a way as to appease Deumert's criticisms: we are not only interested in intentional or conscious goals which drive actions. Under (1), we can thus consider instinctual responses, rational goals, or socially habituated behaviour.

Keller himself specifically states that he considers the most significant mode of action in language change to be the habitual mode (as discussed in section 1.3). This is obviously at odds with rational intention, so at most we can claim that the phrasing of his examples does not conform to his discussion elsewhere. We can thus appeal to "maxims" in a broad sense that can include instinct and habitual behaviour, in

2) Invisible Hand Explanations and Sound Change

2.1 Problems

I think two observations can be made about the invisible hand and about the explanation of sound change, as initial suggestions why no one (as far as I know) has applied the former to the latter. The first regards differences between sound and meaning; the second regards differences between various kinds of sound change.

The invisible hand has been successfully used to explain meaning, and it thus worthwhile examining pertinent differences between sound and meaning. It seems that meaning implies a certain duality that sound does not. That is, it is possible to link some meaningful linguistic units with things in the real world. For example, one can use polite words and thereby act politely: there is a distinction between something that is polite, and something that means 'polite'. Politeness is something independent of language, and though its ontology is obviously not the same as that of physical objects, it is in some sense real. It is thus possible to make a claim such as "I used that word in order to be polite" without committing a tautology. For the importance of this, consider Keller's example of the pejoration of German *Weib* 'woman'. He claims that

> [i]n a society like ours with a courtly tradition, there exists a rule of gallantry towards women... Part and parcel of this gallant behaviour is a tendency, when talking with or about women, to choose expressions that tend to belong to a higher level of style or social standing than a lower one. The maxim is therefore ... 'in case of doubt choose an expression that is a notch too high rather than too low'. In the course of time, the 'next higher' word becomes the normal expression, whereas the formerly normal one is pejorised. (Keller, 1994: 77)

Thus, previously neutral expressions such as *Weib* or *Frau* become pejorised in favour of words like *Dame*, not because people wish to use *Weib* pejoratively, but because they wish to use *Dame* normally. As discussed above, Keller's examples tend to favour intentional maxims, and it is the duality made possible by meaning that allows such intentions to be posited. In this case, for example, a maxim relating polite

meaning and polite behaviour in the use of *Dame* has the causal consequence of *Weib* becoming pejorised.

Since sound lacks such duality, it is difficult to conceive of corresponding maxims for sound. Though it may be difficult, however, it is not impossible. Two options are available here, then. Either we can maintain that sound does not relate to anything language-external, or we discuss what kind of a language-external thing would be suitable. It is by no means the case that only one of these options is possible, though, and this leads into the second point introduced above: that there are different kinds of sound change.

The first option, treating sound as purely linguistic, yields simple maxims for action, but a correspondingly dubious teleology. As discussed above, the sign has some purely linguistic aspects (the phonemic structure of the expression), but as a link between sound and meaning, it is not purely linguistic itself. Since it is the sign itself that is the phenomenon of the third kind, it seems reasonable to suppose that maxims explaining sound change as a phenomenon of the third kind should make reference to the sign. Maxims which relate only to purely linguistic sound, then, are not likely to be satisfactory in dealing with a phenomenon of the third kind.

Consider, for example, the case of lenition. A suitable maxim might be 'speak in such a way that you conserve energy'.[12] I concede that lenition is obviously more complex than this implies, and that this on its own does not explain lenition (failing, as it does, to say anything about the problem of actuation), but however one explains lenition, it cannot be as a phenomenon of the third kind. Firstly, phenomena of the third kind have two levels of causation (distinguishing results and consequences). Lenition, however, is simply causal: this maxim leads directly to a state of lenition. Secondly, phenomena of the third kind stem are rooted in a collectivity. The sign, for example, is a matter of interpersonal custom. Lenition, however, is not interpersonal. It is a matter of expression only (concerning only the speaker and not the hearer), and not of the link between expression and meaning that constitutes the sign.

It thus seems we should follow the second option, and find something extra-linguistic to make sound comparable to meaning. This something would be (roughly) to linguistic sound what social politeness is to a grammatical honorific. Identifying what such a thing might be is the course I will pursue throughout the later part of this

chapter. In searching for an external link, though, I will not yet specify whether this link is meaningful or accidental. That is, I will not yet distinguish between the functionalist and the historical approaches. The former involve maxims of the sort given in Keller's example of *Weib*; the latter on the other hand will be committed to the habitual mode of action outlined in Keller's discussions.

2.2 Initial considerations

Though my argument for sound change of the third kind will be presented in various chapters to come, I can outline at this stage what kind of a sound change should be amenable to description and explanation by Keller's model, and what entities we may have to invoke to do so.

To begin the discussion, let us examine another of Keller's examples, that of the beaten footpath (Keller, 1994: 70-73). Take two buildings that face onto a university lawn, one from the south, the other from the east. Their front doors face onto the lawn and these front doors are linked by concrete paths running in cardinal directions, which thus meet perpendicularly. However, it is likely that many people will walk the shortest path between the doors (i.e. diagonally across the lawn). Given a high amount of traffic, a beaten footpath will emerge. It was the intention of no individual to create such a path, but many people following their own maxims behaved in a similar fashion, and a footpath was created. This is clearly a phenomenon of the third kind. Keller explains it as follows:

> I hypothesise that most people resemble each other in preferring to take a shorter rather than a longer route. I observe that the paved paths do not represent the shortest connection between those points most often frequented by people at a university. I know that a lawn withers in those places that are most often trodden on. I therefore assume that the system of footpaths is the non-intended causal consequence of all those (intentional[13], finalistic) actions which consist in reaching certain goals by foot under the maxim of saving energy. (Keller, 1994: 71)

---

[12] This makes no commitment to the mode of the maxim: it may be habitual, rational or instinctual, or some combination of the three

[13] As already stated, I think that just as good an explanation is possible even if we do not treat these maxims as intentional. That is, the decision to walk the shortest distance is unlikely to be as rational or as conscious as Keller suggests, as Deumert (2003: § 13) points out.

In the explanation of the beaten path, certain maxims have a result: people prefer to walk along the shortest route. This result, then, is a matter of individual psychology and action, and has little to do with the lawn. It is merely because the lawn is the arena where these results are played out, that a path is made. Because the results are about how to walk, and because the lawn is the place where the walking is done, the consequence - the beaten path - follows. To revert to Mandeville's example, we see a similar pattern. Each bee is driven by greed to acquire as much as possible for himself. Again, the results (the bees' actions) are a matter of individual psychology, and have nothing to do with the hive. However, because these actions are played out within the hive, the consequence follows: the hive as a whole prospers. We can distinguish here between the maxims, the bees' avarice; the results, the bees' acquisitive actions; the arena, the hive; and the consequence, the hive's prosperity. In highlighting the arena in this manner, I am not departing significantly from Keller's model; merely making explicit a notion that lies tacit there.

Explanations of this kind must thus make reference to various partially independent but historically contiguous things: walking and a lawn; braking and a highway; bees and a hive. There is no necessity in linking the act of walking with any given lawn, or with lawns in general. The accidental contiguity of actors and arena is what causes the results to lead to the consequence, and the consequence makes reference to both individuals and arena.[14] The initial suggestion, then, is that functionalism (regardless of its status elsewhere) will be unsuitable in explaining a phenomenon of the third kind, for it assumes a non-accidental link between two phenomena.

The explicative importance of this claim can be seen in a comparison with Berg's claims about Popper:

> 'The new theory [the explanation] should proceed from some *simple, new, and powerful unifying idea* about some connection or relation … between hitherto unconnected things … or facts … or new "theoretical entities"' (Popper 1963; his emphasis). Setting aside methodological issues for a moment, explaining means establishing a 'connection between hitherto unconnected things' … The locus of the 'unconnection' is in the

[14] This might clarify the claims about lenition being inexplicable by the invisible hand: lenition does not make reference to two such independent phenomena.

internal world, i.e. the mind of the researcher … The 'things' in the external world must be of an intermediate level of connectivity. On one hand, there must be some kind of connection between them, otherwise the scientist would hunt for a chimera. On the other hand, the 'things' must not be so strongly connected that they lose their independence … (Berg, 1998: 11)

It is not immediately obvious, however, at which levels language is relevant for sound change: relevant maxims or conditions, the level of individual action, the arena where the action is played out, or the consequence. Recall, for example, Keller's explanation of *Weib*. Here, the individuals' maxims involve politeness, and are not necessarily linguistic. Because matters of politeness are played out in language, however (because people wish to be polite when they speak), the maxims have a consequence for language: the word *Weib* becomes pejorative. Here maxims of action yield consequences for meaning, polite action and meaning not being in a cause-and-effect relationship. Language, in this example, then, is relevant as a result, as the arena and as the consequence but not as a maxim. It may be taken for granted, though, that language will always be relevant at the level of consequence, when we are trying to explain language as a phenomenon of the third kind.

The state of affairs just outlined is not logically necessary, however. Labov's description of /r/ in New York provides an alternative set-up. The maxims in this case are largely linguistic; people may or may not wish to imitate another's speech sounds. If a certain variant is seen as positive or negative, the maxim will govern the use of that linguistic variant.[15] The arena where these results are acted out, however, is largely social. This is to say, the variation is structured by social contexts. The consequence, the fate of *r* in this dialect, is naturally linguistic. Keller's example of change in meaning thus involves social maxims being played out in language; Labov's involves linguistic maxims being played out in social interaction.

Though Labov's setup, and not Keller's, seems to be the best point of departure for discussing sound change within the framework of the invisible hand (since it can deal with phonological variation, while Keller's examples are largely semantic), this is not in fact the case. Various criteria are important in making this judgement.

[15] It seems to be the case that some maxims lead to other maxims (my desire to seem socially acceptable leads to a desire to sound like someone else). I will generally only deal with those maxims most closely related to the relevant action itself. In this case, the relevant maxims govern the use of linguistic units.

Primary among these is that the arrangement in Labov's example, though involving sound, is not amenable to explanation of the third kind. If we were to identify the maxims at work in Labov's example, we would find the following. While the specific desires might range from the wish to appear of a socially higher class to the desire to disassociate oneself from metropolitan invaders[16], we can posit two general maxims: either to use a linguistic variable that someone else does (i.e. to talk like someone else), or not to use such a linguistic variable (i.e. not to talk like someone else).

The result of the first of these, for example, is that we do indeed then talk like someone else. The consequence of this result (i.e. of many people following this maxim) is that the variation spreads. There is a direct or causal link, however, between maxims and consequences in this case. Though the two are not logically equivalent, the maxim promotes conformity and the conclusion is a state of conformity. Phenomena of the third kind, however, are not the consequences of maxims themselves. The spreading of this variation, then, is not a phenomenon of the third kind. This point is similar to one made about lenition above: in both cases the consequence and the maxims concern the same phenomenon. Many sound changes, then, need not appeal to language's nature as a phenomenon of the third kind.

Notice that in both Keller's and Labov's examples, the results of individual actions are similar: either a linguistic unit is used (in which case it spreads/survives) or not used (in which case it does not, or becomes pejorative). The conclusions are also linguistic in both cases, so the nature of neither the results nor of the conclusions provides the significant differences between Keller's and Labov's models. What remains to be established, then, is the nature of maxims and of the arena the actions are played out in. It is by departing from both Keller's and Labov's examples of maxims and arenas that we will be able to explain sound change as a phenomenon of the third kind. The trick, then, lies in finding maxims that concern sound, but not so directly that they yield consequences causally related to those maxims.

It is unsurprising that these two features, maxims and the arena, are jointly significant. The arena will determine the ecological conditions affecting individuals, and hence will affect the individuals' responses to these conditions. In walking across a lawn, for example, a sign saying "keep off the lawn" (a factor of the arena) may

[16] As is claimed for Martha's Vineyard (Mesthrie *et al.*, 1999: 80-84).

affect my behaviour. The nature of the arena may thus play a part in determining the kind of ecological conditions that will be relevant.

2.3 Possibilities for independent levels

Within a typical modern linguistic framework, the initial assumption might be that the phonemic and phonetic levels are the two different phenomena we need here. The phonemic underlying representation, as a mental and abstract phenomenon, differs somewhat from the phonetic output, which is physical rather than mental (Berg, 1998: 11). This suggestion is unsatisfactory for various reasons, though. First among these is the causal link between the emic and etic levels; second is the ambiguity with which the etic level is used.[17]

In the example of the beaten path, it is historically accidental that the walking was done on the lawn. Walking has to happen some place, but it does not necessarily have to be on the lawn. This contingency has been highlighted as an essential feature of explanations of the third kind, and the semi-independence of external things has been related to Popper's claims. There is no comparable contingency in the relationship between the phonemic and phonetic levels of sound. Assuming a neutral generativist outlook, one could argue that rules applying to the phonemic level lead causally to the phonetic output. This *predictable* causal link is incompatible with features of phenomena of the third kind.

Nevertheless, it would be acting too hastily to abandon this distinction entirely. To be able to retain it, I need to clarify the sense in which I mean "phonetic": there is, I think, a degree of ambiguity here. The string of phones generally labelled "phonetic", that string shown in square brackets, is sometimes treated as the output to the phonological processes of a language. For example, if we assume that aspiration in English is not part of the underlying representation (being predictable), but is added by a phonological rule, we produce a surface form such as [p^hʌpi] for *puppy*. It is this sense of phonetic that I have rejected in the preceding paragraph.

This output of the phonological process, however, is merely the prototype for the sign, as discussed below. That is, this prototype may be pronounced with more or less

accuracy, as in the case of *university*. These varying pronunciations are also transcribed in square brackets, and also called phonetic,. The word "phonetic" is thus used both for the idealised output of rules (the type), and for the various possible instances of the word (the tokens). This type/token distinction cannot be ignored, however. The former are purely linguistic, while the latter involve non-linguistic considerations, in addition to the purely linguistic. Recall that the use of a token is at least partially shaped by the interaction of speaker and hearer and their context, for example. It is not the case that purely linguistic laws lead to the form [jʊŋvə:stɪ̯].

The situating of a phonetic type within the range of possible tokens will be shown to be the expression of a sign. Having established above that it is the sign that is, strictly speaking, a phenomenon of the third kind, and not language as a whole, it is clear that we ought to investigate the sign as a possible solution to the problems outlined in the previous section. We thus turn to semiotics for further exploration. In what follows, I will make use of Eco's semiotics, rather than Saussure's or Peirce's. I have argued in another essay[18] how Eco's model represents a compromise between those of Saussure and Peirce, providing a sign that is both more realistic, and more useful to linguists. In accepting Eco's theory of the sign, however, I am not committing myself to his general application of semiosis.[19] Nor in advocating a semiotic investigation am I suggesting that the typical semiotic 'one meaning: one form' claim (Lass, 1997: 342; cf. Berg, 1998: § 2.4) will be of use to me.

Eco divides semiotics into two main areas. Firstly, we have a semiotics of signification, which studies the make-up and nature of signs; secondly, a semiotics of communication which studies how signs are used in context (Eco, 1976: 4, *inter alia*). The latter is said to consist in a theory of codes, the former in a theory of sign production. Both will be relevant to my purposes here, though I will refer to the former immediately in §2.4, and to the latter substantially only in §2.5. Both these presentations will have to be fairly rough sketches.

[17] Lass (2003, personal communication) suggests a third reason: the mental is itself ultimately physical in a non-dualist account of mind.
[18] 'Semiotics and Linguistics', written for professor Nigel Love's course "Landmarks in Modern Linguistic Thought", 2003.
[19] Cf. the criticism of Eco's semioticisation of tight jeans in Lass (1997: 340n).

Independently, both Peirce and Saussure can be said to have founded the modern study of signs.[20] Saussure's sign is a dyad, is social, immutable, and exists independently of its use. Peirce's sign is a triad, is logical, mutable, and has no existence outside of its use. These two perspectives of the sign could hardly be more different, but Eco has effected a compromise between the binary and the triad, and has salvaged the most useful features from both theories.

Peirce's triad is based on inference (Eco, 1976: § 2.7). A mind[21] perceives something (a representamen). On the basis of previous experience (a ground), a link (an interpretant) is created in the mind between this representamen and a further something (an object).[22] As an example, a hunter perceives scratches on a tree (these scratches being the representamen). In his experience, such scratches are made by passing deer (this feature of his experience being the ground). A link is thus created in his mind between the scratches and an object, 'deer' (this object is a further sign, and not a physical deer). The scratches, as a result, signify 'deer' to the hunter. The interpretant or link between representamen and object is based on inference motivated by experience, and since the representamen, object and interpretant make up the sign, we can claim that the sign is essentially inferential. However, this allows any act of inference to constitute a sign. Semiotics, then, ends up being a theory of almost everything that humans come into contact with.

Eco limits this unwieldy generality by claiming that a sign must have a socially established ground.[23] He states

> I propose to define as a sign *everything* that, on the grounds of a previously established social convention, can be taken as *something standing for something else*. (Eco, 1976: 16; emphasis Eco's)

[20] I will refer to this as "semiotics" throughout, ignoring the distinction between "semiotics" and "semiology".
[21] To be more strictly peircean, it would be better to say "quasi-mind". Since I am concerned with language only, this will not be necessary, for I imagine that language presupposes a mind.
[22] It would be pointless to define the object with any clarity here, given the complexity of the matter. We can simply take the object as a further sign. Signs, then, are interpreted in terms of other signs.
[23] It is partially because of this that I claimed that Eco's semiotics is more useful to linguists.

> The first doctor who discovered a sort of constant relationship between an array of red spots on the patient's face and a given disease (measles) made an inference: but insofar as this relationship has been made conventional and has been registered as such in medical treatises a *semiotic convention* has been established. (Eco, 1976: 17; emphasis Eco's)

Eco thus effects a compromise between fixed social value and free inference. In that the ground is constructed out of an individual's experience of society, the sign has a social aspect. But since the ground is only one part of the sign, the whole is not given, and inference is still needed. In line with my comments above, it is my experience of other people that using certain sounds has a certain result. On this basis, desiring to effect my own results, I use the same (or similar) sounds. The potential link between sound and meaning is thus built up out of my experience of society; nevertheless, no one reproduces the sounds exactly, and thus a degree of inference is needed on the part of my hearer to associate tokens with particular types.

This ground, then, an association of content and expression, is represented by Eco's Model-Q. Model-Q, however, is not the sign. It consists of a network of possible meanings for possible expressions. In hearing an actual expression, then, an inference is made to its actual meaning, based on previous experience of such sounds and meanings (i.e. on the ground). Showing tokens with a prime stroke, and types without (where the ground is treated as an implication), we thus have the following schema. Here, fuzzy abduction has replaced the deductive *modus ponens*.

$$\begin{array}{l} e \rightarrow c \\ \underline{e'} \\ c' \end{array}$$

Without the primes, this would be deductive. The abduction creeps in when we associate tokens and types. Inference may be needed to varying degrees. Where the token closely resembles the prototype, inference will be needed to a lesser degree than for a more marginal exemplar of the type. Varying amounts of inference, for example, will be needed for the various forms of 'university' transcribed above. The effort expended in producing a sound thus has a direct but inverse counterpart in the effort needed to interpret it.

Though Eco does not state so directly, I think it is possible to conceive of a network of possible expressions, in addition to the standard assumption of a network of possible meanings. I think we can claim that just as a range of previously heard meanings of an expression are stored in Model-Q, so too are a range of previously heard pronunciations. Structured variation (such as competing rhotic and non-rhotic varieties being accessible to the same person in production or interpretation) is perhaps describable in these terms.

In that Model-Q is based on previous experience, it allows feedback in language. This is to say, previous instances of language are stored in memory, and inform present usages. These present usages in turn obviously become part of the same store, though. Model-Q thus provides a framework for treating past stages of a language as inputs to present stages, in tying new memory into older structure. Part of the sign therefore exists *in potentia*, for we have access to our grounds through memory.

However, the sign itself, the association between a particular sound and particular meaning, exists only in the moment of its use. In assigning one value to another, the sign acts as a function. Indeed, Eco (1976: § 2.1) claims that a more proper name would be "sign-function", linking the two functives, sound and meaning in the case of language, or expression and content more generally. In that the sign-function translates elements of one system (meaning) into elements of another (sound), or vice versa, it is a code. Here Eco (1976: § 1.2) distinguishes between a true code, and an s-code.

Both types of code describe relationships between units. In an s-code, these units form part of the same system; in a true code, they are taken from different systems, as above with sound and meaning. Structure in the Saussurean sense, then, is an s-code. The linking of sound and meaning during communication, however, is a true code and is not purely linguistic.

To find an invisible hand explanation, we need to distinguish at least two kinds of phenomena (walking and the lawn; bees' collecting pollen and the hive). If we remain within one system, as in an s-code, such a distinction is impossible. In this light, we can clarify a previous distinction between meaning and lenition. Keller's example of *Weib* involves a link between linguistic politeness and social politeness, where the link is a true code. Lenition, on the other hand, makes reference only to the physical production of sound, and is thus not an instance of a true code. Reference to the sign,

allowing two interacting systems, thus seems useful for treating sound change as a phenomenon of the third kind.

Sound is not yet on as stable a footing as meaning, however. This is to say, there is a duality in meaning which allows us sensibly to discuss maxims and conclusions that both relate to meaning, without them involving the same phenomena. For example, a maxim can concern polite action, which only indirectly has consequences for polite speech. I have not yet presented any such duality for sound, though. The duality of meaning falls directly out of the nature of meaning: we sometimes use words to refer to things. It would be preferable if my discussion of sound rested on as firm a foundation. Fortunately, Eco's conception of semiotics provides such a possibility.

2.5 Language and communication

Recall that semiotics also contains a theory of sign production, entailed by the semiotics of communication. A semiotics of communication studies, in part, how information is transferred to its interpreter: 'there is ... a communication process when the possibilities provided by a signification system are exploited in order to produce expressions for many practical purposes' (Eco, 1976: 4). Information here is not merely linguistic. Nor is communication synonymous with language, though language presupposes communication. Bruner's studies show how children learn to communicate before they learn language (Taylor, 2001: 74), and clear instances of communication without language include bee dances, human body language, and two computers transferring information. It is may thus be sensible to distinguish between the transfer of information, and the information itself.

Eco (1976: ch 1) gives the following example of such a distinction, though this discussion is fairly standard (cf. also Keller, 1994: 108-109). He describes how an engineer needs to know whether the level of water in a watershed upstream is too high or not, i.e. whether it has passed what he has specified as a danger level. A simple code for transferring this information would be for the watershed to send a +A signal when the level has been passed; otherwise it sends a -A signal. However, this system is open to problems from

the presence of potential *noise* on the channel, which is to say any disturbance that could alter the nature of the signals, making them difficult to detect, or producing +A when –A is intended or vice versa. Therefore the engineer has to complicate his code. (Eco, 1976: 34; emphasis Eco's)

Eco then describes how adding a B signal in addition to the A signal reduces the change of interference: +A+B can be taken as an indication that the danger level has been reached, while at least some of the other three possibilities can be left devoid of content, reducing the scope of error. Adding a C and a D signal reduces this chance even further, and allows the possibility of further messages, not merely whether the water level has passed the danger point. Through all of this, the information remains the same (whether the water has passed a certain level). The manner of the information's communication, through a set of features A-D, is not the same.

The interpreter of the message uses a code to link the expression with its content, but this expression can be more complex and hence clearer, or less complex and hence less clear. The redundancy used in the message results from the nature of communication, not necessarily from the nature of the information to be communicated. Varying methods of communication will convey different aspects of the information with varying degrees of accuracy, and there is thus no necessary logical link between any message and a particular means of getting it across. Compare the following words, for example, taken from different stages in the history of English: **mu:siz*, *my:si* and *my:s* all mean 'mice'. They all thus contain the same information – 'plural' – but this information is communicated differently in each: in the first instance in the ending *-iz*, in the second in the combination of vowel rounding and the ending *–i*, and in the third only in vowel rounding. At the moment, it may be unclear what I mean by "information" here, as opposed to "meaning", but this should become clearer in chapters 4 and 5.

No linguistic rule governs the distinction between tokens such as [jəɱvə:stɪ̯] and [junɪvɜ:sɪtɪ]. Since both tokens can be successful, we can thus speak of "success" rather than "correctness" in communication. If a speaker wishes to get a message across to a hearer, he will select a suitable word, that is a pairing of meaning and phonetic type. The sound type will be conveyed by a more or less accurate token. The correctness of the type depends only on the speaker's intentions and on linguistic

rules. However, the success of the use of the token depends on the interaction between speaker and hearer.

Communication and language, then, are two semi-independent phenomena, but are conceptually contiguous, as is needed for explanations of the third kind. One is a set of a certain kind of information, the other is the arena where that information is transferred. They are judged by different criteria, and follow from different maxims. The two sets of maxims are only indirectly related by the sign's association of type and token, and of content and expression. It is therefore not stretching the metaphor too far to compare the lawn as the place where walking happens with communication as the place where speech happens.

Various conclusions follow from the above discussion. Firstly, only those sound changes that need to make overt reference to the communication of certain information should be accounted for by an explanation of the third kind. Lenition, remember, involves a maxim of production, but not of transfer: it involves only the speaker, not the interaction of speaker and hearer. Secondly, just as a "don't walk on the lawn" sign can present ecological factors affecting action, the nature of communication will have consequences for specific actions of speaking. In chapter five, I will thus examine some communicational maxims. Thirdly, since we cannot discover specific historical acts of communication (with a view to explaining sound change), we will need to present the historical context of the sound change with some detail. This will be the focus of chapter three.

3) A Brief History of *I*-umlaut

I-umlaut will be described in this section, first historically, then phonologically. A clear description is essential before explanation can take place. In the following sections, the historical context of *i*-umlaut will be presented: the phonological context in 3.2 and the semantic context in 3.3.

3.1 An introduction to i*-umlaut*

Descriptively, umlaut is not very different from the vowel harmony seen in Finnish: they both involve the spread of a feature from one vowel to another, forwards in Finnish and backwards in Germanic. Nevertheless, umlaut can be distinguished from Finnish vowel harmony in how we explain it. Finnish vowel harmony is simply a fact of the language: just as PIE had ablaut, so does Finnish have vowel harmony. There is thus little sense in wanting to *explain* Finnish vowel harmony in any system-external sense. Germanic umlaut, on the other hand, was a process more limited than Finnish vowel harmony.

It was implemented in a specific and short-lived historical period, while Finnish harmony continues as a synchronic process. Further, umlaut only spreads a feature from very specific endings, while Finnish harmony is less morphologically constrained. We can thus examine the historical and morphological contexts of umlaut to provide an explanation that has some system-external features, in suggesting why it happened at a certain time and not another, or why it involves a specific set of endings. This will provide a fuller explanation than merely saying that Germanic underwent a period of vowel harmony, or that it anticipated certain features.

Both PIE and PGmc lacked front rounded vowels. At some stage after the split up of the various early Germanic languages, they all (with the exception of Gothic) developed front rounded vowels by the same process. In general, between the sixth and seventh centuries, a high front vocalic unit (i.e. either an *i* or a *j*) in a suffix caused a preceding back vowel to front, or a front vowel to raise (Lass, 1994: 61-2). Since the non-low back vowels were rounded (as is quite common), this process led to the creation of front rounded vowels. For a while, the front variety was merely a allophonic alternation. In time, however, the environment was lost, leaving the

fronted or raised vowel with a morphophonological role. The following brief outline from Old English shows these various stages (Campbell, 1998: 217):

Proto-Germanic	**mu:s*	**mu:s-iz*
Early Pre-English	*mu:s*	*mu:s-i*
Umlaut	-	*my:s-i*
Loss of *–i*	-	*my:s*

Gothic was not subject to this process simply because what we mean by "Gothic" is a highly specific literary language standardised in the fourth century (Wright, 1924: 2). This language was thus too early to have shown *i*-umlaut. More can be said about why it can be considered too early, but a fuller discussion of this will have to wait until a clearer description of *i*-umlaut is possible. In brief, though, *i*-umlaut can be considered in relation to another Germanic rule, mora loss. Gothic only experienced one application of this rule, while the other languages experienced a second application shortly after umlaut. Because the second application of mora loss occurred a few centuries after the first, and long after Gothic had been standardised, I will discuss how mora loss is at least a partial cause for *i*-umlaut between its first and second iterations. In that case, we might be a bit more certain that Gothic did not merely fail to show umlaut in its orthography. When I speak of umlaut as a Germanic process, then, I mean as a process of a certain era.

I-umlaut is generally characterised in two ways: as a raising process for front vowels, and as a fronting process for back vowels. Treating this as a rule with features, we still need to define two processes: for front vowels, the [+high] or [-low] of the *i* causes raising; for back vowels, the [-back] of the *i* causes fronting (Lass, 1994: 60).

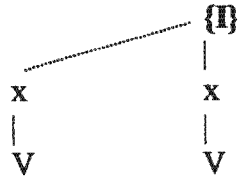
If we treat this as a rule with particles[24], however, only one process is needed. I take it as theoretically axiomatic that a simpler description, though managing to take into account all the facts, is preferable to a more complex description, and I thus prefer to use particles instead of features in this case.[25] The addition of an {I} particle to a low front vowel will raise it, and the addition of the same to a back vowel will

[24] "Particles" is used by Schane (1984, quoted in Ewen & van der Hulst, 2001: 103); "elements" by Harris (1997: 342). I will use the former throughout, discriminating only when necessary.
[25] This is insofar as particles are useful in this instance – I make no claim about the superiority of particles over features in general.

front it.[26] In the following, I will show particles in bold face, and assume that the three particles **{I}**, **{A}**, **{U}** combine to give different vowels in the following way. This table is based on Schane (1984, quoted in Ewen & van der Hulst, 2001: 103).[27]

particle	vowel	particle	vowel	particle	vowel
I	i	**I,U**	y	**U**	u
I,A	e	**I,U,A**	ø	**U,A**	o
I,A,A	ɛ	**I,U,A,A**	œ	**U,A,A**	ɔ
				A	ɑ

Under such an interpretation, *i*-umlaut is merely the spreading of an **{I}** particle in an ending leftwards to an adjacent position on the vocalic tier (Ewen & van der Hulst, 2001: § 2.4.1):



Although the use of particles is simpler than the use of features, it introduces a further useful point. In classical terms, features are an undifferentiated block. /i/ and /ɛ/ are thus, respectively

$$\begin{bmatrix} +\text{son} \\ -\text{cons} \\ +\text{voice} \\ -\text{back} \\ +\text{high} \\ -\text{low} \\ +\text{ATR} \\ -\text{round} \end{bmatrix} \text{ and } \begin{bmatrix} +\text{son} \\ -\text{cons} \\ +\text{voice} \\ -\text{back} \\ -\text{high} \\ -\text{low} \\ -\text{ATR} \\ -\text{round} \end{bmatrix}$$

[26] On this basis, I exclude the change */e/→/i/ here. Although it fits the general pattern of umlaut, it happened centuries earlier, and is not describable in terms of the addition of an **{I}** particle.

[27] Ewen and van der Hulst suggest different options for representation, however. Other accounts include the notion of dependency or headedness, so that the distinction between [e] and [ɛ] is that the former includes I in a head position: **{I, A}**, the latter in a non-head position: **{I,A}**, where the underlining shows headedness (Harris, 1997: 360). My account below, in its essentials, is compatible with either option. My table above is a hybrid of Schane and Harris, in that it treats /ɑ/ as **{A}** following the latter, and not **{I,A,A,A}** following the former.

Both these vowels contain the same amount of information under this interpretation. In particle theory, on the other hand, /i/ is merely **{I}**, while /e/ is **{I,A}** and /ɛ/ is **{I,A,A}** (Schane, 1984, quoted in Ewen & van der Hulst, 2001: 103). Here, clearly /ɛ/ contains more information than /e/, which in turn contains more information than /i/. The same holds true for the elemental view from government phonology, though in the distinction between /e/ and /ɛ/ there is the headedness of **{I}** in the former (Harris, 1997: 360). Though of a different type, this headedness is nevertheless additional information. Though I make no claims about the comparative values of Harris and Schane in any objective sense, I will generally prefer the latter's account in that it treats /e/ as being phonemically simpler than /ɛ/, for the addition of headedness achieves the reverse in Harris.[28]

A particle approach thus brings us naturally onto the notion of information itself. As discussed in the previous chapter, a sound change that needs to make reference to information in its own right is exactly what we are looking for, and a particle account of umlaut will allow us this option (though it remains to be seen why this notion of information is relevant here). Naturally, some accounts of features using redundancy or underspecification achieve a similar result (as does feature geometry), treating some sounds as having more information than others. It is still the case, however, that particles provide a simpler description of *i*-umlaut than even these reduced features.

3.2 Phonological loss[29]

In its barest essentials, the history of the phonology of Germanic endings is one of loss. Many consonants were lost in a word-final position, and we find great reduction in the range of possible vowels. A particular process has been identified in inflectional vowels, however: mora loss. This section presents a description of mora loss that departs in various respects from the traditional view (which I take to be either

[28] It may be possible, however, to interpret headedness as a default, in which case /ɛ/ is less simple in that it departs from the default (i.e. a lack of headedness is information that needs to be specified, while presence of headedness is otherwise assumed), in which case both accounts make similar claims. In this case, I would still use Schane's particles for their notational simplicity.

[29] The contents of this section are summarised from a more detailed essay, "Gothic thematic vowels and mora loss", written for a course in Germanic philology under professor Roger Lass, 2003.

Prokosch, 1939: § 49; or Voyles, 1992: § 3.1.12). It clarifies the question of how we should describe the rule, and also how we should define a mora in this case.

I think that whatever is being lost in the process described below is not exactly equivalent to the usual sense of "mora" elsewhere in phonology. Calling the rule "mora loss" is then somewhat of a misnomer. It should be understood that any reference I make to a mora here is merely a reference to that thing which is lost in the rule commonly called "mora loss", and not to morae as generally understood. It may thus be best to use scare quotes every time I say "mora", but I do so only sporadically, so as not to soak the next few paragraphs in punctuation.

My central example here is taken from Gothic. This may seem odd, in that I will eventually relate mora loss and umlaut, a process that does not seem to have happened in Gothic. However, since Gothic only experienced one application of mora loss while other languages experienced more than two, its relative simplicity in this regard makes it a clearer example. Its endings are more clearly comparable with PIE endings in as short a space as I have available here. More specifically, I am limiting myself to nominal paradigms. This is not theoretically problematic, since these form a system distinct from verbs, and I can thus cover a semi-autonomous set of data properly in as short a section as this.

A mora is usually treated as a unit of vowel or syllable length (Crystal, 2003: 299-300). In PIE, we find vowels with one, two, or three morae. I will thus confine myself to "a unit of vowel length" (and not syllable length), since vowels in open syllables contain as wide a range of morae as I will need to make reference to. This is to say, we often find both a bi-moraic and a tri-moraic ending without a consonant in the rhyme, so the distinction in such cases rests in the vowels alone. For example, cf. the Gk. *κρίνω* 'I judge' with *κρινῶ* 'I will judge', where the latter has three morae in the omega, and the former has two (at least, under the interpretation of Prokosch, 1939: § 49). It is not the case, however, that we find a corresponding three-way distinction in vowel length, which we might expect if a mora were a unit of vowel length. Indeed, a three-way distinction in vowel length is typologically unusual.

PIE's three-way distinction in morae carried on into Greek: there we find vowels that are short (1 mora), long (2 morae) and circumflex (3 morae). On this basis, linguists suppose that a tri-moraic vowel in PIE had a circumflex accent genetically comparable to that in Greek (though the PIE accent is one of pitch, a type gradually lost in later Greek). Already this statement is problematic, since "short" and "long"

are used to describe length, while "circumflex" is the name of an accent: length and accent are independent concepts, and there is no reason to suppose that if *x* is "long" then *x*+1 is "circumflex". Circumflex vowels had the same length as non-circumflex long vowels. Given the distinction between a two-way length contrast and a three-way mora contrast, we can only conclude that a mora is not merely a unit of vowel length (or that, whatever it is, it is only indirectly related to length).

For the moment, *faute de mieux*, I will suppose that a mora is instead a unit of vowel information, though this supposition will be supported by the Germanic data. I will also assume that this information is best represented as a particle. This is largely because it makes the following account simpler, and because I have used particles in my description of umlaut. At this point I narrow my focus from morae in general to those in Germanic endings, remembering that the sense of "mora" here is merely that which is lost in the rule described in this section. I propose that such a "mora" is a slot that can bear one particle. A tri-moraic or circumflex vowel, then, can carry up to three particles, a mono-moraic vowel only one.[30]

On this basis we can make a fairly strong prediction for Gothic. Short vowels (vowels with one mora) in inflectional endings in Gothic should be limited to /i/, /a/ and /u/, these being the realisation of **{I}**, **{A}** and **{U}** respectively. Similarly, in long vowels (having two morae), we would expect the combinations **{I,A}** = /e:/ and **{U,A}** = /o:/. **{I,U}** = /y:/ is logically possible, but is excluded here simply because Gothic lacked front rounded vowels, and because this gap is nothing to do with mora loss. Since PIE had at most three morae in a vowel in an ending, we would expect at most two in Gothic after the application of mora loss, to be described below. We thus would not expect the vowels /ɛ/ = **{I,A,A}** or /ɔ/ = **{U,A,A}** in an ending.

All of this is in fact exactly what we find to be the case. In this last claim, I am following Wright's rather than Voyles' interpretation of the dative singular of Gmc *ō*-stems: that it is the diphthong /ai/ rather than the vowel /ɛ:/ (Voyles, 1992: 230; Wright, 1924: 90). This confusion arises out of the sometimes parsimonious orthographic system used by Gothic. I am not merely making this decision to avoid a

[30] This interpretation, however, is not incompatible in its essentials with a more standard interpretation of morae. In Greek, if a circumflex vowel has HLH on the tonal tier, while an acute vowel has HH or HL, it is still the case that the circumflex contains more information than the acute. A fuller discussion might thus be able to compare Germanic and Greek on the basis that the morae in the latter are units of information on the tonal tiers, in the former on the melodic tier.

hole in my argument, however: I will discuss why Voyles' account is otherwise unsatisfactory.

Traditional statements of mora loss tend either to be able to capture a significant generalisation, or to provide historical detail, but not both. Compare the following two statements of the rule, which are typical in this regard. Prokosch states

> During the first two or three centuries A.D., i.e., soon after the Germanic accent shift, final syllables lost one mora. About five hundred years later a second mora was lost; another five hundred years later, a third. Naturally, this is an over-mechanical statement and a good deal of allowance must be made for deviations. (Prokosch, 1939: 133)

Voyles, on the other hand, lists the following rules (Voyles, 1992: § 3.1.12). These are abbreviations of his rules, since they otherwise demonstrate a complexity unnecessary for my purposes here.

Deletion of Unstressed Short Vowels

$$\begin{bmatrix} V \\ \text{- nasal (NWG only)} \\ \text{- long} \\ \text{- stressed} \end{bmatrix} \rightarrow \emptyset$$

Shortening of Unstressed Long Vowels

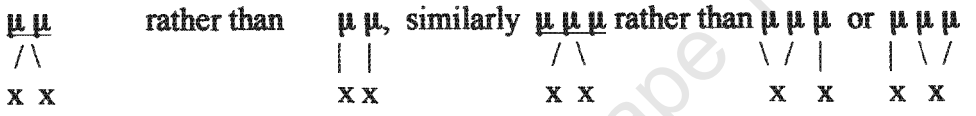
$$\begin{bmatrix} V \\ \text{- nasal (NWG only)} \\ \text{+ long} \\ \text{- stressed} \end{bmatrix} \rightarrow [\text{- long}] \,/\, \underline{\quad}\#\#$$

The change from a circumflex to a non-circumflex vowel is not shown as part of mora loss. Voyles' account thus treats the changes from 3 to 2 morae, from 2 to 1 and from 1 to 0 as different processes. Prokosch treats them as the same process, but does not provide any detail as to how this might be the case. I take it that a description of the rule that achieves Prokosch's level of generality and (something near) Voyles' level of detail is theoretically better than either of these on their own. If Voyles treats a mora as a unit of length, he cannot generalise across the various changes subsumed under mora loss. Only 1 to 0 and 2 to 1 involve shortening; a change from 3 to 2

morae involves a change from a circumflex to a non-circumflex vowel of the same length.

We can generalise across these processes, however, if we treat morae as units of information. Further, under this interpretation, various changes previously considered exceptional are made regular, and it thus has various independent claims to recommend it. Under such an interpretation, however, we would not expect a low-mid vowel in a bi-moraic position. It can be argued, then, that Wright's assumption of a diphthong in the dat. sing. of *ō*-stems is better than Voyles' suggestion of a low-mid vowel, on the basis that an informatic sense of mora makes more sense out of mora loss than a definition based on vowel length.

The following assumes that an intervening "moraic" tier exists between the skeletal tier and the melodic tier (Crystal, 2003: 299-300). Morae are linked as a group to skeletal positions, a group being all the morae associated with one vowel:



Since the morae indicate the amount of information present in the vowel, and since the vowel itself represents a conglomeration of this information, linked in its entirety to the skeletal positions, this assumption is unproblematic. However, we still need a way of describing how a change from 2 to 1 or 1 to 0 morae involves a change in length, but a change from 3 to 2 does not. Let us suppose that a skeletal position attached to a syllable nucleus cannot hold less than 1 mora. This is a fairly arbitrary supposition, I concede, but it gets the job done. Perhaps more argument on its behalf would be possible if this discussion of morae were not a mere detour here.

I propose that the Gothic rules are as follows, and that they apply in this order:

1) Mora Loss (ML): The rightmost mora in a word is deleted.

2) Coalescence (C): The information of vowels in an ending coalesces.

3) Final Syllable Neutralisation (FSN): A vowel can have no more particles than morae. Additional morae are deleted as follows.

3.1) $\{x, x\} \rightarrow \{x\}$

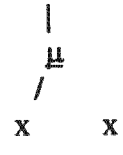
3.2) **U** is deleted rather than **A**; A is deleted rather than I.

Rule (1) is justifiable independently of the data. Mora loss acts on the final syllable of a word. It thus positions itself with regard to syllable by counting from the right. It is simpler to suppose that it positions itself within the syllable in the same way. Counting from the right, it would delete the rightmost mora. It would be silly to suggest, all else being equal, that it counts from the right to find the correct syllable, and then from the left to find the correct mora within that syllable, deleting the first mora in the final syllable. Note that this rule deletes a mora, but not the information carried by that mora. Rule (2) is roughly equivalent to a prohibition on hiatus between a thematic vowel and inflectional vowel. We do not find two such vowels in hiatus in Gothic, and thus this rule is unproblematic, too.

Rule (3) is based on the definition of a mora in its general principles, and on the data in its details. I will present various examples of these rules in Gothic below, beginning with a completely regular example to show the simplicity of this model, followed by what is assumed to be an irregular example, to show its strength.

Nom. sing. *ō*-stems:

ML[31]:



FSN (3.2): **{A,U} → {A}**

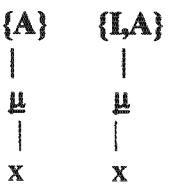
Hence *geb-a*, 'gift nom. sing.'

The PGmc genitive singular of *a*-stems is usually posited as **–eso* with no thematic vowel. This ending is thus irregular on two accounts: firstly, it uses a

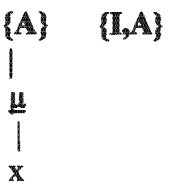
[31] At this stage, the rightmost skeletal position is detached, because of the supposition above that a skeletal position cannot hold less than one mora. If it were to remain in this case, one mora would be shared between two skeletal positions.

pronominal form of the genitive (with a final *–o*) rather than the normal nominal *–*e/os*. Secondly, it lacks a thematic vowel. Since the Gothic form in *–is* has only one mora, we cannot assume more than two morae for PGmc. However, under the above rules, a regular gen. sing. in **–a-es* (i.e. with the regular thematic vowel, and with a proper nominal ending) would yield Gothic *–is*, as follows. Here I will not show consonants, as the final *s* does not change during derivation:

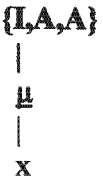
Gen. sing. *a*-stems:



ML:



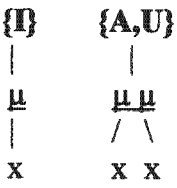
C:



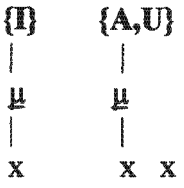
FSN (3.1): **{I,A,A} → {I,A}**
FSN (3.2): **{I,A} → {I}**

Hence *wulf-is* 'wolf gen. sing.' I think the only suspicious move here might be the coalescence of the remaining information. However, since this rule allows an extra case to be more regular on two counts, I see no problem in positing it, especially since this is merely a detour. The usefulness of the rule can, however, be seen in the derivation of the gen. pl. of *i*-stems. Both Prokosch and Wright claim that this ending is irregular, and that it had not yet been properly explained (cf. Wright, 1924: § 179). If the above rules manage to regularise this ending, this in itself might be taken as sufficient evidence for them. The following thus posits PGmc *i-ōm* with a regular thematic vowel and case ending. Again, the consonants are not shown (though the *m* was deleted at some intervening stage).

Gen. pl. *i*-stems:



ML:



C:

FSN (3.2): **{I,A,U} → {I,A}**

Hence *gast-ē* 'guest gen. pl.' I will not push the analysis much further for other languages however. All that I hoped to demonstrate here is that mora loss is a rule affecting information: it reduces the amount of information possible in an ending, and this point could have been made more simply just by observing that mora loss has led to neutralisation in final syllables. The exact details of this loss in other languages are thus unimportant, and even the rules given above are merely intended to show roughly how mora loss affects information. Furthermore, it is psychologically unrealistic to suppose that the rule acted in exactly the same way millennia later, for how would speakers of a Germanic language around 800 AD know what had happened around 300 AD? The details of the rule might thus have changed, though the core idea (that of information loss) remains stable.

As a brief (and unsubstantiated) aside, however, I would like to make the following suggestions, to show that this account is (in its essentials) applicable to other languages. Much Old English data can be accounted for be re-ranking the particles in rule 3.2. My use of "rank" here merely suggests historical ordering, and should not be taken to imply any association with Optimality Theory. If the Old English version of 3.2 is "**A** is deleted rather than **U**; **U** is deleted rather than **I**", then we would expect PGmc. *–o* to become *u* in OE, but *a* in Gothic. This fits with the OE nom. sing. *giefu* vs. Goth. *geba* 'gift'.

A final phonological process that needs mentioning is the loss of /j/ in certain contexts. The description of West Germanic Gemination is as follows (from Lass, 1994: 35):

-VCjV- > -VCCjV-
condition: C is not /r/

Further, the first V is short. After causing this gemination, the *j* was deleted, leaving pairs like *sellan* < **/sal-jan/* 'sell' vs. *sealde* < **/sal-ða/* 'sell: 1,3 pret sing.' (Lass, 1994: 70). The deletion of the *j* was entirely independent of mora loss, but just as the *i* deleted by mora loss mutated the preceding vowel, so did this *j*. This *j* often carried grammatical information before its deletion, however. In the above example, since the present has a *j* and the preterit does not, it might be possible to claim that the *j* helped signal a tense distinction. Though the information it carried was thus redundant (given the typical coronal preterit marker), redundancy is by no means unusual in language. In other instances, the *j* indicated derivational information, showing denominal or deadjectival verbs. Lass (1994: 71) gives as examples *trymman* < */trum-jɑn/ 'strengthen' beside *trum* 'strong' and *blēdan* < */βlo:ð-jɑn/ 'bleed' beside *blōd* 'blood'. Similarly, the *j* sometimes indicated a causal verb.

3.3 Semantic loss

I will have considerably less to say about semantic loss than I did about phonological loss. In part, this reflects the focus of this thesis as a whole: sound change. However, a detailed discussion of semantic habits follows in chapter 4. It is generally true that Germanic lost many of the inflectional distinctions of PIE, though I will not go into the details here. Rather, I wish to draw attention to a particular process, again focusing on nominal paradigms.

In PIE, we generally posit a set of thematic vowels and a set of inflectional endings. The thematic vowels, when used, are attached directly to the stems, and are class markers: they specify to which paradigm the stem belongs. As a result, though, they occasionally have additional grammatical meaning, in that they might indicate that the form is a causative or an inchoative by virtue of the fact that they display

membership of a class that is causative or inchoative. The inflectional endings, on the other hand, follow the thematic vowels (though naturally if there are no thematic vowels in a certain paradigm, the inflection is attached directly to the stem. Cf. Lat. *rek-s* 'king: nom. sing.'). While the thematic vowels mark membership of a paradigm, the inflectional vowels mark position within a paradigm: they show the case, for example, of nouns. The thematic vowel and the inflectional ending, then, display different kinds of information.

As a result of phonological loss (a specific instance of which, mora loss, has been discussed above), the inflectional ending often disappears entirely. The Go. acc. sing. *haird-i*, 'herdsman', for example, shows the *i* thematic vowel of its class, but has lost the nasal accusative marker, as seen in PIE **-om* (Voyles, 1992: 228). Nevertheless, this *i* now signals accusative case, in that it contrasts with nom. sing. *hairdeis*. It seems, then, that the case information has shifted leftwards. As expression of the endings erodes (in those languages where this happens), two options are possible. Either the grammatical information carried by that ending can be lost, or the information can shift leftwards.

The latter strategy is what we see in the above example, while the former is seen in the change from Latin to Old French nominal paradigms. Regular sound changes in this case involve the merger of the lax high and the tense mid vowels: /ĭ/ merges with /ē/, for example, and /ŭ/ with /ō/. Further, final nasals were deleted. The formal distinction between 2nd decl. acc. *– ŭm* and dat. *–ō* was thus obscured, and the semantic distinction was also lost.

Since both these options are possible, the erosion of the ending does not explain why the information shifts leftwards in Germanic, or why the information was lost in Old French. However, the leftward movement is part of a description of what happened, and this descriptive fact of a leftward movement of semantic units can be used as part of a further explanation. The distinction between Germanic and Old French in this regard will be discussed in chapter 5.

4) Language Habit

It has been suggested above that the notion of information is relevant for *i*-umlaut. The historical context involves mora loss, an informatic change, and the rule formalism appeals to particles, which allow for an informatic treatment. Before I can give informatic maxims, however, I will need to examine what kind of maxims we should be using. Keller's maxims were found to encompass a fairly broad category, including instinctual, rational and habitual behaviour. Though the phrasing in his examples tends to emphasise the intentional or rational mode, elsewhere he does state clearly that this is not how he conceives of action (Keller, 1994: 40-41). This chapter develops an account of what language habit is, before examples of habitual maxims can be presented in the following chapter. The importance of habitual behaviour in treating language change in general will be outlined in 4.2, comparisons being made with the claims of Lass (1997: ch. 7). Habit will be compared with Deumert's discussion of structuration in section 4.3.

4.1 Meta-theory caveat

Much of this discussion of information strategies is the result of abductive diagnosis. That is to say, there is little independent evidence for some of my claims in the following chapters, apart from the fact that they give a sensible account of *i*-umlaut. Explanations of phenomena of the third kind typically provide this form of insight, however. My claim about information strategies will certainly fit the empirical data, but I suppose that many other claims could do the same. The proof of this pudding, then, lies in extent to which my eventual explanation of *i*-umlaut is accepted. This is not entirely too problematic, as my claims about information strategies do not step very far beyond the limits of the empirical, and they are based on clearly argued discussions from Lass (1997: ch. 7) and Deumert (2003). My claims may, however, be open to the criticism that they are either *ad hoc* or tautological, but the extent to which this is true or even relevant remains to be seen.

Berg gives an example of such criticism:

> In a discussion of onset and coda structures in various languages, Goldsmith (1990) states that a greater number of phonological contrasts are usually accommodated in

onset than in coda positions ... Goldsmith accounts for this asymmetry by introducing the concept of 'licensing'. He asserts that onset and coda segments are licensed by different 'licensing agencies', the former by a primary licenser (the syllable) and the latter by a secondary licenser (the coda itself). The greater restrictions on codas are 'explained' by their being less freely licensed. This is either a (tautological) restatement of the basic observation or the invocation of hard-to-justify assumptions whose ad hoc nature severely curtails their explanatory value. (Berg, 1998: 1)

A criticism of *ad hoc*-ness translates into the problem of unfalsifiability. However, given the nature of the form of explanation I am pursuing here, falsifiability is not automatically the problem it is in a (pseudo-)scientific account. The habits I am about to investigate are admittedly unobservable in themselves. However, as shall be discussed, they are directly perceptible in human action, or through historical analysis (as distinct from scientific practice). Take the case of *Weib* presented above. At an intermediate state of the language, where the word is becoming pejorative, we can know this either by observing that people tend to use it more pejoratively as time passes (i.e. by observing human action). Alternatively, we might compare two states, one where it is not pejorative, and a later one where it is, and supposing that it became pejorative at some intervening stage (i.e. by conducting an historical analysis).

The accusation that my claims might be tautological is more relevant and potentially more damaging. In Berg's criticism of Goldsmith, however, we see that the tautology lies in the fact that something linguistic is explained by something else linguistic, at a greater degree of abstraction. In that rules apply to abstract forms to yield less abstract forms, so that the one is related to the other in a predictably causal fashion, it is possible to suppose a large degree of isomorphism in such instances. Nothing new is added, in contrast with Popper's injunction:

> The new theory should proceed from some *simple, new and powerful unifying idea* about some connection or relation ... between hitherto unconnected things ... or facts ... or new "theoretical entities". (Popper, 1963; quoted in Berg, 1998: 11)

By explaining linguistic phenomena in terms of extra-linguistic phenomena, then, this isomorphism is made impossible, and the chance of tautology is reduced. Berg claims that a good explanation of linguistic phenomena must ultimately come from

outside of linguistics (Berg, 1998: 18). Though I may not agree that this must always be the case, my discussion here at least complies with such claims. Communication, language and social habits are related, but not fully equivalent. They are ontologically and conceptually different, and any translation from one system must at best be only partially successful. Though an informatic claim might appear very similar to a linguistic one, a degree of methodological separation underlies the similarity.

4.2 Language habits[32]

Given the entire gamut of language types, from an agglutinating language such as Inuit to an almost completely isolating language such as Mandarin, or from SVO to VOS, or across any other such continuum, it is notoriously difficult to find absolute universals (Greenberg's universals often being implicational rather than absolute). In terms of grammatical distinctions, many languages get by without tenses (Biblical Hebrew), others without number in nouns (Mandarin, with a few exceptional animate plurals). Even semiotic properties are difficult to treat as universals: the typical claim about language optimally relating one form and one meaning (Lass, 1997: 342) cannot stand up against the creation of a huge number of homophones in Mandarin, or IE inflectional endings (which typically encode various pieces of information such as case, number, gender, or paradigm in one morpheme.

In the case of homophones, we might re-evaluate an earlier schema, and represent the fuzzy or abductive *modus ponens* used in interpreting a sign as follows:

$$e \rightarrow c_1 \ldots c_n$$
$$\underline{e'}$$
$$c_x'$$

Which is to say, our experience has taught us that a particular expression is generally linked with various meanings, c_1 to c_n. We hear something which we infer to

[32] The material covered in this chapter stems from a non-existent dialogue conducted somewhere between my computer and Popper's world 3. I had formulated the contents of this chapter pretty much *in toto*, and then proceeded to read Lass (1997: ch 7). I found that much of my material was similar to Lass's. Nevertheless, given the greater subtlety and cohesion in Lass's chapter, this reading brought to my attention various criticisms professor Lass would no doubt have made of my original if he were to have read it, leading me to tweak various details in producing the current version. In particular, I owe the introduction of hermeneutics here to the reading of his chapter.

be a token of that expression (e′), and having made this decision, we infer which of these meanings of the homophone (c_x′ such that $1 \leq x \geq n$) is intended in context, though of course we may be mistaken in this decision. By adding homophones to the schema, then, we are merely increasing the amount of inference needed in interpretation. Inference is needed in any case in establishing that a certain token represents a certain type, and in deciding exactly what subtle sense of any given c is intended in context, as c′. Allowing for various contents to be associated with one expression thus does not depart much from the normal workings of semiotics, contrary to the usual semiotic claim about one-to-one optimality. A sign (as suggested by the data from the world's languages) need not be an exclusive link between one sound and one meaning. It must, however, involve some inference. Some signs (those pronounced badly, or those that are homophonous) will need more inference than others.

Imagine, however, a lone speaker of English who (for whatever reason) suddenly ceased to contrast voiced and voiceless stops. A large number of homophones could suddenly be created (for example, the distinction between *big*, *pick*, *pig* and *Bic* would be lost), and any normal speaker of English would be flummoxed when presented with such speech. We are simply not used to having to perform as much inference as this would entail. This is not to say, however, that we cannot perform such inference, or that that we would not eventually learn to cope with it if we moved into an area where such a fictitious rule had actually come into effect.[33] There is little reason to suppose that we would not learn to do something that Mandarin speakers do as a matter of course. We in fact do something of this sort all the time, to a lesser degree, whenever we are faced with a new variety of our own language.

Never having heard the South African /lɑ:k/ for 'like', a speaker of RP would have to infer what was meant (though the context would help in establishing that whatever was meant, it probably wasn't 'lark'). If the RP speaker heard it only once every 10 years (and consequently did not remember it), he might have to perform as much inference each time he heard it. But if exposed to it regularly, it would become part of his experience or ground (in that he would be more likely to remember it). Once part of his ground, the strain on inference would be lessened. The distinction in

[33] Lass (2003, personal communication) claims that this has indeed happened in some dialects of German.

RP between /ɑ:/ and /aɪ/ and the lack of this distinction in many forms of South African English are thus both meaningless in semiotic terms, for in both cases the hearer can infer the c′ from the e′. The inflected distinction between future and present in Classical Greek and the lack of such a distinction in English is just as meaningless.

Simply because a sentence in Mandarin does not indicate tense as such, it does not follow that a Mandarin listener has no idea of the time of action in a sentence of his language: that would be subscribing to the untenable Whorfian hypothesis. Such a hearer may infer the time of action, if it is relevant, and if it has not been expressed through periphrasis. Similarly, a French speaker can infer from context whether a present tense verb is punctual or continuous, if there is any need to (and a speaker can use periphrasis to indicate such information, as *je suis en train de* ..., if such information needs to be conveyed with absolute clarity).

Semiotic inference is an important part of a non-Whorfian picture of language, for we must assume that hearers have some means of becoming aware of information that has not been overtly expressed in language. We might simply claim that speakers of Mandarin or French are more used to performing a certain type of inference than speakers of English, or conversely, that speakers of English are more accustomed to producing information about the aspects of the present tense. Similarly, Mandarin speakers are used to inferring which of many homophones are intended, while English speakers are not accustomed to interpreting homophones on so grand a scale.

This is just a more abstract form of the old chestnut about the conservation of energy in speech. Just as the less energy a speaker expends in articulating clearly, the more effort a hearer expends in working out what was said; so too the less information a speaker gives (about tense, aspect, number, gender, etc.), the more inference a hearer has to perform (if they wish to work out such information). From a sensible semiotic point of view (and not the one-to-one-optimality point of view), then, the English distinction between present and present continuous involves expending time and energy on expression; failing to make such a distinction involves expending time and energy on inference.

What language could possibly ensure total optimality between these two extremes at all times, though? How would any individual compare how much energy he expends in expression when speaking with how much he expends in inference when

hearing? The entire process must be rather loosely construed, possibly as a trial-and-error process. We thus would expect instances of language to see-saw back and forth across a theoretically optimal point, without any speaker being aware of such a point. More particularly, language being such a manifold structure, we would expect various sub-systems of language to be at different sections of the semiotic continuum.

The above discussion, however, suggests a way of reducing the chance aspect just mentioned. We might claim that habit situates a speaker/hearer's behaviour somewhere along this continuum for any given aspect of his language. Just as Mandarin speakers are accustomed to inferring tense or number, and French and German speakers are used to inferring the aspect of the present, and Finnish speakers can infer whether a noun is definite or indefinite, English speakers are in the habit of distinguishing tense, number, present aspect, or the definiteness of nouns. Expression of such distinctions and inference in the absence of physical expression are both habitual.

There is little evidence to suggest that any such setup is superior to the others (and I will shortly discuss why such a functionalist move would actually be anti-semiotic), so a speaker of any of these languages follows the habits of his language simply because that's what his fellow speakers do. Inferential habit is thus similar to the semiotic ground: both are habits or social practices. Ontologically, French's semantic distinction of *chaise* and *fauteuil* is not very different from English's grammatical distinction of present and present continuous. Speakers learn to do both since they must conform to social practice in order to communicate effectively, and though such habits may change slowly over time, a current French speaker is unlikely suddenly to cease distinguishing *chaise* and *fauteuil*, just as an English speaker has no choice to continue distinguishing punctual and continuous aspects as long as his community continues to do so.

My discussion above can thus be anchored in Lass (1997: ch. 7). Lass argues that

> We live perpetually with the 'decisions' of past generations. Somebody, somewhere (as it were) decided in the eighteenth century or thereabouts that the expression of progressive aspect should be obligatory in English, and as an English speaker I'm simply stuck with it. It is not correct to say that my 'world-view' somehow focally involves the distinction between continuous and habitual action (to take one possible interpretation of a kind of flaky Whorfianism), or that this distinction is 'meaningful'

for me; any more than it would be to say that a German 'doesn't care'. The fact is that what our language happens to encode grammatically has no necessary relevance for us, or any 'significance', (a) because language does not control or influence thought ..., and (b) because many important 'decisions' about contemporary linguistic characters were made generations ago, and we have no more control of them that we do of the number of our vertebrae.[34] (Lass, 1997: 368).

It is in this sense that we should accept

a view of language as being at least partly 'transcendental' or metapersonal, a system with which speakers 'interact', but which is in some sense 'outside' them, and 'extra-mental reality'. (Lass, 1997: 365)

Or the fact that

[a language] will also have a lot of material that simply *makes* speakers do things, whether or not there is any (functional, discourse, pragmatic) 'need' to do them. (Lass, 1997: 367-368)

And it is in this sense that I have discussed the possibility of treating language as an environmental consideration within Keller's model (as well as something internal, for a different sense of "language"). Language makes speakers do things, regardless whether these things are independent signs, or only grammatical distinctions. Keller's model, then, in its essentials, is compatible with Lass's arguments. A few differences remain to be ironed out, which I will briefly do here.

A central distinction between Lass (1997: ch. 7) and Keller (1994) is their treatment of speakers, though this boils down the different senses with which they mean "action". This is one of the problems of using ordinary words to refer to unordinary things. Lass argues, for various reasons, that language change is not the consequence of action, whereas Keller argues that speaker action is at the heart of language change. However, it seems Lass's sense of "action" is that of the Boden quotation above (the rational mode), while Keller clearly generally means "habitual

[34] One distinction between my language and my vertebrae, though, is that I (subconsciously) choose to communicate using the codes of my environment: there is no biological imperative forcing this choice.

behaviour" (though this doesn't seem apparent in the phrasing of his examples). Both Lass and Keller agree in principle that speakers have little agency in the intentional sense. Further, in Keller's model, language change is the immediate consequence of results, not directly of actions.

Various points made by Lass also fall out of a Keller-communicative hybrid. Firstly, although a community is just the collectivity of speakers in a physical sense, it has various properties *qua* collectivity that reside in no individual speaker. It is by virtue of the fact that actions happen in a collectivity that a phenomenon of the third kind is possible (Deumert, 2003: 61), and this fact is not reducible to a feature of the behaviour of any given individual speaker. This can be contrasted with Lass's claim that a theory of language change taking speakers into account is committed to the fact that 'a speech-community is nothing more than the sum of its members; i.e. it has no emergent properties, or properties any different from those of any individual speaker. (Lass, 1997: 362)' The actions of each individual are irrelevant in themselves for a *description* of the macro-process, though the speakers are important in a *causal* sense for the level of results.

Further, given the fact of communication, language is made something quite different from fashion. I settle on fashion here to provide a contrast with Eco's semioticisation of tight jeans (Lass, 1997: 340n). Both language and fashion allow a user a certain amount of leeway. However, my clothing is something of my own choice, something over which I have control, and (most importantly) not created by my interaction with anyone else. If I dress outrageously, I may isolate myself from my peers, but I still succeed in the action of clothing myself. Communication, however, demands that at least two people use a comparable code: it is intersubjective in a way that fashion is not. If I depart too seriously from linguistic norms, I will not be able to claim that communication has happened at all. There is a limited amount of possible variation (structured other otherwise) before I cease speaking my language at all, so that each person's desire to communicate enforces a form of population integrity. In this sense, the requirements of communication add something to the correctness of grammar.

On this point, Lass draws an analogy between language and evolution. He describes a language as a quasi-species. 'A quasi-species is somewhat oversimply a hypervariable but (limitedly) self-stabilising population' (Lass, 1997: 375). It is self-

stabilising in that the limitations imposed by the need to communicate using a similar code reduce the amount of error in copying, to the extent that

> The region in sequence space can be visualised as a cloud with a center of gravity at the sequence from which all the mutations arose. It is a self-sustaining population of sequences that reproduce themselves imperfectly but well enough to retain a collective identify over time … That cloud is a quasispecies. (Eigen, 1993; quoted in Lass, 1997: 375)

As the delimitation of a species as a population depends on the condition of breeding, so too does the delimitation of a language depend on communication (though the analogy is quite loose). In that communication in this sense presupposes compatible grounds, and in so far as a ground is a habit, it seems reasonable to suggest that the concept of language habit and a compatible semiotics are both useful for Lass's 'modest ontological proposal' (the heading of Lass, 1997: § 7.6).

Both Lass's and Keller's models allow speakers a certain amount of leeway within such a quasi-species. A habit is not a physical law like that of gravity; and it is reasonable to suppose that habits may alter with time. However, it is also reasonable to suppose that at a given time, certain habits are more deeply ingrained than others. It might thus be possible to identify certain habits as being essential to a given language, so that if such habits are contravened, a speaker cannot reasonably said to be speaking that same language any more. If I suddenly fail to distinguish singular and plural, or durative and non-durative present, I will still be understood (with difficulty), but in what sense if any will I still be speaking English? More extremely, if I use or use /dɒg/ to mean 'elephant' or 'blunderbuss' instead of the usual 'dog', I will not even be understood.

In the first example, a larger than usual amount of inference will be needed; in the second example, no amount of inference is likely to help work out what such a speaker meant. There thus seems to be an inference continuum, ranging from usage heavily supported by experience to usage that is incomprehensible. At some point along this continuum, we might speak of bad English; at a further point we would cease to refer to it as English at all. Naturally, this is not the place to formalise a methodology for dividing up this continuum, if indeed it can even be done.

Nevertheless, on the basis the above, it seems reasonable to suppose that habits delimiting such inference[35] can form a core aspect of what it means to be speaking a particular language.

I will make an initial suggestion here, to be expanded upon below, that different habits constrain behaviour along certain sections of this inference continuum. The following refers to a speaker of New York English, as described in Labov (1966; cited in Mesthrie *et al.*, 1999: 84-91). Such a person would regularly use **/r/** more often than others in certain contexts. Though obviously one form may be considered more prestigious than another, the use or non-use of **/r/** are both possible within the rules of grammar of one particular variety. Here experience supports a wide range of usage. On the other hand, with something like the present tense, one is constrained in having to distinguish the durative aspect from the punctual: experience here supports a more constrained usage. We might thus distinguish strong habits (which brook no exception) from weaker habits (which are generally implemented, but allow exception).

4.3 Habit and structuration

Although I have tried to show that I mean "habit" in a very specific sense above, a sense taken from both Keller and Lass, it is often tricky using everyday words for technical purposes. Furthermore, though my claims about habit seem coherent with various facts, I have given little independent support for them. In this section, therefore, I will attempt to clarify both these issues, and show that my insistence on habit as an important concept is more than mere quibbling, by presenting a brief discussion of structuration. This will put me on a firm enough footing to finally address a distinction I made in section 2.1 between functional and historical approaches.

This is not the place to present a full exposition of structuration and its assumptions, so instead of arguing on its behalf, I will take the value of its claims as assumed, and will thus generally limit myself to quoting claims from Deumert and to relating these claims to points I have made above. While Keller's discussion provides

[35] Examples given above include Mandarin speakers' habitual inference of number or English speakers' being unaccustomed to infer aspect in the present tense.

us with a clearer ontology, Deumert's provides a clearer description of agency. Reading both accounts complimentarily thus allows a broader picture of the issues.

From diverse perspectives (including the philosophy of action, sociology and neuro-psychology), Deumert (2003) presents a thorough argument showing that the intentional mode of action is more limited than is generally thought. Often, in attributing intentionality to their own actions, humans are merely accounting *post hoc* for something that was not caused by the attributed intention (Deumert, 2003: §§ 6, 7): the intention is a psychological illusion. She thus claims that

> there is no need to assume *a priori* that the structure of conscious reasoning (which furthermore seem to rely on abilities of verbalisation couched in some kind of hypothetical mentalese) adequately describes the non-conscious cognitive (and bodily) processes which underlie human agency and which moreover occur in a brain which is best characterised as a complex parallel system and not a digital computer sequentially manipulating symbols according to specified rules. (Deumert, 2003: 35-36)

Clearly, "agency" here is meant in a similar sense to Keller's discussion (1994: 40), i.e. including the rational, instinctual and habitual modes defined above: the term thus focuses on human doing and not the reasons for the doing. In studying human action within this framework, then, instead of asking why an agent did something, we ask what it is that he did (i.e. what kind of thing it is an instance of). In certain cases, as in language, what he does is an instance of social practice and thus, simultaneously, a cause of social practice.

> In other words, individuals through their actions engage in social practices, thus creating, reproducing and changing social structures. At the same time, these structures form the contingencies (rewards, possibilities, constraints, etc.) which shape human agency. (Deumert, 2003: 53)

In the case of language, it is the condition of communication that enforces this reciprocity. As a result, the social practice and the action, rather than being separate entities, exist in a type-token relationship: the action merely instantiates the practice. Though this last might be a slight oversimplification, it will cohere well with my semiotic claims. Experience of other people's communication forms an experiential

ground in one's memory. One then does something based on that ground – one speaks. This act of speaking, however, in turn becomes part of one's listener's experience, and if it remains in memory, it is part of their ground. Compare this claim with the previous quotation from Deumert for the similarities between sign and social practice: the sign's ground, then, is merely a lexical case of social practice.

Each act of speaking uses tokens, of which the ground is the type. Above, however, I claimed that social practice is the type. The former is mental, but the latter is external to any speaker, and they are thus not identical. However, as discussed above, the desire to communicate will ensure that subjective grounds cannot not deviate too far from each other within a speech community. There is thus the possibility of assigning the social practice a virtual reality, and though this is thus ontologically different from the ground, the two are very similar in terms of content. I thus think the use of "type" to refer to either is acceptable.

It would be ridiculous to say that the type is the reason for the token, or that it causes the token. A speaker desires to say something, and this commits him to employ a certain type, of which his actions will be tokens. It is the desire to speak that is causal, not the ground. Similarly, the social practice does not cause action (unlike an intention), for the action merely instantiates the practice in suitable contexts. Structuration's discussion of social practice is thus significantly different from the functionalist approach.

This move to structuration as a clearer explication of habit brings with it several methodological advantages. Consider:

> To see human action in terms of *doing* rather than *intending* allows Giddens to proceed from a third-person, external observer's perspective (characteristic of all science) and to avoid the pitfalls of a first-person epistemology which aims at reconstructing the possible subjective intentions of actors based on the introspective, first-person reasoning of the sociologist or linguist … (Deumert, 2003: 55; emphasis Deumert's)

Unlike a hermeneutic approach, then, this avoids the pitfalls inherent in claiming we can know the subjective states of other minds, which is what we do in attributing intentions to speakers (Deumert, 2003: 41). Similarly, the attempt to argue from such intentions is itself fraught with subjectivity: 'other people's abductions … may fail to convince, since the nature of a particular abduction depends on contingent attributes

of the abducer' (Lass, 1997: 335). Above, I advocated diagnosis (abduction) as an important feature of explanations from the invisible hand. Now this can be reformulated to take cognisance of these new distinctions.

If a child copies someone else's behaviour, what do we (as third party observers) know? It may have copied the action because of a promise of reward or punishment, because of boredom, admiration, mockery, or any number of reasons. The child's mental state, which gives hermeneutic meaning to the action, is unobservable. We can argue with a greater degree of certainty, however, that the child's action is based on its experience of the action that it is copying (even if copying creatively). This experience resides in the child's mind in some form, and so we do abduce to something mental even in this case. These two kinds of mental thing, however, do not relate in a similar way to the observations and our theories about these observations.

In supposing an intention, we are creating a link between two completely different things: an act and its cause. In supposing the experience, however, we are linking something empirical with a mental copy of itself (i.e. we are linking a type and its token): this thus fails the conditions for a true code, and thus for meaning. The existence of the copy can be argued for in non-mental terms. For if a child has never heard or read "to be or not to be", it is very unlikely that it will spontaneously say /tʰubiʔɔnɒttubi/. If, on the other hand, we hear it say such a thing, we can reasonably suppose its mind contains the memory/experience of having heard it. This supposition, however, is based on the fallacy of affirming the consequent, and it is thus still abductive.

The first case, a link between cause and effect, allows for hermeneutic meaning. The second, a link between type and token, does not. We can thus distinguish the hermeneutic approach, which abduces to a meaning, from the historical approach, which still abduces to habit, without claiming that such a thing is meaningful. There does exist a link between the mental habit and something else, but this something is external to the subject. This is analysable on historical grounds, though, being the social practice of a certain socio-historic context (Deumert, 2003: 53). The hermeneutic approach claims that its abduction explains something; the historical approach *on its own* does not. This distinction is at the root of my discussion of the criticism of *ad hoc*ness in section 4.1.

The positing of a habit, then, merely generalises this empirical, third-party kind of abduction to cover a group of people rather than one individual. The desire to communicate lends this move from individuals to a society a great deal of support, for it gives us a reason for supposing that the habits of individuals speaking the same variety of the same language are comparable. The habits do reside in the mind, but they merely link types and tokens according to the principles of semiotics.

Structuration thus highlights a sensible methodology in the wake of Weinreich *et al.*. However, in conflating the micro and macro[36], it is less successful. Lass (1997: 363n) claims that there is heuristic value in maintaining a distinction here. It is useful for us to do linguistics based only on the macro-structures, for these structures can sensibly studied on their own. Invoking speakers for every problem at every level of analysis would be prohibitively time-consuming and complicated. We wouldn't see the wood for the trees, and it would be an inelegant way for the discipline to proceed. Furthermore, there is little reason to suppose that all language is a matter of social practice. For such reasons, some distinction between micro and macro ought to be preserved.

In Keller's model, such a contrast is preserved. Speakers' actions are relevant at the level of results. These results lead to a linguistic consequence which can be studied in its own right. The linguistic structure at the level of consequence, however, is part of the speakers' ecological conditions. Speakers' actions, collectively, have consequences for structure, and structure has an effect on speakers' actions. Keller's model of explanations allows this important feedback without equating action and structure under the heading "social practice".

Having thus established my position more clearly than before, I can turn to the important matter of contrasting my position with an opposing view. In section 2.1, I outlined various possibilities for dealing with maxims of sound change. Although I have expanded "maxim" into something that includes instinctual, rational and habitual behaviour, and have since focused on habitual behaviour, the distinction still stands. I argued that, to explain sound change, we need to link sounds with something language-external. This link can be meaningful or accidental. Functionalism supposes the former; Lass, Deumert and Keller suppose the latter.

[36] Which it does on the basis that individual action and social practice are no longer distinct entities.

Arguments against functionalism have already been made in the paragraphs preceding this, and more lucid arguments of this nature are given in Lass (1997: ch. 7) and Deumert (2003). Rather than rehash these arguments inexpertly here, I will try a different tack. The arguments in Lass and Deumert show why it is false to assume that all acts mean something, or that all acts are motivated, as the hermeneuticians and functionalists seem to think. While I agree wholeheartedly with these criticisms, I will grant hermeneutics its claim in the following, for the sake of argument. I hope to show that this still does not help their cause.

This is mainly motivated by my espousal of the semiotic cause, for semiotics has also been pressed into service by hermeneutics and functionalism. If I wish to argue that semiotics is useful in explaining sound change, it would be preferable to clarify how I think this might be the case by excluding competing interpretations. In the following, I will argue that functionalists typically make an unwarranted assumption in dealing with experience and grammatical distinction. My argument stems from an analysis of teleology as a higher order semiosis, and from a comparison between this and normal (sound/meaning or linguistic) semiosis.

Both the hermeneutic and the functionalist approaches can be characterised as assigning meaning to action. This is overtly clear in the case of hermeneutics:

> The concept-dependent existence of cultural objects implies that the proper scientific attitude towards them is the hermeneutic one in which we attempt to understand their meaning, and that explanation of them and of the relations of human agents to them will be given in the vocabulary of beliefs, reasons, motives and purposes rather than in the mechanistic vocabulary of causes and laws. (Pateman, 1987; quoted in Lass, 1997: 337)

Structuration shows that Pateman's assumption of such a dichotomy here is unwarranted: Keller refers to this as the 'prison of dichotomies' (Keller, 1994: ch. 3). I will let the hermeneutic claims stand, though, to turn to functionalism. This form of argument is more covertly preoccupied with meaning, though the sense of "meaning" is quite similar to what we find in hermeneutics. Functionalism supposes that we should understand or explain an action in terms of its function. In linking two entities in this way, we are clearly dealing with a true code. More particularly, in that this model thus interprets one thing in terms of something else, we are clearly dealing with the process of semiosis, similar to hermeneutics' assumption that acts are meaningful.

We can thus further clarify the difference between the structurative approach and the functional: the former links individual act and social practice in a type-token relationship, the latter in an expression-content relationship.

Recall that semiotics links expression and content on the basis of a ground, which is constituted out of experience. Recall, too, Saussure's claim that '*le lien unissant le significant au signifié est arbitraire*' (Saussure, 1922: 100). It is not merely the coding of expression and content that is arbitrary, for the structure of the s-code itself is, too. For example, there is little reason for French to distinguish a *chaise* from a *fauteuil*. If one makes such a distinction, it is merely because one must when speaking French. Distinction (and experience of such distinction) is thus language specific, according to the semiotic nature of an s-code. This is true regardless whether the distinction is semantic or grammatical. Just as French simply happens to distinguish *chaise* and *fauteuil*, Greek simply happens to distinguish present and future with inflection.

The claims of semiotics, however, reduce the functionalist and hermeneutic accounts to tautology. Let us say that speakers of Greek do *x*, *y* or *z* because they want to keep the future and present distinct. According to the claims of semiotics and structuration, however, the only reason we can give for their wanting to keep present and future distinct is that they speak Greek (for it is only their experience of Greek that causes them to make this distinction). Thus, the functionalist claim is that speakers of Greek do *x*, *y* or *z* because they speak Greek. Further, assuming that *x*, *y* or *z* fall under the heading of speaking Greek, the claim boils down to "speakers of Greek do things that involve speaking Greek because they speak Greek".

It seems that function can only explain action if the function can be argued for independently of the action. A typically hermeneutic claim here might be that it can be argued for from 'the common experience of being human' (Antilla, 1989; quoted in Lass, 1997: 338). However, structuration and semiotics show that claims of function are reducible to the following of custom, so that even if we allow that people do things in order bring about a result, the result is not independent of the action. The result is the following of custom, and custom is created out of the various actions of individuals within the community holding that custom.

Having thus presented what I hope is a clearer outline of what I mean by "language habit", we can turn in the following chapter to examine examples of such habits, and see how they may be used in explaining things. It is the case that people

follow habits blindly, but such habits may conflict with each other. The following chapter outlines how speakers may respond in the case of such a clash.

5) Habitual Maxims

Having argued against functional explanation in the previous chapter, I will turn here to showing how habit provides an alternative account of certain changes usually treated as functional. In the following, I will gradually introduce points about a process in Classical Greek, but these points will necessarily be interpolated with a great amount of discussion. I thus ask the reader's patience for what will be a disjointed account of an apparently simple story.

All Indo-European languages, to greater or lesser degrees, suffer from loss in their inflectional endings. Some languages retain a relatively high number of PIE inflections, others maintain very few of the total. Some, on the other hand, developed new inflection. Greek, for example, developed a sigmatic future (so called because the regular future is distinguished from the present by the addition of a sigma before the personal endings). This future was a post-PIE development, and it remains in modern Greek (though there the simple future combines inflection and periphrasis).

There was a regular sound change in classical Greek, however, that deleted intervocalic sigmas. This would have led to a loss of distinction between present and the regular future in stems ending with vowels. If a sigma was the only marker of the future, then, the deletion was blocked. We thus find words such as *αἱρήσω* 'I will take' beside *αἱρῶ* < *αἱρεω* 'I take', where the former has retained an intervocalic sigma. Similarly, *λύσω* 'I will loose' retains the sigma that distinguishes it from *λύω* 'I loose'. A typical functionalist claim is that Greek (or speakers of Greek) took preventative measures, and blocked the deletion in order to maintain the distinction.

Lass (1997: § 7.4.2) discusses why this is unsatisfactory for various reasons. Sigmas were retained in the aorist, though they were not the only method of distinguishing the aorist (for it contained the historical augment and used a different set of personal endings). Secondly, languages lose distinction all the time, and there is little independent reason why Greek should have to retain this distinction. For example, Greek did not distinguish indicative from subjunctive in the first person singular present. The retention of the sigma is thus not explained by the structural distinction between present and future.

The situation is improved by the introduction of habit. We can claim that Greek of a certain era habitually differentiated present and future, in as far as its speakers

regularly made this distinction in their speech. The only reason for supposing this is that the speakers do it, following the principles of structuration theory. We do not abduce to any need or intention to distinguish forms, but merely assume the entirely unproblematic claim that speech presupposes social norms, and that speakers must follow these norms to some extent in order to be able to communicate. It is thus a purely empirical matter. Further, given its presence in modern Greek, we can assume that this particular habit has particularly great historical momentum. Let me unpack this.

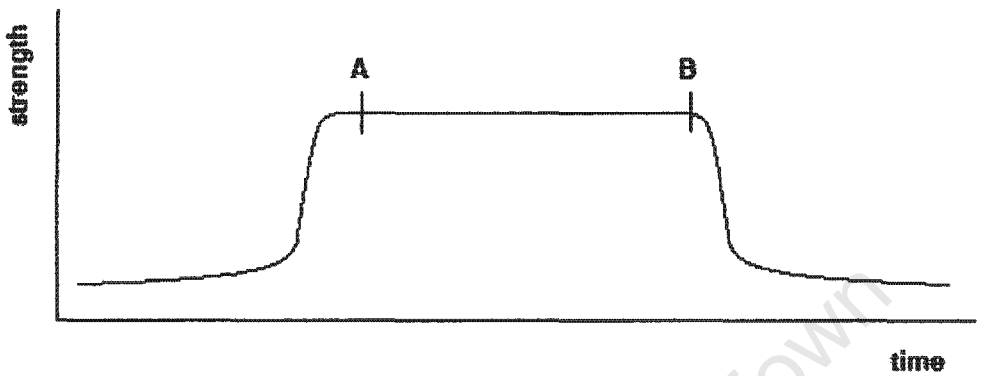
I think it might be argued that habits have a greater strength at some times than at others, for I have discussed how habits do not remain constant over all time. Strength is merely the regularity with which a distinction is implemented. We can assume, for example, that the continuous/non-continuous distinction in English was gradually introduced, and that during the initial period, the distinction would only have been made sporadically. In such an instance, people could obviously fail to make the distinction without anyone noticing or finding the fact odd. Even if someone noted the lack of distinction, I imagine it would have been considered merely unusual rather than ungrammatical. If a distinction is only used sometimes (i.e. the habit is weak), failure to use it departs almost insignificantly from social norm. If the norm is firmly established, however, failure to follow it departs from the norm in a more marked degree. In current circumstances, my failure to use the continuous present in its proper context would be quite a significant grammar mistake.

The same can be said for a habit in decline. In its prime, classical Latin distinguished up to six cases. By Old French, these distinctions had faded away to two cases. Failing to distinguish the genitive and dative cases at a certain time was quite drastic: it meant that one was no longer speaking Latin as such. Failing to distinguish these cases a few centuries later, on the other hand, was by no means as serious: we can imagine that some people simply did not often make such distinctions, even before these distinctions died out altogether.

A habit thus begins weak, grows strong, and then grows weak again, and dies (and there is little sense in wanting to explain why it goes through this life-cycle, or at which points it begins to grow or die). This addition of the dimension of time to the strength of a habit gives us its momentum. Since the addition of time is what allows us to draw the s-curve that tracks the rise and fall of the habit, momentum is merely what we see when we read the history of a change off such a graph. This momentum

is not an empty concept, however, for it can provide us with a way of explaining speakers' reactions that would not be possible if all we knew was the strength of a habit at a given point.

Consider the following very rough outline of what might charitably be called an s-curve, showing the rise and fall of a habit:



The strengths at times A and B are obviously the same: in both cases the habit was regularly implemented. In looking at the momentum, however, we can see that the habit continues regularly after A, but drops sharply after B. We might thus claim that the momentum around time A is greater than that around time B. In looking at a section of the curve instead of a point, momentum thus adds something to strength, though naturally it is only possible to speak of momentum from an historical perspective.

Momentum, as the rise and fall of a habit, is a matter of historical accident. There is no reason why English should have developed the continuous aspect when German did not. Similarly, there is no reason why Greek has maintained an inflected future, while the Romance languages have lost Latin's inflected future.[37] This has an important implication, though. Since momentum is an historical accident, it seems counter-intuitive to claim that the strength of a habit can be changed significantly in a causal fashion.

Compare, for example, the independent s-curves representing sigma deletion and the distinction of the sigmatic future. The sigmatic future was obviously a

[37] Claiming that the Romantic loss of the future is just part of a larger trend of loss merely begs the further question why these languages should display such loss at all, or why other distinctions (such as the participles) have not been lost in the same languages. These further instances of loss are simply

development pre-dating sigma deletion. There was thus a period when future was firmly established, while sigma deletion was incipient. At such a time, sigma deletion would be represented by the lag section of the s-curve. In this case, speakers would apply it in isolated instances, and failure to apply it would not be a severe violation of social practice.

The distinction of the future, on the other hand, would be represented by the plateau section of the curve, and more particularly by something like the section around point A in the above diagram, where the habit has great momentum. We can make this claim because Modern Greek still distinguishes a sigmatic future, suggesting that the last few thousand years of its history has not yet reached something like point B. In that the habit had great momentum, it was applied regularly, and failure to apply it would be a severe violation of social practice.

I have suggested that people are not likely to commit such anti-habitual acts, since the following of habit is what allows linguistic communication in a society. This is only true, though, in a period of high momentum, for in a period like that around point B, sporadic departure from norms is not impossible. Greek speakers of a certain era would thus be no more able *suddenly* to stop distinguishing present and future than a French speaker would be to use *chaise* to mean 'armchair' in the present situation.

The fact that an independent sound change was eroding the sigma would certainly have little to do with the habitual use of the future in its own right. We should not attribute unnecessary predictive or calculative skills to speakers, and can thus justifiably assume that no speaker foresaw the impact of sigma deletion on the future at the time when sigma deletion was incipient (and therefore sporadic), and the existence of the latter habit is thus not sufficient reason for speakers to abandon the former habit. One habit concerns a certain sound, the other habit concerns a certain distinction. It is only indirectly (though the sign) that this distinction is related to that change. The latter thus cannot have a causal effect on the former.

Regardless of the exact shape of the curve, and regardless of when sigma-deletion made its sharp rise, we can at least be fairly certain (given the limits on historical accuracy set by the quality and quantity of the historical witnesses) that at one stage sigma-deletion had the weak momentum of an incipient change, while the sigmatic future had the strong momentum of a regular habit. The very nature of creating speech

---

other habits, and their fall is thus inexplicable in the same way. I don't think that a conglomeration of

forces us to produce acts that instantiate social norms, or habit. Blocking sigma deletion in this particular historical instance, then, involved a significantly smaller departure from social norms than failing to distinguish present and future.[38]

It is thus reasonable to suppose that, after a certain amount of trial and error, people's experience began to include the fact that a future lacking a sigma was not in line with habit nor socially successful. As the application of sigma-deletion became less sporadic, it also became more patterned; people would not have habitually applied it in the case of the future, for such speech was closer to social practice. The general habitual application of sigma deletion increased, but this increase was based on previous usage. Since this previous usage was weaker in the future than it was in other environments, it did not grow in the future as it did in other cases. This suppression in certain environments amounts to environmental conditioning over time. The consequence, some time later, was that sigma-deletion had been blocked in the future.

This does not assume that speakers were able to perform complex computations to evaluate the risks of various options; rather it falls out of the trial-and-error process that constitutes an inferential semiotics. At the time when sigma-deletion was slowly beginning, it would have been applied in various isolated instances. Where this incipient change accidentally happened in the case of an established sigmatic future, the particular acts of speech deviated from social practice, which is to say that people were habituated to a certain balance between expression and inference in this area of grammar (without assuming that this balance was in anyway optimal, as discussed). They were accustomed to expressing the future, and unaccustomed to having to infer it. I have already discussed how this habit is not easily violable within a short period of time

It is a fact that people will follow social norms (with a certain smallish amount of leeway), given their desire to communicate efficiently within their society. Arguing for the blocking of sigma on the basis of habit, then, is considerably different from the hermeneutic or functional approach, for the above desire is the only mental entity we posit. We do not assign meaning to any act, or desirability to any distinction, and

---

inexplicables makes for a particularly good explanation.

[38] Although we do not have historical data to quantitatively assess the strength of the social norms in each case, we can nevertheless make such a claim because the very nature of an s-curve enforces a marked distinction between the plateau and lag sections.

merely proceed from the basic nature of speech to a discussion of how this nature constrains people's behaviour.

The situation in Greek can be contrasted with the history of Latin cases. We already know as a matter of historical analysis that speakers of Vulgar Latin towards the end of the Empire did not habitually distinguish the full range of classical Latin cases (and this was part of a larger trend of substituting periphrasis for inflection). Similarly, we know that there were certain sound changes (such as the merger of short high and long mid vowels) well under way by the same period. These two processes were entirely independent, as with sigma-deletion and the sigmatic future. In this case, though, the grammatical distinctions did not have much more momentum than the sound change, and the sound change could proceed without compromising the habit governing case distinction.

In that both sound and the form continue with the expected s-curve, this seems to require less explanation than the Greek example, When comparing Latin cases and the Greek future, then, we are not pressed to explain anything by hypothesising beyond the historical. In both cases, the histories of the distinctions and the histories of the sound changes are independent. We simply observe though historical analysis that the future in Greek at the relevant time was something more like point A (or in shorthand: had more momentum), while the cases in Latin at the relevant time were something more like point B (or in shorthand: had less momentum). There is no scientific reason why the future remained strong, and the Latin cases did not: these are merely historical facts. These patterns are naturally a matter of coincidence, and we thus account for the different processes in purely historical terms.

Though my phrasing for most of the above is taken from structuration theory (with the addition of "momentum" as a useful shorthand), the explanation I have given here is one from the invisible hand. Consider the following, more formal, expression of the explanation:

1) As regards the micro-processes:
    a) Internal considerations:
        i) speakers intend to communicate
    b) External considerations:

i) In the context of pre-Classical Greek of a certain era, a speaker is obliged to distinguish present and future, by virtue of the fact that he is speaking that lect.

ii) In the same context, a speaker has a comparative amount of freedom in deleting sigmas, by virtue of the fact that he is speaking that lect.

2) As regards the collectivity of speakers:

a) Speakers coming together to communicate will favour social norms in as far as these allow for easy communication between the parties

b) Favouring i) over ii) does not depart significantly from social norms; favouring ii) over i) departs significantly from social norms.

c) In communication, randomly having tried either of the above options initially (since speakers do not compute such things at the outset), speakers would settle on the former.

3) Without any speaker intending to retain a sigma in the future, the consequence of many people acting on these habits is the blocking of sigma-deletion in the future.

I think this is a reasonable explanation of a bit of morphophonology, or at least that it is a better explanation than is often given. In explaining this morphophonology, however, we do not explain the sound change itself: we simply take the rise of sigma-deletion or Vulgar Latin vowel merger as a given fact (and there is no reason I can think of for our wanting to do otherwise). However, since I do want to use this model of explanation to account for a sound change itself, I will have to look elsewhere: in the following chapter I will turn to umlaut.

To return to my main concern I should make some claims about Germanic that will be useful in the next chapter in identifying some relevant habits. I have discussed how Indo-European languages often lost PIE inflection. Nevertheless, it is the case that a core group of inflections has remained in Germanic. This includes the nom. ~ acc. distinction in pronouns (though it has become nom. ~ obl.), the present ~ past[39], and the singular ~ plural in nouns. Indeed, it may be possible to treat some of these

[39] "Past" here refers to "strong past", which is indirectly a continuation of the PIE perfect.

(as innovations from PIE) as defining features of Germanic. The development of a tense system is one such example.

There are exceptional nouns, though, that do not show a singular ~ plural distinction (*fish* and *water* in English: most of these are mass nouns). However, these exceptions do not constitute an organised declension. This grammatical distinction is a core one for Germanic, and it has always had great momentum from PIE to the present. It would thus be remarkable to find an entire noun declension (when English still had such things in any productive sense) without a singular ~ plural distinction, and we can safely claim that Germanic habitually distinguishes singular and plural in words such as *mouse*, *foot* or *goose*.

In this light, a few comments can be made about the history of athematic consonant stems, having a plural in *–i* < PGmc **-iz* < PIE * *-es* . Compare the PGmc nom. sing. **mu:s* and pl. **mu:siz* 'mouse' above. The *z* was eventually lost, leaving the *i* as the most significant marker of the plural in this case and in the accusative. The loss of this *i* (without the intervention of another process such as umlaut) would thus have involved the loss of distinction between singular and plural for these nouns in these cases, and we would have had the unusual situation of a Germanic declension with no number distinction in the highly salient nominative and accusative cases. Nevertheless, it is the case that this *i* was eventually lost. The sing. ~ pl. distinction was however not lost: umlaut occurred and the vowel alternation took on the role of signalling this distinction.

This seems to be more than coincidental, but any of the following statements would nevertheless be inadvisable: 'Germanic underwent umlaut to preserve this important distinction' or 'Germanic could lose the *i* because the umlauted vowels were sufficient to signal plural in these particular contexts'. Such functionalist claims seem to be invoking a *Sprachgeist*, or to be positing knowledge of speaker's mental states, unlike the historically verifiable claim that this distinction had great momentum in Germanic. Much work is still needed, though, in using this momentum to explain umlaut.

6) Explanations of Sound Change

Having laid out the various concepts needed for an explanation from the invisible hand, we can finally turn to an explanation of sound change. In 6.1 I will discuss a theory of tonogenesis as an example of such explanation, followed by an account of umlaut in 6.2. In 6.3, I will compare my discussion with Aitchison's snowballing.

6.1 Phonemic shift

In the previous chapter, two types of examples were discussed (the Greek future and Latin noun cases). Both involved a sound change entailing phonemic loss, and an independent grammatical distinction which is expressed by the affected phonemes. Further, both examples traced the historical momentum of these independent processes. In the case of the sigmatic future, the distinction had more momentum than an incipient sound change, and the sound change was blocked (remembering that these claims are based on conformity with social practice, and not the desire to keep distinctions). In the Latin cases, the distinction among oblique forms had little momentum, while the relevant sound change had greater, and the sound change proceeded as usual.

In neither case did speakers do anything new, as such: the explanation deals with processes already given, for it does not (and could not) explain why sigma deletion or the Vulgar Latin vowel merger arose in the first place. In order to explain a sound change itself (i.e. to say how a change may have arisen), we would have to explain why the speakers did (or were made to do) something new. We will thus have to depart from the conditions of the previous examples.

What is interesting from the point of view of the invisible hand, then, is when a distinction and a counter-acting sound change co-occur, both with a great amount of momentum. What would we expect to happen if the plateaux of conflicting s-curves coincide, when a distinction is firmly entrenched in usage, but a counter-acting sound change is too strong to be blocked? In the above examples, when two social practices conflicted, the weaker was not followed; in the just-mentioned case, however, it seems that both conflicting practices would have to have been followed, on the basis

that they had great momentum, and that speakers could not consciously depart from such practices.

Since the high momentum of the habitual distinction implies that it will be maintained[40], but since the expression associated with that content will be lost, we might suppose that the grammatical information will have to shift elsewhere. This shift would be something new, and would involve sound. The accidental flux of social habits, then, would force a specific response, but this response would be unintended. If the assumptions here are true (and the following suggests that they are), this will have explained a sound change by the invisible hand. Previously, I had vaguely suggested that the invisible hand should apply to those changes which dealt with information in its own right; now I can be more specific in suggesting that it should apply to those changes which involve a shift of information from one site to another as the result of blindly following habit.

It is possible to interpret Hombert, Ohala & Ewan's account of tonogenesis as an example of such a situation. They claim that

> [t]he development of contrastive tones on vowels because of the loss of a voicing distinction on obstruents in prevocalic position is probably the best-documented type of tonogenesis. When such a development occurs, a relatively lower tone develops on vowels following a previously voiced series, and a relatively higher tone is found after a previously voiceless (or voiceless aspirated) series. (Hombert *et al.*, 1979: 38)

Here, we must posit three independent processes. Firstly, as is argued throughout Hombert *et al.*, there is the above correspondence between voicing and pitch. Hombert *et al.* suggest various physical reasons why this might be the case, given the structure of the human vocal tract, but such claims are irrelevant here. Secondly, again as mentioned above, there is a loss of distinction between voiced and voiceless consonants. Further, however, we must suppose that social practice distinguishes certain relevant words. These words are distinguished on the basis of voicing at one stage, and on the basis of tone at another, and by redundancy through both voicing

[40] Recall from section 4.2 that a strong habitual distinction implies a regularity in expression: English's habitual distinction of the continuous aspect implies that speakers have to express it, and cannot merely allow it to be inferred. The loss of the expression, though the information might still be inferred, is a departure from the habit.

and tone at some intervening stage, but the distinction remains independently of these various means. The distinction is therefore a habit with great momentum.

An invisible hand explanation of this might run as follows. Let us divide the continuum into times A, B and C. At time A, we find phonemic distinction in voicing[41], and a corresponding (though not terribly salient) pitch difference. At time C, the pitch is distinctive, but voicing is not. Presumably, then, at an intermediate time B, the pitch and voicing were equally salient. The following might be a possible explanation sometime after A:

1. As regards micro-processes:
    a. Internal considerations
        i. Speakers desire to communicate through language
    b. External considerations
        i. With great regularity, speakers habitually distinguish certain words. At some time, this distinction happens to consist in a voicing contrast
        ii. With great regularity, speakers habitually pronounce vowels following voiced consonants with a lower tone than vowels following voiceless consonants
        iii. Sporadically, speakers fail to distinguish voiced from voiceless consonants. Following the s-curve, though, this practice becomes increasingly regular
2. As regards the collectivity of speakers:
    i. Speakers follow social practice by the very nature of linguistic communication.
    ii. Following i) does not involve a commitment to a particular form of distinction (though a change from one form to another should be gradual)
3. As regards the macro-processes:

bii) suggests a possible redundancy. Because of this redundancy, as biii) became increasingly regular (i.e. as the voicing distinction was

[41] I do not propose to go into what would have to be a lengthy detour discussing the status of the phoneme. For my purposes here, it would be sufficient to distinguish the case in English (where we

used with decreasing regularity), the pitch distinction became more salient, in compliance with bi). Eventually, the pitch distinction was the only feature distinguishing these words.

Again, no complex calculating ability is assigned to speakers through this claim. The following picture seems reasonable. There are independent processes in such a language: one losing voicing contrast, the other creating a small amount of pitch difference corresponding to voicing. At the beginning of the process (when the voicing-loss rule is incipient, or on the lag section of the s-curve), voicing may be lost sporadically in usage. Similarly, the pitch will be sporadically more marked in some usages than in others; no speaker will reproduce exactly the same pitch in each instance. Accidentally, then, a sporadic lack of voicing will sometimes occur with a slight difference in pitch, at other times with a more significant difference in pitch.

In that the latter instances keep closer to social practice than the former (in distinguishing the relevant words more successfully), and in that people blindly follow social practice in speech, a process of trial and error would be expected to favour the latter, for speakers are sometimes able to realise whether they communicated successfully. We would thus expect people to start producing significant distinction in pitch with habitual regularity, so that at time B, we find a salient distinction in pitch, and a salient (though unstable) contrast in voicing.

What we see here, then, is a shift in phonemic status. Initially, there was phonetic redundancy. The pitch of the vowels was not contrastive, however, while the voicing was. By the end of the process, the pitch had slowly become contrastive. This was the intention of no individual, but it arose through a collectivity following their own desires (in this case, the desire to communicate, entailing a commitment to the social practice of distinguishing certain words). It is thus an explanation from the invisible hand. It is an explanation in that if we can suppose that the premises are historically true, it is reasonable to assume that people acted in the manner described.

The examples from Greek and Latin dealt with morphophonology, starting the move towards an explanation of sound change. This example of tonogenesis continues that move, showing how low-level phonetic distinction might become phonemic. As such, it is an explanation of sound change, but it is possible to go even further. The

---

find a pitch difference corresponding to voicing, though it is not contrastive) from the case in Bantu or

low-level phonetic distinction here arises from some property of the vocal tract (and Homert *et al.* discuss what property this might be). The invisible hand thus merely explains how this low-level phonetic difference came to be phonemic, but not how the phonetic difference first came about. If the context discussed above is as constraining as I have suggested it is, then it may be possible to explain the phonetic change itself.

6.2 I-umlaut

I think it reasonable to assume that no speaker of a Germanic language ever intended to bring about umlaut. This is insufficient, however, for arguing that umlaut is a change only explicable by the invisible hand. No one actually intends to bring about lenition either, but lenition should not be explicable by the invisible hand. As discussed, lenition is a matter of expression, and not of the sign as a whole: it is not interpersonal in the way phenomena of the third kind must be, and the relevant actions lead directly to the consequences. I have laid out various introductory points above (mora loss, the habitual distinction of singular and plural, etc.) and it should be clear that none of these lead to umlaut in any direct fashion. If, however, I can show that people, reacting to these conditions, can reasonably be expected to bring about umlaut, then I will have explained umlaut as through the workings of the invisible hand.[42] As Keller discusses (1994: 72-73), I would not necessarily need to predict that people faced with these circumstances would necessarily bring about umlaut. Rather, I should suggest that, given our knowledge of how speakers behave, it would be reasonable for us to think that they behaved in this way.

Unlike the typical method of a functionalist or hermeneutic account, the only knowledge of speaker's mental states that I will suppose is that they desire to communicate linguistically, which leads them to have to follow social convention in using signs, by the very definition of a sign. In that this holds true in speech, regardless whether the particular act of speech occurred in the earlier middle ages or in the modern world, our expectations of speakers' behaviour in this regard is not bound by historical period.

Sino-Tibetan (where we find contrastive pitch).

[42] I will begin with an explanation of umlaut in plural nominative and accusative endings, before briefly showing how this argument can be extended to other instances.

The following table has already been quoted above (from Campbell, 1998: 217), though here I have added labels A-C to specify different time periods.

Proto-Germanic	**mu:s*	**mu:s-iz*	
Early Pre-English	*mu:s*	*mu:s-i*	A
Umlaut	-	*my:s-i*	B
Loss of *–i*	-	*my:s*	C

This process is highly idealised, however, and should be taken as a phonological and not historical outline, for it suggests that *i* was lost only after umlaut had become stable. Long before the *i* was regularly lost, it would have been lost sporadically. The period from A to C, then, should be characterised by the increasingly regular loss of a short /i/ in an ending. This can be taken as part of a larger process with great momentum: that speakers habitually express less information in endings as time increases. In the same period, a speaker would be obliged to distinguish singular and plural in all nouns, except for a limited set of mass nouns.

The plateaux of both processes coincide, though this only involves a conflict once both processes had gained significant momentum. Mora loss had reached its plateau by roughly 300 AD (or at least, it is possible to interpret Prokosch in this way) and this state of affairs continued until around 1300. Similarly, singular and plural had been distinguished since PIE, and this has continued until the present. In the period long before 800 AD, then, both habits could co-exist quite happily. However, in the centuries preceding 800, mora loss started causing the loss of a Germanic short /i/ in endings. Unlike the future in Greek, where a problematic sound change was incipient while a distinction was firmly entrenched, here we find two processes that only clash once they have both gathered significant momentum.

As discussed throughout chapter 4, in desiring to speak, people would have had no choice in following social practice. One such social practice is the habitual distinction of the class of words including the ancestors of *mouse*, *goose* and *book*. Another such practice is the habitual loss of a particle in an ending over time. Given the speakers' passive roles in relation to these practices, the speakers cannot be expected to identify the clash that would be caused by these processes, and to rationalise some way of avoiding it. Nevertheless, the practices must be followed. Germanic speakers of this era, then, are faced with the knotty problem of losing

phonemic information in one location, while wanting to hang on to a semantic distinction signified by that phonemic information.

In the case of this noun declension, it is the {I} that signals plural. Shifting this {I} leftwards would hypothetically solve speakers' problems, though. We can thus distinguish two parts in this explanation. Firstly, if we take the just-mentioned shift as a sporadic possibility, the invisible hand should explain how the shift came to be a regular habit in the context of early Germanic, with the consequence that an umlauted vowel eventually became the main marker of certain combinations of case and number. Secondly, we might explain how the {I} came to be shifted sporadically in the first place. The first part is similar to the discussion of tonogenesis above: it shows how a non-contrastive phonetic distinction passed through a stage of redundancy, and came to attain phonemic status. The second part contrasts with tonogenesis, on the basis that the argument for the shift there is articulatory, while here it is from the invisible hand.

Taking the possibility of a sporadic shift as assumed, the following discusses the first part of the explanation, so that a comparison with tonogenesis can be established before proceeding with the new claims. The discussion here is obviously more complex than the simple statement that we see vowel harmony, but this is only to be expected. In Finnish, we can take vowel harmony as a descriptive fact. With Germanic umlaut, on the other hand, we run into the actuation problem. The simple description of vowel harmony does not explain why Germanic of a specific era came to have a limited amount of vowel harmony, which lasted for a comparatively short period in its history.[43]

At an initial stage (labelled A above), we can imagine people sporadically failing to produce the final *–i*. Accidentally, this would sometimes occur with a shift of the {I}, and sometimes without. A word with this shift would be more likely to be distinguished as plural than a word without it, since it contains the information which indicates 'plural' for its class (an {I} particle). The act of using such a word would thus conform more with social practice than the act of using a word without the shift. In that the desire to speak necessitates that people will follow social practice quite closely, the form with the shifted {I} would come to be used habitually, through a process of trial and error. Using it habitually implies that it will eventually become

[43] Though we might claim that vowel harmony remains as a regular habit *in* Icelandic.

regular, and that its use is not reasoned out. The increased regularity of this usage, then, does not need to be argued for in any way except from non-rational need to follow social practice in general. This is typical of the sense of social practice given by structuration theory, and the habit is accounted for by the invisible hand, for no one actually intended to make the process regular.

Therefore, given the sporadic shifting of {I} in isolated instances, the above context will bring about regular umlaut. However, this sporadic shifting needs to be examined in itself. That is, it may be accidental or it may be caused by some property of the speakers (as is the case with the correlation between pitch and voicing). Alternatively, however, it may result from the working of the invisible hand: it is possible that the above context will lead speakers spontaneously to shift the particle in the first place, and then cause this sporadic shifting to become a regular habit. I do not think there is an easy or clear answer to this question, for it may be the case that various factors contributed to the causing of umlaut. Nevertheless, the following suggests that the invisible hand is at least one of these factors. First, I will examine a few independent accounts, which might explain the spontaneous initial shift as an accident, or a result of the speakers' brains. If these latter accounts succeed, then we would have a causal explanation of the initial shift, and an invisible hand explanation of how this shift becomes regular umlaut. In the alternative account, we will have an explanation from the invisible hand for both the initial shift, and the habituation of this shift.

The null-hypothesis might be that the original shift is entirely accidental, but that once this move was found to be possible, the working of the invisible hand caused this accident to become habitual. As discussed above, this *might* still lead to regular umlaut in the specified context, but it may be unsatisfactory for other reasons. Among these is the fact that umlaut looks far from accidental. We are unsure as to exactly when umlaut came to be. In Old English, the first evidence for it is from the sixth and seventh centuries (Lass, 1994: 61-62), but this does not necessarily imply that it only occurred then. Either umlaut began sporadically during the period of Proto-Germanic, though only becoming visible orthographically centuries later, or it occurred after the split up of the Germanic languages, in a period later than Gothic.

Given the current state of knowledge, we cannot be sure which of these pictures is closest to the truth (and it is possible that we never will). If the initial shifting was accidental, though, it would be compatible with the first option, but not with the

second. That is, if the shift was accidental, then it would be surprising that various languages underwent the same processes independently. If, however, the shift was caused by some property of the language or languages undergoing umlaut (or some property of the speakers of these languages), then the claim would be compatible with either scenario. Since we would not be justified in committing ourselves absolutely to either scenario, but since claiming that the shift was accidental does commit us to the first option, it seems be more in keeping with our ignorance to claim that it is possible that the shift may have been caused, rather than that it was accidental.

Such a causal account is given in Berg (1998: § 4.11). Arguing from principles of language processing, he says that

> [s]peech is a sequential activity which happens in real time. At any particular moment, some stretches of an utterance have already been completely articulated while others have yet to be produced. Speakers are well advised to allocate their attentional resources more to the future than to the past because planned units have to be outputted while used elements are no longer needed ... So, in response to the requirements of speech, speakers' primary focus will normally be upon upcoming events and their secondary focus upon past events ... In general terms, it can be said that imminent material is more highly activated than that which has been already executed ... Given this greater activational strength of not-yet-produced units relative to already-produced ones, it can be predicted that regressive assimilation will be more common than the progressive type. (Berg, 1998: 113)

On this basis, we might argue that speakers of Germanic of a certain era, having similar processing techniques, would sometimes anticipate the {I} because of its increased activation. A few such anticipations would then become habitual by the same process as outlined for tonogenesis. However, this is not crucially different from the null hypothesis. Berg's account of anticipation rests largely on data from speech errors. Such errors, however, are entirely accidental. Even though considerations from processing might underlie the shift, its actuation will still be accidental.

Furthermore, this interpretation is cannot deal with the similarities between umlaut and tonogenesis. Although the former passes information backwards, and the latter passes it forwards, both can be interpreted as being reactions to the same environment. Berg's explanation is limited to instances of the former. Again, it is

therefore at least possible that umlaut is something more than the result of increased activation.

Given that umlaut may have happened in all Germanic languages of a certain era, after they had split up, perhaps we should look to the commonalities among them for a possible cause. One such commonality is that they shared the habitual context outlined above: they all still underwent mora loss, for it had great momentum before the split, and this momentum continued centuries after they had become independent. Likewise, they all distinguished singular and plural in this declension.

I have suggested that this same context (a regular distinction being compromised by a regular sound change, both with great momentum) affects the low-level phonetic process in tonogenesis, causing pitch to become more salient. Given the similarities, is it maybe possible to claim that the same context *caused* the {I} to shift in the first place? That is, by effecting a low-level phonetic process in the stem vowel, and then causing it to become habitual? Since the phonemic information was being lost in a final syllable, though this information signified an important distinction, and since speakers could not do anything about this state of affairs, can we maybe say that speakers reasonably responded (or can be cogently said to respond) in this way?

If we take the notion of social practice seriously, we are committed to the idea that

> a great deal of linguistic structure does not respond to 'speakers' needs' in any intelligible way; it's simply there, and using the language is playing by the rules … Structure is given as a historical fact; the speaker's job is to get it to do (in utterances) what he wants to do, using the available machinery. (Lass, 1997: 368-369)

If this is true (and I think Lass, Deumert and Keller have argued convincingly that it is), then we have little reason to suppose that speakers have any more control over the structures of their language when practices are in conflict than they do when practices are either isolated or complementary. Therefore, if one practice deletes a certain unit of expression which bears grammatical information, and another practice requires the same grammatical information to be expressed, a speaker should not be able to respond rationally to resolve the situation; social practices are not rational phenomena. This distinction and this sound change are social habit in the same sense

that a semiotic ground is. A speaker is no more capable of working against social habit than he is of deciding that /dɒg/ should signify 'blunderbuss'.

If a speaker must express certain information, but cannot express it in its original location, then it will simply have to be expressed elsewhere. If we take the original position within a word to be B, then elsewhere in the word will be ~B. We can claim that such information as the plural here can logically be expressed in B or ~B, but it must be expressed somewhere (since it is clear from historical information that it had momentum). Since it cannot be expressed in B, it follows that it will have to be expressed in ~B.

This has already been seen in the case of Germanic case endings. Since the case endings were continually less well expressed, the case information ended up being expressed in the thematic vowel. Passing information leftward is thus not at all atypical in Germanic. In the history of French, on the other hand, the case distinctions themselves were weak. Old French thus did not have the same context as Germanic, and consequently no reason to pass information leftwards.

In the case of umlaut, the {I} is what contains the information 'plural'. This {I} was being lost in its original position, but the information 'plural' had to be retained (in order to follow social practice, and not for any functional reason). Since the expression could not be retained in its original site, it was moved to another position in the word. However, because it was only dropped sporadically at first, there was a stage (time B in the above schema) when the information was redundant, before the process reached its endpoint (time C). Notice the distinction here between information and purely linguistic processes. The loss of an *i*, the fronting of a back vowel, and the raising of a front vowel are three different processes, in structural terms. It is only when we consider this to be the preservation of a certain piece of information that we can view it as a single process.

This explanation thus has two important aims. Firstly, it suggests how people may have come to spread the {I} in the first instance; secondly, it shows how this sporadic behaviour became a matter of regular social habit in its own right. Umlaut, then, can be seen as the cumulative consequence of a non-rational response on the part of speakers to their linguistic context. Information that had to be expressed was moved from a position rapidly losing salience to a position with a great deal of salience (an initial, stressed position), as a result of a commitment to social norms.

If it is the case that the aforementioned context caused umlaut, then we might expect a certain difference between umlaut and Labovian change. Labovian change spreads through a social network. If umlaut consists in a response to a certain environment, however, it would be reasonable to suppose that various speakers in such a context would respond in this way, without having had experience of each other. Put more simply: umlaut would not *have to* spread through a social network (though obviously this does not imply that it cannot do so in addition). In such a case, then, the appearance of umlaut in independent languages, not linked by a strong social network, is unsurprising.

My discussion may be criticised on the basis that the nom. and acc. plurals of the athematic declension are merely one instance of umlaut (and fairly minor ones, at that). Nevertheless, it is this form of umlaut that is most commonly quoted in textbooks. Lass (1994: 61) and Campbell (1998: 22-23) are representative examples here: they both give examples of umlaut in nominal paradigms before expanding the discussion to include the other data. Lehmann (1962: 165) also presents significantly more examples of this type of umlaut than of any other. It is thus by no means unusual that I should focus on this sort of umlaut.

My purpose here has not been to present a full history of umlaut, but rather to show that it is possible to explain sound change from the invisible hand. Given the constraints of length imposed on this paper, I have chosen to argue for this point on the basis of a limited but detailed example, rather than through a more general but correspondingly less detailed picture. Since it is the detail of the habitual context that is important here (such as mora loss), I think this bias is condonable.

Nevertheless, it is possible to suggest how other instances of umlaut are explicable in the same way. Lass claims that 'many OE inflexional and derivational suffixes historically contained */i, j/; IU [*i*-umlaut] therefore produced an important set of morphophonemic alternations' (Lass, 1994: 70). This suggests that speakers habitually expressed the morphological information contained by these sounds (examples of which I have listed in ch. 3). We can also identify an independent process deleting these instances of *j*. All such instances of *i*-umlaut are thus similar to my example above, in that they are situated between conflicting habits: one expressing grammatical information (irrespective of whether it is inflectional or derivational), the other deleting the expression that carries that information (regardless whether it is an *i* or a *j*).

A possible criticism is that this extension of the argument is unwarranted, on the basis that the plural is more salient or more important than, for example, the marker of a causative conjugation, and therefore more likely to be preserved. This criticism, however, makes functionalist assumptions that are incompatible with structuration. No distinction is necessary. If any distinction is habitually made, it is only because speakers make such a distinction in their speech. Conversely, if speakers made certain distinctions, then this is sufficient evidence for the above argument.

6.3 Habitual blindness and snowballing

My claims about the flux of habits causing change can be compared with snowballing (Aitchison, 1987: § 5) for various reasons. Firstly, this will tease out some implications of the blind following of habit (suggesting briefly just how blind it in fact is). Secondly, if snowballing can be shown to result from the following of habit and is explicable by the invisible hand, then historical linguistics needs to make recourse to fewer independent types of change in explaining its data (even if it is just one fewer, for the moment). Thirdly, it will provide an initial suggestion concerning further ways in which the invisible hand can affect sound change.

I have mentioned repeatedly that speakers blindly follow habit (echoing Lass, 1997: ch. 7), and was thus pleased to find that Wittgenstein himself wrote: 'when I obey a rule, I do not choose. I obey the rule *blindly*' (Wittgenstein, 1953;quoted in Lähteenmäki, 2003: 55). It would be sensible, then, to tease out various implications of this claim for explaining language change, especially since Lähteenmäki (2003: 58) explicitly connects this aspect of Wittgenstein's thought with Keller's claims, and with emergentism.

He discusses how,

> during his mature period, Wittgenstein came to realise that knowing a rule cannot be identified with having a correct mental representation of a rule or knowing a symbolic expression for it. He rejects the idea of rules as formula-like reified objects and conceives them as routine-like skills to act in a normatively correct way in various types of situations. ... In this specific sense of 'rule', as controversial as it may sound, rules can be followed without *explicitly* knowing what they are. Thus, although all rules are

potentially explicable, they do not *necessarily* have to be formulated. … Although the distinction between genuinely following a rule and coincidental acts in accord with a rule seems conceptually important, it becomes problematic when considered from the point of view of the criteria on which our judgements about somebody's following a rule are based. (Lähteenmäki, 2003: 55-56; emphasis Lähteenmäki's)

Given the problems of knowing other minds and the habitual nature of certain rules, speakers' internal methods of producing certain results need not be crucial in discussing how various speakers come to produce the same results. Because they are engaging in a normative set of actions in a community, it is sometimes sufficient to claim that speakers reproduce social behaviour, without inquiring into the speaker-internal processes that lead to this behaviour. Naturally, this all depends on what one's focus in investigating language is. It may be useful to hypothesise about the mental processes that produce linguistic results (as the mainstream in linguistics intends to do), but it does not follow that all linguistic explanation needs to be built on such analysis.

Recall, for example, Keller's discussion of the traffic jam. For the purposes of explaining the jam (from a certain perspective), it really does not matter what driver A's mental processes were in braking were, for other drivers respond on the basis of his actions, and due to the reasons for those actions. It is possible to treat some linguistic structures in the same way: we can explain phenomena of the third kind on the basis of accumulative actions, regardless of how speakers' internal processes produce such actions.

For those aspects of language that are primarily habitual, the degree to which speakers are likely to do something is directly related to how often or how regularly it is done around them. As discussed, this can be phrased as follows. The oftener or more regularly something is expressed in their environment, the stronger the corresponding habit is for such speakers. The stronger the habit, the more it in turn constrains their actions. If something independently increases the usage associated with a certain habit, it follows that the habit will consequently become stronger, and it will in turn effect more regular usage. The fact of the usage is what yields this increase in strength, and not the processes whereby the usage came about.

In a linguistic conspiracy, various processes might yield the same or similar results on the surface (Kisseberth, 1970; cited in Aitchison, 1987: 20). If it is true that

speakers do certain things because the rest of their community does them, it follows that speakers need not distinguish these results as arising from different processes. If they do not distinguish the causes of such actions, then the set of actions must be associated indiscriminately with the various habits. Each habit will therefore be linked with a greater degree of apparent usage than it would if it were isolated. As discussed, the greater apparent usage entails an increase in the strength of the habit.

The chance similarity of results, then, will be a factor in increasing the strengths of the relevant habits. This last claim is compatible with the essentials of snowballing:

> If one assumes that language maximizes structures that happen to be the end of a number of different processes, then it will in all probability get encased in a snowball from which there is unlikely to be an escape. (Aitchison, 1987: 26)

Further, Aitchison claims that 'that once the snowball started to roll, it gathered momentum and became unstoppable' (1987: 26), where her use of "momentum" seems to have similar implications to mine. It should not be forgotten, however, that "momentum" as I use it is only ever the product of historical analysis. As a result, I am currently doubtful about Aitchison's claim that snowballs can be used to predict language change. It is nevertheless interesting that Aitchison claims this process will give 'the impression that there is some preordained blueprint' (1987: 26), since this is remarkably similar to the motivation for the use of the phrase "invisible hand".

What habit adds to this picture is a basis for making the assumption in the above quotation: language maximises such structures *because* people follow habit blindly. This implies that they can produce usage in line with what they see around them, regardless of the other-mind processes underlying that usage. In the absence of such a framework, Aitchison supposes that 'ultimately, snowballs will need to be related to properties of the human mind' (1987: 28), whereas my discussion above has based the discussion on properties of communication.

The bulk of this thesis has examined conflicting habits, when one has low and the other high momentum, or when they both have high momentum. Snowballing, on the other hand, deals with habits that are harmonious, rather than conflicting. Irrespective of the predictive values of snowballing, we can thus use it to explain past changes. Having observed an increase in the momentum of a particular process, and having identified that the historical environment of this increase contained various

independent processes producing similar results, we might suppose that speakers' habitual reactions to this context brought about the increase in momentum. The increased momentum is the consequence of a collectivity of speakers acting in ways not motivated by this consequence, and is thus a matter of the invisible hand.

In the case of umlaut, for example, although mora loss and *j* deletion are two independent processes, their historical contiguity would have promoted the spread of {I} more than would have been the case if either deletion were to have happened on its own. In other words, the accidental co-occurrence of *j* deletion and mora loss may be said to increase *the usage* of a leftwards spread of {I}. The increased usage would in turn reinforce the strength of the umlauting habit, which in turn would lead to more tokens with umlaut. Similarly, the leftward movement of Germanic case content (in that cases were sometimes expressed in what was previously only a thematic vowel) is independent of umlaut, but the existence of the former as a leftward shifting habit may conceivably have given impetus to the initial sporadic spreading of {I}.

The following offers some suggestions about how invisible-hand snowballing might be used to explain the tendency to palatalise in Russian. Since my aim here has been to show that it is possible to explain sound change from the invisible hand, and not to present a complete discussion of all the ways in which it is possible, the following will have to be a brief outline rather than a watertight discussion (and the explanation is correspondingly weak). It will thus focus on distinguishing this method of explanation from more naturalist forms of argumentation, to highlight what kind of an explanation this is, rather than on showing exactly how it works by justifying each step conclusively.

Entwistle & Morrison (1959: §§ 36-38) propose three independent processes fronting velars in Russian. They call the first '*satem* palatalisation' (§ 36), after the well-known distinction between the *centum* and *satem* languages within Indo-European. This will not concern me further, however. The second (the 'first Slavic palatalisation', § 37) might be stated as

$$\text{PIE} \begin{Bmatrix} *k \\ *g \end{Bmatrix} \rightarrow \text{PSl} \begin{Bmatrix} *t\int \\ *d\text{ʒ} \end{Bmatrix} \; / \; \_\_\_\_ e, i$$

The third (the 'second Slavic palatalisation', § 38) as

$$\text{PSl}\begin{Bmatrix}*k\\*g\\*x\end{Bmatrix} \rightarrow \text{Ru}\begin{Bmatrix}ts\\(d)z\\s\end{Bmatrix} \Big/ \begin{matrix}\text{i) i, } \begin{Bmatrix}t, \tilde{e}\end{Bmatrix} \_\_\_\_\\ \text{ii) } \_\_\_\_ \begin{Bmatrix}e, i\end{Bmatrix}\end{matrix}$$

Now, palatalisation in the environment of a high front vowel is a fairly common occurrence. It is nevertheless inappropriate to claim that the high front vowel causes the palatalisation. The English pair *shirt* and *skirt*, different languages' reflexes of the same original form, shows that such an environment is not a sufficient cause for palatalisation.[44] At most, the co-articulation of high front vowels and velars presents an option that a language might follow, though it is by no means the case that any given language is compelled to follow this option. In other words, the articulatory properties of certain sounds may well give rise to sporadic fronting, but this sporadic fronting need never become a regular habit.

All else being equal, let us suppose that it is historically accidental whether a language ends up habituating such sporadic fronting. This is not a stochastic claim, supposing that any given language has *x*% chance of palatalising in such an environment. Rather, it simply means that a claim such as 'NGmc did not palatalise where OE did' is similar to 'English distinguishes punctual and continuous aspects of the present, while German does not', in that neither admit functional notions.

Snowballing comes in when it is not the case that all else is equal. Taking the first Slavic palatalisation as a *fait accompli*, we can hypothesise about speakers' behaviour at the time of the second palatalisation. At such a time, speakers would regularly produce palatal affricates that resulted from the first process. At the same time, they may accidentally front their velars in other environments (and this fronting would be sporadic). Both first palatalisation and this accidental behaviour would produce similar fronted sounds in the environment of a non-low front vowel.

Lähteenmäki (2003) has stated that speakers merely need to reproduce behaviour in line with the usage surrounding them, and that accidentally similar behaviour need not be distinguished as such. I have discussed how the claims of Aitchison (1987) follow from such a point, for she shows that the chance similarities resulting from

[44] The *sk-* form is a NGmc loan; the *sh-* is natively English. The role of the high vowel in this particular example is reduced even further, since English has a palatal corresponding to a NGmc velar even before back vowels, cf, OE *scofl* and OSw *skofl* 'shovel' (Lass, 1994: 58-59).

independent processes bring about an increase in the momentum of those processes, in creating a snowball. If these fronting actions are habitual and similar, then it might be claimed that acts resulting from the first palatalisation contributed to the strength of the second palatalisation, through speakers' indiscriminate compliance with social practice.

It still needs to be established, however, whether the above actions in Russian can be construed as habitual or as similar behaviour in such a specific sense. Naturally, even if such a thing can be established, it would certainly take longer to do so than is possible here. An integrated account of habits would have to be discussed, making reference to neurology and phonology, and empirical study would have to show that such an account sensibly constrains the levels of abstractness that we can attribute to speakers in their following habit. It may turn out, for example, that the fronting of velars in the two Slavic palatalisations are only abstractly similar[45], and that speakers would therefore not make a connection between the two kinds of usage.

Nevertheless, given that this is merely an initial suggestion, it is worthwhile to point out at least the possibility that the two processes in Russian might be considered similar (from a habitually-blind speaker's point of view, that is). If this is the case, then we would have reason to suppose that the accidental flux of habits has lead to harmony between them, which consequently increases momentum in the manner described. In other words, the fact that the language has already followed the option presented by palatalisation once before might conceivably have an effect on whether it follows the option again at a later time, by virtue of the similar results produced by the processes. If it does, then we can interpret this as a snowball.

The various sections of this chapter, then, have ranged between explaining a change by the invisible hand (as in umlaut), or merely showing that it might be possible to explain certain kinds of change in this way (as in a tendency to palatalise). In all such cases, it is the accidental flux of habits that prompts a reaction from speakers, based on their need to produce tokens of social practice when communicating linguistically. The consequences of their actions are not factors motivating their action, and the processes are thus explicable by the invisible hand.

[45] One being alveolar and the other palatal, they would share the feature [coronal], or the element {R}.

7) Concluding Remarks

The central aim of this thesis has been to extend Keller's model, so that it can include sound change. A series of attempts along this line has been made, tackling phenomena ranging from morphophonology right down to phonetics. Hopefully, this has allowed Keller's model to provide a more general account of how micro-variation becomes macro-structure.

I have left several questions open for further research (since I have had to present a complex matter in a comparatively short space) and can make further suggestions along these lines. In discussing conflicting habits, it was usually the case that one of the habits involved a grammatical distinction, whether inflectional, derivational or periphrastic. Many of the changes above followed from the need to preserve such a distinction when its original site could no longer contain it.

It would be interesting, therefore, to examine the applicability of this model to non-grammatical forms of distinction. It is possible to describe phonemes as having a primarily contrastive status: they distinguish different words. It might thus be the case that a similar mode of explanation could be used in discussing the preservation of phonemic distinction when an independent habit threatens this distinction. My interpretation of tonogenesis took it as assumed that this move was possible.

A suitable starting point for such a discussion might be the data from Andersen's (1973) abductive change. Briefly, since co-articulation is not a sufficient cause of change, it might be argued that the Latin change *du-* → *b-*[46] involves the preservation of an {U} that could no longer be expressed as a secondary articulation. If it is the case that Andersen's data is explicable in terms of the invisible hand, then the types of explanation proposed in historical linguistics can be reduced even further in number. Similarly, colloquial Tibetan shows a process remarkably similar to *i*-umlaut, in that vowels underwent exactly the same change (*u* → *y*, *o* → *ø*, *a* → *æ*). However, they did so in the context of a phonemic loss at the end of the syllable. A fuller discussion might thus be able to show how the habit associated with such distinction is comparable with habitual expression of the plural in Germanic.

It was mentioned above that such an explanation serves us through diagnosis, and it thus important to outline the kinds of diagnosis I have made. For example,

[46] This results, for example, in the Classical pair *duo* 'two' and *bis* < *duis* 'twice' (Andersen, 1973).

explaining umlaut necessitated a clearer understanding of mora loss. In such a case, diagnosis provides a clearer piece of history than we would otherwise have recourse to. In other instances, such explanations might clarify the role of habit in change, and thus provide additional support for a concept that is widely used within the discipline. Finally, if abductive change is eventually shown to be reducible to habit (as snowballing is), then a bit of methodological housework would have been done, and the tools of the discipline would become correspondingly simpler. Diagnosis is thus important in helping historical linguistics understand its object (in providing a clearer history) and in suggesting how it should understand its object (in providing suggestions for methodology).

Bibliography

Adamska-Sałaciak, A. 1991. Language Change as a Phenomenon of the Third Kind. *Folia Linguistica Historica* XII (1-2).159-80.

Aitchison, J. 1987. The Language Lifegame. In W. Koopman *et al.*, 1987: 11-32.

Andersen, H. 1973. Abductive and Deductive Change. *Language* 49 (3).765-93.

Anttila, R. 1992. Review of *Sprachwandel. Von der unsichtbaren Hand in der Sprache*, by Rudi Keller. *Studies in Language* 16.213-18.

Backhurst, D. & Shanker, S.G. 2001. *Jerome Bruner: language, culture, self.* London: Sage.

Berg, T. 1998. *Linguistic Structure and Change: an explanation from language Processing.* Oxford University Press.

Borges, J. L. 1979. *The Book of Sand.* London: Penguin

Campbell, L. 1998. *Historical Linguistics: an introduction.* Edinburgh University Press.

Crystal, D. 2003. *A Dictionary of Linguistics and Phonetics.* Oxford University Press.

Deumert, A. 2003. Bringing Speakers Back In? Epistemological reflections on speaker-oriented explanations of language change. *Language Sciences* 25.15-76.

Eco, U. 1976. *A Theory of Semiotics.* Indiana University Press.

1984. *Semiotics and the Philosophy of Language.* London: Macmillan.

Entwistle, W. J. & Morison, W. A. 1959. *Russian and the Slavonic Languages.* London: Faber and Faber.

Ewen, C. J. & van der Hulst, H. 2001. *The Phonological Structure of Words.* Cambridge University Press.

Guthrie, W. K. C. 1978. *A History of Greek Philosophy.*5 vols. Cambridge University Press.

Harris, J. 1997. Licensing Inheritance: an integrated theory of neutralisation. *Phonology* 14.315-370.

Hombert, J.-M., Ohala, J. J. & Ewan, W. G. 1979. Phonetic Explanations for the Development of Tones. *Language* 55 (1).37-58.

Keller, R. 1994: *On Language Change: the invisible hand in language.* London: Routledge.

Koopman, W., van der Leek, F., Fischer, O. & Eaton, R. (eds.) 1987. *Explanation and Linguistic Change.* Amsterdam: John Benjamins.

Lähteenmäki, M. 2003. On rules and rule-following: obeying rules blindly. *Language & Communication* 23.45-61.

Lass, R. 1980. *On explaining language change*. Cambridge University Press.

1994. *Old English: a historical linguistic companion*. Cambridge University Press.

1997. *Historical linguistics and language change*. Cambridge University Press.

Lehmann, W. P. 1962. *Historical Linguistics: an introduction*. New York: Holt, Rinehart, and Winston.

Mesthrie, R., Swann, J., Deumert, A., Leap, W. L. (eds.) 1999. *Introducing Sociolinguistics*. Edinburgh University Press.

Nyman, M. 1994. Language Change and the 'Invisible Hand'. *Diachronica* XI (2).231-58.

Prokosch, E. 1939. *A Comparative Germanic Grammar*. Philadelphia: Linguistic Society of America.

de Saussure, F. 1922. *Cours de linguistique générale*. Paris: Payot.

Taylor, J. R. 1995 (2nd ed.). *Linguistic Categorisation: prototypes in linguistic theory*. Oxford University Press.

Taylor, T. J. 2001. Bruner & Condillac on Learning to Think. In D. Backhurst & S.G. Shanker, 2001: 71-87.

Voyles, J. B. 1992. *Early Germanic Grammar: pre-, proto-, and post-Germanic languages*. San Diego: Academic Press.

Wright, J. 1924. *Grammar of the Gothic Language*. Oxford University Press.