How to live naturally and not be bothered by economy

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/127902 since

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)
Markedness plays a central role within Natural Morphology and, more generally, in any functionalist approach to language. This is not intended to deny that other forces, more or less conflicting or competing with markedness, may also play a role in shaping a natural language. However, the main concern of this paper is that identifying naturalness, namely “what promotes the true nature of a thing”, amounts to grasping the teleology of (any module of) a language grammar. Concepts such as frequency or economy do not provide by themselves any deep insight into the essence of language, unless they are taken in the right perspective of being in a way symptoms of naturalness. In particular, economy must be related to markedness reduction in order to capture its role within the architecture of grammar.

“Naturalness = what promotes and completes the true nature of a thing. This teleological or philosophical concept, which is closely bound up with the concept of markedness, must be contrasted with the quite different statistical concept of normality. (Pollution may be statistically normal, but it is not natural.)” (Bailey 1996: 374).

1. Introduction

Natural Morphology (= NM) views markedness as a basic ingredient for any functionalist approach to language. This is in a way the sense of the term ‘natural’, the counterpart of marked, developed in every possible respect within Natural Linguistics. Thus, Natural Phonology treats sounds and sound alternations as strictly regulated by the phonetic (including perceptual) endowment of human beings, leaving only a limited space for more “abstract” concepts such as phonological rules to operate. As pointed out by Donegan (2002: 59), “[i]n the generativist view, phonetic interpretation occurs only after the categorical phonological substitutions, which are also viewed as governed by language-specific phonological rules. In the naturalist view, some substitutions may indeed be governed by language-specific rules, but such rules are part of lexicon or
grammar rather than of phonology”. For Natural Syntax, principles like Natural Serialization have been proposed (cf. Vennemann 1974), although much more needs to be done in this area.

The latter caveat, which is often added at the beginning of review papers on Natural Linguistics (e.g., cf. Dressler 1990: 87), has now become less pressing, given that basic ideas common to Natural Linguists are widespread among many functionalist approaches which are currently being developed. Or, to put it in more emphatic terms, many functionalist approaches (to syntax, and in general, e.g., Givón 1991) can be seen as variazioni su tema with respect to the topics discussed within Naturalist circles, and vice versa. Similar considerations hold also true for Text Linguistics (cf. Dressler 1996). At any rate, NM undoubtedly represents the most significant achievement of Natural Linguistics, let alone the one which has been best worked out. Therefore, most exemplifications will come from there, even though examples from other fields will also be mentioned.

In this paper I will discuss the central role played by markedness within NM and, more generally, in any functionalist approach to language. This is not intended to deny that other forces, more or less conflicting or competing with markedness, may also play a role in shaping a natural language. However, the main concern of this paper is that identifying naturalness, namely “what promotes the nature of something” (Bailey 1996: 91), amounts to grasping the teleology of (any module of) a language grammar. Furthermore, usage-based concepts such as frequency or economy do not provide by themselves any deep insight into the essence of language, unless they are taken in the right perspective of being in a way symptoms of naturalness.

2. Markedness and the iconic dimension of morphology

One of the basic tenets of NM is the concept of constructional iconism. The latter can be considered as the foundational basis of NM, and incorporates a version of the Jakobsonian theory of markedness summarized by the ‘more meaning – more form’ principle and elaborated by means of C. S. Peirce’s semiotics (cf. Jakobson 1965). Jakobson’s view of markedness has changed over time, but it has always remained faithful to a substantive principle that goes beyond the limits of classical structuralism and is a candidate for being universal: in natural languages cognitive complexity is mirrored by means of formal (i.e., signal-specific) complexity. In this way, Jakobson lays down a semiotic foundation of language as an aspect (maybe the most relevant one) of the human beings’ ability as sign-builders. Jakobson (1962) explicitly chooses an anti-Saussurean point of view, which refers to what has been called the substantive body of language as opposed to the formalistic “Aristotelian” view of the Swiss linguist (cf. Simone 1990, 1995). This concrete, substantialist position advocates as an ideal ancestor Platon’s Cratylus, in which a naturalistic view of linguistic signs is supported.\textsuperscript{1}
Therefore, markedness intended on a universal basis and naturalness (in the sense of ‘naturalistic’) turn out to converge.

This is in essence Stampe’s (1973) main criticism against any abstract conception of markedness like the one developed in Chomsky & Halle (1968): markedness cannot be defined by means of abstract stipulations, but should be observed in a naturalistic context – for instance, in the behavior of people under the effects of alcohol. This famous example shows quite clearly the limits of classical generative phonology: it is rather implausible to assume that the phonology of drunks becomes more complex because of a higher number of rules of deletion, lenition, etc., than the phonology of an ideal speaker. The paradox is evident: under less controlled conditions (which should point to a simplified grammar), classical generative phonology needs to assume a higher number of mental operations, i.e. rules, at work. Interestingly, Chomsky & Halle’s (1968) conception also reflects the influence of Jakobson’s markedness. Recently, the debt paid to Jakobson has been explicitly recognized by Chomsky (2005: 6), who mentions that in the early Fifties in search of “principles of data analysis that might be used in language acquisition and other domains, [...] the primitive step of analysis of linguistic experience would be feature-based phonetic analysis, along lines described by Roman Jakobson and his associates”, albeit that method has now turned out to be “illusory”.

In spite of the apparent paradox of being a source of inspiration both for formalist and functionalist approaches to language, it must be stressed that Jakobson conceived markedness in very concrete terms as related to the way in which the speech signal is coded, let alone the way in which the world is perceived. Clearly, this has a number of different implementations depending on the linguistic level at which the coding process takes place. Thus, at a phonological level markedness is directly connected with the way sounds are produced or perceived; at a morphological level markedness is related to the way in which meaning chunks are coded into signs; at a syntactic level markedness has to do with the way information chunks (i.e., sign strings) are linearized or grouped.

There is of course an asymmetry between the sound level and the levels concerning meaning, i.e. morphology and syntax. The crucial difference is evidently given on the one hand by the notion of sign, and on the other by the strictly phonetic motivation underlying phonological markedness. Therefore, the other levels entirely lie within the realm of semiotics, whereas phonology is only indirectly connected with it. This does not exclude further connections of a semiotic kind between phonology and other components or modules of language. For instance, Jakobson (1965: 352) suggests that the range of phonemes involved in building inflectional morphemes is quite limited, and usually includes the most frequent ones. However, this kind of connection is external, as it were, to the essence, i.e. the nature, of phonology.
Jakobson’s notion of markedness is explicitly adopted by NM (cf. Wurzel 1992, 1998, Dressler 1989, 2002), and equated with naturalness/iconicity. More marked means less natural and less iconic. At any rate, this is only one possible source of markedness, which is universal since it affects the way the speech signal is structured, as laid down by Jakobson. Let us abstract away from other possible sources of markedness for a while, and concentrate on this universal level.

The extent to which a sign is good on a scale of markedness – naturalness is crucial in order to evaluate its possibility of being expanded as a model throughout the lexicon, i.e. of being productive. In this sense, let us consider the synoptic table offered by Bauer (2001: 60), in which all possible factors influencing productivity are summarized:

As the author observes, “there are extremely complex chains of causation involved here”, among which “naturalness itself is a function of a large number of factors, including transparency, and frequency is a result of naturalness” (Bauer 2001: 60). The basic idea behind NM is that “naturalness increases productivity because only if a morphological process is maximally natural is it maximally analysable and maximally computable. That is, the more natural a morphological
process is, the more likely it is that forms using it will be readily understood and will be produced with ease by speakers. Note that this implies a gradient view of productivity” (Bauer 2001: 60). As observed above, this assumption directly descends from Jakobson’s semiotic view of complex signs.

One advantage of this approach is that it can be empirically falsified. In fact, as shown in Gaeta (2002a, 2005a) on the basis of a detailed analysis of a significant portion of Italian word formation, namely action nouns, only some of the natural preferences relating naturalness and productivity (cf. Dressler 1987) are empirically supported.

Thus, productivity and naturalness seem to correlate directly: those word formation devices producing most natural derivatives (to be evaluated along a language-specific naturalness scale, we will come to this point back later) are more productive than the ones generating less natural derivatives. However, the correlation only holds in terms of type frequency, namely relating to the occurrence and/or increase of types. Much less so for the token number. As for the latter, the opposite holds true: a lower degree of productivity is found for derivatives showing a higher token frequency. The latter also include those derivatives which are evaluated as less natural in terms of the naturalness scale.

Thus, naturalness correlates with (or simply influences) type frequency productivity in positive terms, whereas it does not with token frequency. At least not necessarily. In other words, the mere fact of repeating a single type does not influence the possibility for a speaker to extract a productive schema and to expand it throughout the lexicon. On the other hand, for a schema to be extracted and expanded it must be natural, i.e. transparent.

This evidence crucially supports Jakobson’s view of an isomorphic relation between both faces of the linguistic sign. It is this isomorphic relation that is considered unmarked for morphological coding (not in general: in fact, it is marked from other points of view, because it gives rise to very long lexical units in agglutinating languages where the diagrammatic principle is fairly prominent). And this is labelled as natural for morphology. The more transparent the isomorphic relation within a complex diagram is, the more natural and preferable that behavior will be. Natural Linguistics always emphasizes the processual dimension of linguistic competence. In this sense, it inherits important concepts from structuralism, as we have seen above, but it goes beyond it focusing on the generative capacity of speakers, which was absent in the era of classical structuralism.

Recently, there have been several efforts to reduce markedness to “usage-based” notions such as economy of use relating to frequency, the latter being understood as a primitive (cf. Bybee 2001, Blevins 2004, Haspelmath forthcoming). I will not discuss the question whether frequency is to be seen as a primitive, or rather as the result of deeper forces. For my purposes, it is
sufficient to repeat the crucial distinction drawn above between type and token frequency, and to recall the positive correlation between higher type frequency and transparency/naturalness.

However, it is questionable whether frequency alone can explain (not necessarily) linguistic behavior. As shown by a well-known example in perceptual psychology which has evident implications for lexical semantics, an eagle is perceived as a much more prototypical instantiation of the category ‘bird’ than a chicken despite its lower ‘text frequency’ (or experiential entrenchment). I have never seen an eagle in my life, nor am I interested in ornithology; nonetheless, I share the very same feeling although I eat chicken once a week. Thus, frequency may surely have an impact on categorization, but it determines neither the prototype nor the shape of the categories (cf. Violi 1997: 183-186). In other words, although a chicken is statistically more frequent, i.e. a more normal bird, it is a less natural representative of the natural category ‘bird’ than an eagle. To mention a linguistic example pointing out the limits of frequency, let us briefly consider a paper by Hippisley et al. (2004), who have recently investigated suppletion on a large scale. They emphasize that frequency alone cannot be considered the key factor for suppletion to be maintained or developed across languages. On the one hand, there are other relevant factors such as the type of inflectional category involved (inherent vs. contextual), and the nature of the distribution of stems. On the other, they quote languages where suppletion is attested in presence of rather infrequent lexemes, as in Archi, in which the plural of the scarcely salient word for ‘corner of sack’ is suppletive: bič’ni (sg.) l boždo (pl.) (cf. Hippisley et al. 2004: 394).

Discussing several meanings attributed to markedness, and correctly criticizing the amount of confusion often enclosing this concept, Haspelmath (forthcoming: 29) with regard to the question of iconicity, and implicitly of naturalness, observes that “the basic generalization is easily explained by economy: What is used more frequently is shorter in any rational communication system. No appeal to iconicity principle is needed”. He means that for instance the tendency for a singular to be coded by the shorter (actually basic) form and the plural by the longer (affixed) one is due to economy: since singulars are more frequent, they are coded in the shorter way.

I see two questionable assumptions here. First, it is claimed that frequency as a primitive property of words directly influences the way in which morphological information is coded. In particular it structures the way information is transmitted obeying G. K. Zipf’s least effort principle: more frequent means shorter. To be sure, frequency correlates with the mental status of words. This is undeniable given the increasing empirical evidence accumulated in psycholinguistic studies, but as such it does not offer any cue for understanding how the morphology of a language works. Let us consider for instance the general observation that
derivatives are usually less frequent than their bases (cf. Hay 2001). This does not offer any key for explaining why conversion is less preferred than affixation cross-linguistically, unless a connection with a notion such as naturalness is assumed. Here, Haspelmath appeals to the concept of overt marking vs. zero marking, and proposes replacing markedness with overt marking since descriptively more adequate. However, the relation with transparency and with isomorphic coding of form and meaning is completely lost.

Finally, Haspelmath (forthcoming: 30) claims that “type frequency is one of the key factors determining morphological productivity, for well-understood psychological reasons”. And this should be the final word on the usefulness of markedness for morphology, since type frequency decides whether a morphological device is productive or not. Unfortunately, this assumption as such does not hold against the empirical evidence. A good example is again provided by the Italian nominalizing suffixes -mento and -(z)ione (for a more detailed analysis, see Gaeta 2002a: 66-77; 2005a). The first one displays a much lower type (and token) frequency than the second, but it is more productive (at least relying on a quantitative measure of productivity such as the one proposed by Baayen 2001 and revised by Gaeta & Ricca 2006). It does not strike us as a surprise that -mento derivatives also are fairly more transparent than -(z)ione formations.11

Leaving aside the problem of economy for the moment, a second questionable assumption implicit in Haspelmath’s approach concerns the shortness of plurals with respect to singulars, which introduces us to the question of universality and naturalness. In fact, Haspelmath (forthcoming: 19) observes that “coding length has no or very little influence in cases like the singular–plural distinction”, quoting data from languages like Sanskrit, Latin or French (and we can add Italian) where plurals are by no means longer than singulars, and still “in these languages, too, the plural is much rarer than the singular”. Again, text frequency is claimed to explain markedness, and, even worse, iconism, i.e. the ‘more meaning – more form’ principle, apparently proves false.

As repeatedly emphasized (cf. Wurzel 1984, 1987), constructional iconism cannot be taken for granted in a simplistic way. In fact, there may be differences between the universal tendencies and the language-specific structural conditions at play. In this respect, Wurzel’s research has highlighted the conditions for grasping what system normalcy means for a given language. Thus, in a language like Latin the system-defining structural conditions are organized around a number of properties among which the concept of stem-based inflection plays a relevant role. Compared to Turkish, the Latin morphological system is therefore less natural from the viewpoint of universal naturalness.

However, this is only relevant from a typological-comparative point of view (on the role of typological adequacy in NM, cf. Dressler 1985a, 1988); to establish the real conditions of markedness for a specific system, such a universal
level is less significant. Let me be more explicit by drawing a parallel with the typology of basic word order. The universal tendency predicts a consistent linearization of constituents and accordingly VO should correlate with NA, whereas OV corresponds to AN. The fact that many languages (e.g., English!) violate such correlation does not falsify the general picture; rather, it highlights the complex role played by other factors (e.g., language contact, sound change, etc.) in shaping a specific language system.

Thus, the general preference holds at a universal level; how a single system is organized must be carefully investigated taking into account the specific (among others historical) conditions which influenced the development of that system. Wurzel’s concept of system normalcy aims at spelling out how a specific morphological system is usually organized.

A general organizational procedure of any inflectional system is given by paradigms: they tend to be organized by means of paradigm structure conditions which establish implicational relations among the different paradigmatic slots; this makes the system consistent and therefore easier to learn by children. Moreover, inflectional classes tend to be anchored at extra-morphological (i.e., phonological, syntactic, semantic) properties which warrant for their stability and productivity. In order to make this concept clear by means of a clear-cut example, let me quote the case of German modals (cf. Gaeta 2002b).

The verb which is nowadays represented in German by the modal *wollen* ‘will’ goes back to an old optative of the old athematic class of the so-called *mi*-ending verbs (cf. Braune 1987: 307). It consequently displayed the endings that usually appeared in the preterite subjunctive of the other inflectional classes. This verb happened to be attracted by the inflectional family of modals in Old High German as well as in other Germanic languages.

The latter group was characteristically (but not uniquely, see Gaeta 2002b for a more detailed analysis) formed by the so-called preterite-presents displaying a number of particular inflectional features (= IF); the latter are summarized by means of the following Wurzelian paradigm-structure condition (= PSC):

\[
PSC: \left[ \begin{array}{c} V \\ + \text{modal} \end{array} \right] \supseteq \left[ \begin{array}{c} \text{IF1} - \emptyset/1./3.\text{ps.sg.pres.ind.} \\ \text{IF2} - t/2.\text{ps.sg.pres.ind.} \\ \text{IF3} - A/\text{sg.pres.ind.} \\ \text{IF4} - n/1./3.\text{ps.pl.pres.ind.} \\ \text{IF5} - i - /\text{subj.} \end{array} \right]
\]

If its development since the oldest stages of German is considered and compared to other Germanic languages (such as Gothic), *wollen* reveals a triumphant march to come close to the modal family, as can be gathered from the following table,
in which the Gothic *wiljan* is shown near the Old and Middle High German *wellen* and contrasted with the MHG modal *durfen* ‘need’:

(3)

<table>
<thead>
<tr>
<th>inf.</th>
<th>Goth. <em>wiljan</em></th>
<th>OHG <em>wellen</em></th>
<th>MHG <em>wellen</em></th>
<th>MHG <em>durfen</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>pres.ind.</td>
<td>1.sg.</td>
<td><em>wiljau</em></td>
<td><em>willu</em></td>
<td><em>will</em></td>
</tr>
<tr>
<td></td>
<td>2.</td>
<td><em>wileis</em></td>
<td><em>wili</em></td>
<td><em>wilt</em></td>
</tr>
<tr>
<td></td>
<td>3.</td>
<td><em>wili</em></td>
<td><em>wili</em></td>
<td><em>wil</em></td>
</tr>
<tr>
<td>1.pl.</td>
<td><em>wileima</em></td>
<td><em>wellemēs</em></td>
<td><em>wellen</em></td>
<td><em>durfen</em></td>
</tr>
<tr>
<td></td>
<td>2.</td>
<td><em>wileih</em></td>
<td><em>wellet</em></td>
<td><em>wellet</em></td>
</tr>
<tr>
<td></td>
<td>3.</td>
<td><em>wileina</em></td>
<td><em>wellent</em></td>
<td><em>wellent</em></td>
</tr>
</tbody>
</table>

Leaving aside IF5 which is quite irrelevant here, *wellen* has almost entirely approached the family of modal verbs, as recapitulated by the following table:

(4)

| IF | Ø/1./3.ps.sg.pres.ind. | NO!: *willu* vs. *wili* | NO!: *wil* | YES: *wil* | YES: *wilt* | YES: *will*- vs. *well-* | NO!: *wellemēs* vs. *wellent* | NO!: *wellen* vs. *wollen(t)* |
|----|----------------------|--------------------------|------------|-----------|-------------|--------------------------|-------------------------------|

Except for the deviant IF4, which is however on the way of being levelled because forms such as *wellen* are also attested (cf. Paul 1989: 268), the MHG *wellen* has entered the inflectional family of modals. Clearly, the motivation for such inflectional changes cannot be sought in overt vs. zero coding, because in some cases the introduction of zero coding takes place (cf. OHG *will-u* vs. MHG *wil*); nor may type frequency as such have played any role because modals were a small nest of verbs, displaying a highly idiosyncratic inflection. Moreover, as for economy of usage, modals, including *wellen*, were likely to be all highly frequent verbs, which leaves little space for an explanation in terms of the least effort.

If, however, one assumes that *wellen* acquired the extra-morphological property of being a modal, then the morphological changes follow, since they contributed to levelling markedness off the system by anchoring the paradigm of *wellen* at the solid PSC proper of modals. One can argue that in this way an economic result is obtained because of the reduction of an overload of idiosyncratic inflectional information. In fact, naturalness IS economic, if correctly considered in its component-specific processual dimension (cf. Wurzel 1997,
Bender 1998, Croft 2003: 117). It is only along this dimension that economy, i.e. a cost–benefit assessing procedure, can be evaluated and made fruitful. This is actually what diachrony shows: structural changes appear to be motivated by economy, i.e. markedness reduction!

On the other hand, the change exemplified by *wollen* resembles analogy: the latter has been traditionally assumed to restore regularity within paradigms counterbalancing the destructive effect of sound laws. In effect, Wurzel (1988c) has pointed out the debt paid by NM to H. Paul’s analogy theory. Compared to Paul’s approach, however, NM is far more pretentious, because it aims at providing a clue for predicting the direction of analogical change by making reference to universal principles such as the ones highlighted above.

Admittedly, in spite of the explanatory force of the principles advocated by NM there are “recht wenige Sprachen mit einer maximal natürlichen und insofern ‘idealen’ Morphologie” (Wurzel 1988c: 544). However, this depends on phonological change, that “im Sinne des bekannten Paulschen Dualismus von Lautgesetz und Analogie ständig die ‘Symmetrie des Formensystems’ stört und zerstört” (*ibidem*). The reference to Paul’s dualism makes the complex dialectics between the preferences proper of morphology and the other levels of language analysis (including typological adequacy, E. Coseriu’s norm, etc.) explicit, which has a central role for Natural Linguistics (as well as for Jakobson) in steadily moulding the shape of natural languages. In this light, exceptions are not a serious problem (cf. Wurzel 1988a).

However, one might object that since, in the complex dialectics between universal and system-related morphological naturalness, the latter is claimed to prevail (cf. Wurzel 1984: 110-113), a potential source for unnaturalness shows up here, as we have seen in the case of zero marking for the MHG *wellen*. In this vein, I am not aware of any attempt to investigate the naturalness of system normalcy, namely to check whether system normalcy for a given language may be systematically unnatural. Although this is an empirical question, one may lean towards a negative answer, because morphology is permanently reinforced by grammaticalization, which by itself has little to do with markedness and naturalness (cf. Lehmann 1989). Nonetheless, it steadily contributes to supply iconic morphological coding (cf. Bybee ms. for a similar view).¹⁴

For NM, markedness is the direct result of viewing form and meaning as strictly related, even though in a non-trivial way. Denying the very concept of markedness amounts to make such claim empty, and give up the (Jakobsonian) belief that form and meaning may be thought of as isomorphic. As a consequence, a strong anti-functionalist ingredient enters the theory of language, which is by the way shared by most formal (for instance generative) linguistics.¹⁵ In this light, is this denial really useful?
3. Economy and the fallaciousness of irregularization

We have already touched upon the possible role played by economy in shaping language. In particular, it has been observed that markedness reduction has the effect of increasing the naturalness of a morphological system by rendering the system more consistently organized along universal principles such as system adequacy, extra-morphological motivation, etc. By doing so, the power of analogy is exploited, which is however guided by the naturalness principles and counterbalances the destructive force of phonological change, as in Paul’s dualism.

This view of economy of a morphological system should not be confused with other sources for economy coming from other components, and specifically phonology. In fact, phonological economy may be obtained in two ways: reduction of articulatory effort and increase of perceptual ease. The former amounts to saving energy in producing shorter signals; the latter may also cause the increase of energy costs – for instance, in the case of realignment of an iconic relationship between phonemes and allophones.

A case in point is given by r-insertion in several English varieties, where a non-historical [r] occurs between (a number of) lax word-final vowels and following word-initial vowels: I like Rosa, but Rosa[r] is leaving now; I saw Bill, but I saw[r] Allison, etc. The insertion aligns these tokens with the cases where /r/ is underlingly present and then deleted: Your brother[r] is coming, but I saw your broth[œ]. From a historical viewpoint, it is clear that [r] was lost from post-vocalic position within the syllable-rhyme: bird > [bœ:d], bar > [ba:], brother > [‘brœðə], etc. The surface [r]/Ø alternations resulting from this sound change were interpreted “as instances of deletion or insertion by language learners, since, in both cases, the phonological environments were transparent, and could be stated in terms of natural classes of sounds” (Blevins 2004: 68).

This rule inversion (cf. Vennemann 1972) has the effect of re-establishing a more iconic relationship between phonemes and allophones: in fact, a good allophone stands in a biunique relation to a phoneme. In this case, de-iconization is due to the fact that the allophones resulting from the lenition process are partially homophonous with allophones of other phonemes (cf. Dressler 1985b: 307ff., Gaeta 2001: 109). Thus, iconism (even though of a different nature with respect to morphological iconism, which – in the light of the previous discussion – should not surprise us) produces an anti-economic effect, because the speech signal lengthened. However, this anti-economicity is only apparent: at the level of the cost-benefit evaluation an iconic phoneme–allophone relation is preferred over a bothering scanning procedure which had to distinguish between cases of deletion and cases where no /r/ was underlingly present. This unnatural, anti-economic state of affairs is maintained in the standard variety by means of a strong normative effort (cf. Donegan 1993: 118).
Thus, naturalness (intended as the opposite of markedness) and economy aim at the same goal. This is summarized by Wurzel’s (1997: 306) “Grammatical Economy Maxim”: “Rede so, dass du soweit möglich stärker markierte grammatische Erscheinungen vermeidest”. The maxim incorporates Keller’s (1990) view of language change as resulting from the interaction of the Gricean maxims underlying the speakers’ behavior in normal communicative contexts. Accordingly, language change is held to be the unintentional result of single intentional acts, i.e. an invisible-hand phenomenon. On the other hand, Wurzel’s maxim lays down the role of markedness and its bearing on the cost-benefit evaluation by appealing to a general markedness universal: “Grammatische Erscheinungen belasten die menschliche Sprachkapazität umso mehr, je stärker sie markiert sind” (Wurzel 1997: 307).

In radical contrast with this view, which sees economy and naturalness as strictly interconnected, economy has been claimed to be a basic anti-naturalistic force, aiming at reaching formal differentiation by increasing irregularization.

Nübling (2000, 2001, 2005) postulates the existence of two different irregularization strategies, which explain the strong degree of idiosyncracy typical of highly frequent verbs such as have and say in the Germanic languages (and in many other language families, too) – the so-called reductive and non-reductive irregularizations. Together with a number of other strategies (such as overdifferentiation, boundary crossings, etc., cf. Nübling 2001, 2005), they constitute the core of Economy Theory. Recognizing the benefit of irregular relics, Economy Theory “considers irregularity including suppletion to be increased formal distinctivity [sic!]” (Nübling 2001: 54). Irregularization strongly correlates with high token frequency, which in its turn correlates with shortness of expression as well known since Zipf’s least effort law mentioned above. The basic tenet of Economy Theory is that “irregularity not only (passively) develops by the preservation of nonfunctional relics, but can also be ‘created’ actively by innovative processes” (Nübling 2001: 54). The increase of distinctivity is the alleged goal underlying such irregularizations.

To support this claim reductive and non-reductive irregularizations are mentioned, which should show the active role of irregularization in preserving or increasing formal distinctivity. The latter “has the advantage of protecting the forms which become increasingly shorter under the effects of high token frequency from homophony (syncretism)” (Nübling 2001: 69). Again, homophony is the expected result of increasing word shortness; however, “[t]he more strongly and further forwards the word is differentiated, [...] the more strongly it can be reduced without the danger of homonymy” (Nübling 2001: 69).

Before discussing the empirical evidence brought about to support this view, it must be emphasized that nobody denies the reductive role played by phonological change against morphological coding, nor underestimates the effect of
such reductions on highly frequent words. This has been repeatedly discussed within NM, showing the complex interaction between phonology and morphology (cf. Wurzel 1980, 1990, Dressler 1985b). Therefore, the positive correlation between token frequency and phonological reduction which gives rise to morphological irregularity is fairly well accommodated within naturalness theory. The question is rather to understand whether irregularization is the mere result of such conflicting tendencies, or whether it can also be created regardless of natural (i.e. either phonological or morphological) changes. In other words, the question is whether morphological change can be intrinsically unnatural producing a higher degree of irregularity. It should be pointed out that this view also challenges Paul's dualism introducing an anti-analogical finalism.

In order to support this claim, reductive and non-reductive irregularizations are quoted by Nübling (2001: 65). However, the first type is unable either to verify or to falsify the theory, since it is independently justified by the reductive role played by phonological change. This is true for the Norwegian la < late ‘let’, the NHG hast/hat < *habst/*habt ‘(thou) hast / has’. That such changes happen is not surprising in view of the high token frequency of these verbs. There are indeed cases where phonological reduction leads to a reanalysis in morphological terms, for instance the Old Frisian hade > Modern Frisian hie ‘had’ under the influence of die ‘did’. As admitted by Nübling (2001: 65), “it is difficult to decide whether the analogy was motivated by the shortness of the word”.

The second type, namely the non-reductive irregularization, should crucially highlight the active role of the regularization strategy. Let us review in some detail whether this evidence provides the desired support to Economy Theory. In this respect, Nübling (2001: 59) mentions the case of the verbs HAVE and SAY in Faroese and Icelandic, where a “stable instability” is observed, in that the respective verbs swing between the original ē-class and the very productive ja-class of Germanic weak verbs. Again, in the light of the high productivity of the ja-class, it does not strike us as a surprise to observe that the more productive model expands its influence even on very frequent verbs. How far such an influence goes can only be ascertained on the basis of a thorough analysis of the system carried out with the help of Wurzelian PSCs. The actual mixed paradigms are therefore highly irregular; one wonders, however, whether this irregularity (or overdifferentiation) is the goal of the change or rather results from the action of conflicting forces, namely the extension of a productive pattern and the lexical inertia typical of highly frequent words (cf. Gaeta 2005b).

Even worse is the case of the Swedish ge ‘give’ where “former versions of the verb containing e or i combined to a new and thus more strongly differentiated paradigm: gel/ger – gav – gavit ‘give/gives – gave – given’” (Nübling 2001: 65). In fact, as more extensively discussed in Nübling (2000: 117-119), the stem ge- has given rise to other more “natural” forms (e.g., the preterite gedde), which presently
compete with the oldest ones at different sociolinguistic levels. On the other hand, “[i]m Zuge der Standardisierung wurden diese Varianten [scil. forms retaining i or e] gemischt” (Nübling 2000: 118). Therefore, the diachrony of this verb is fairly complex, and shows the interaction of grammatical and extra-grammatical factors. Clearly, no naturalness theory may account for the latter (cf. Wurzel 1994).

A final example for non-reductive irregularization is provided by the Frisian jaan ‘give’, where a preterite joech occurs instead of the expected *jef analogically formed on the basis of the rhyming verb slaan – sloech ‘hit’. Similar to Fris. hie ‘had’ mentioned above, this example illustrates possible analogical remodelling explainable in terms of what Joseph (2005) has recently called “lateral shifts”. He means that morpheme reshuffling as witnessed by Middle Greek 3PL nonactive past ending -ondustan from earlier -ondusan based on 1PL -mastan and 2PL -sastan is actually a change type not going back to any movement from less grammatical to more grammatical as in typical grammaticalization. Rather, they must be thought of as lateral shifts, “since it is not that an element is moving from more to less grammatical” (Joseph 2005). The Frisian examples are in a way similar to Joseph’s morpheme reshuffling, because they all arise by reshaping morphemes on the basis of a complex network of “local” relations occurring among stored forms. In this view, they do not constitute an argument against markedness or its counterpart naturalness; rather, they highlight the possible interconnections arising between two morphemes because of their contiguity within a morphological paradigm or across two close paradigms, as for the rhyming jaan and slaan. After all, the fact that specific models or templates influence the organization of paradigms is another way to spell out PSCs, namely the specificity of morphology “by itself”.

At any rate, all this does not go beyond Paul’s dualism: no active strategy for irregularization exists. Rather, analogy, which points to a better organization of a paradigm, or of paradigm nests, operates at a local level for highly frequent words (especially when serially ordered, e.g. numerals) introducing local optimization. The latter basically follows the same strategy of more general PSCs, namely to save energy costs of lexical storage by generalizing morphological types. These local optimizations arise on the basis of morphological models which are strongly entrenched (highly frequent); however, their aim is not to increase distinctivity, but rather to reduce formal differentiation. Paul’s dualism still resists; it is only applied at a local level.

Further research is needed to understand whether analogy and contiguity are only different in terms of generality, or whether there are other and for the moment less foreseeable forces at play here. This is not to deny that high frequency may cause the compression of the speech signal both in phonological and morphological (in one word: morphonological) terms. Therefore, I subscribe to Werner’s (1990: 169) claim that “je frequenter die Oppositionen vorkommen, um
so stärker, komprimierter und irregulärer [werden] sie zum Ausdruck gebracht”,
with the further observation that it remains to be understood whether compres-
sion also plays a role for lateral shifts. However, I do not see any empirical
evidence supporting an active role of irregularization as a coding strategy.

4. Conclusion on teleology

Naturalness theory does not deny the teleological frontiers of functional expla-
nations: by doing this, it counts on its side Jakobson (1962: 652), who explicitly
recognized the value of teleology: “Since in the process of change its two terms,
the start and the finish, necessarily co-occur and can be compared as to their
place and function in the system, we are enabled and even compelled to seek
the purpose of the change. If mutations are a constituent part of the purposive
linguistic system, then the application of a ‘teleological criterion’ to the analysis
of phonemic changes must be accepted”. However, the alleged principle of
differentiation discussed in the previous section points to a more abstract tele-
ology, according to which independently motivated changes aim at improving
the distinctive function of morphemes, as if the latter were engaged in a sort
of Darwinian struggle for life. Or, even worse, as if the speakers intentionally
changed the shape of morphemes in order to make them more differentiated. As
has been shown, there is no evidence for assuming such an abstract teleology.
Rather, the increased regularization, surely correlating with high frequency, is
the unintentional result of local optimizations, i.e. of the opposite goal: reducing
differentiation by analogically creating local networks.

A similar teleological view may also be ascribed to Leiss’ (1997, 2005) idea
that any homonymy should be rather treated as polysemy. On the one hand, she
refuses the NM principle of uniform coding, because it is factually inadequate
to mirror the wide polyfunctionality/homonymy of morphemes commonly
observed across languages. On the other, she claims that homonymy must be
functionally motivated. I will leave aside the question of the uniform coding
principle, which is less relevant to the perspective adopted here and is admit-
tedly problematic: presumably, that principle should be low-ranked with respect
to other naturalness parameters (cf. Wurzel 1984: 111). Rather, I will briefly
discuss the second claim, which in its extreme formulation is rather surprising
because it rescues the same uniform coding principle by making it general: “Man
nimmt die Form ernst, wenn man hinter gleichen Formen (präziser: hinter einer
Form) auch einen gleichen Inhalt vermutet” (Leiss 1997: 154).

Accordingly, Leiss attempts to show that several cases of supposed ho-
monymy can be rather treated as polysemy, i.e. syncretism, on the condition
that a more abstract (and general) semantic analysis is carried out which adopts
Jakobsonian concepts like archigrams and aims at grasping “die gramm-
atische Architektonik” of a language. Moreover, she entirely discards the
etymological argument, according to which the plausibility of a functional relationship between two homonymous forms crucially depends on the existence of a common etymon (cf. Leiss 1997: 145). Against such traditional approach, Leiss on the one hand objects that Paul’s dualism discussed above is misconceived because analogical processes may also act in accordance with sound laws (cf. Leiss 1997: 146). On the other, she asserts the priority of analogical (or functionally motivated) changes over sound laws and claims: “Lautgesetze wirken danach immer nur dann ‘blind’, solange sie den funktionalen Bereich nicht stören” (Leiss 1997: 147). Thus, for instance, the loss of case markers in German should be interpreted as functionally motivated because their role was taken over by prepositions, whereas the concomitant sound changes simply are processes of erosion of forms bleached long before.

Against the priority of sound laws, Leiss supports the view that forms tend to merge because of content similarities. Accordingly, she reconstructs a syncretism involving “homonymous” -er in German (cf. Leiss 2005): plural suffix (Kind ‘child’ → Kind-er ‘children’); derivational suffix (husten ‘to cough’ → Hust-er ‘cough’); comparative suffix (schön ‘beautiful’ → schön-er ‘more beautiful’). Even without analyzing her theory in detail, its basic framework is unconvincing. In fact, the three German suffixes mentioned by Leiss represent the developments of three different morphemes (resp. an old Germanic stem formative *-iz- going back to IE *-es; an OHG loan suffix -ari borrowed from Latin -ārius, and the old Germanic comparative suffix *-iz- going back to IE *-jes-, cf. Ramat 1986: passim). Even if one can agree with Leiss that mergers may sometimes take place because of content similarities, one wonders how far such an analysis can go. Thus, content similarity is surely responsible for the rising of suppletion in the comparative and superlative degree of the German gern ‘willingly’ – lieber – am liebsten (cf. Ronneberger-Sibold 1988). It seems much more difficult to interpret the rise of homonymy in the case of the OHG -ari coming from the Latin -ārius and for instance the OHG -ir- in jungiro ‘younger’ as due to functional similarity, unless we assume a strong teleology which forces forms to fit into a precompiled semantic canon. The latter is, however, only justified by the existence of the striking homonymy, which gives a flavor of circularity to the whole enterprise.

Therefore, Leiss’ research program, which aims at restricting the extension of homonymy in natural languages, is surely worth pursuing. However, it seems methodologically sound to have explanations of goal-directed activities rooted in causal laws; or, in the weaker case of simple function ascription, to “make evident one role some item plays in a given system” (cf. Nagel 1979: 315). Otherwise, we all run the risk of Pangloss in Voltaire’s Candide (see Dressler 2002 for a discussion), who thought the nose form to be teleologically shaped in order to carry spectacles.
Notes

1 Simone (1995: x) opposes the Platonic Paradigm to the Aristotelian-Saussurean Paradigm. According to the former, “language and reality must resemble each other to some extent if we want to be able to speak of reality without necessarily recurring to it directly”. The latter “claims that language and reality are quite independent of, and do not resemble, each other; this is claimed to be so for reasons of economy and ‘handiness’, since no language could be used if not arbitrarily structured.”

2 The apparent paradox disappears if one takes note of Holenstein’s (1990) reconstruction of Jakobson’s education in the Russia of the late 19th century, deeply influenced by a particular version of Hegelism, and the later encounter with H. Pos’ phenomenological thought. Relying on this philosophical background, Jakobson views a structure, i.e. a form, as indissolubly tied up with the substance composing it: “[s]i chiama sostanza o materia qualcosa che è stato costruito per essere l’elemento di una relazione, sia essa formale o funzionale. Nel momento in cui tale entità sostanziale viene a sua volta analizzata, risulterà essere anch’essa una rete di relazioni” (Holenstein 1990: 23).

3 It is mandatory to quote here from Jakobson’s famous letter to Trubetzkoy (1975: 162-163): “I am becoming ever more convinced that your idea that a correlation is always a relation between a marked and unmarked series is one of your most remarkable ideas. I think it will turn out to be important not only in linguistics, but also in anthropology and cultural history, and that historico-cultural correlations as life and death, liberty and bondage, sin and virtue, holiday and workday, and the like are always reducible to a vs. non-a relations, and that it is important to establish for each period, group, nation, etc., which is the marked series. For instance, for Majakovskij life was the marked series, realized only when motivated, for him not death, but life demanded a motivation”.


5 There is a further claim underlying this observation, which relates to the cross-linguistically widespread phenomenon of affixal homonymy. Namely, the latter is claimed to be due to economy, i.e. parsimony of storage, a general cognitive factor in language design (cf. Dressler 1999: 140, Ronneberger-Sibold 1980). The question is fairly complex because of the reductive effect exerted by phonological change. Recent studies on grammaticalization have put forward the hypothesis that this state of affairs might be due to the overwhelming force of phonological erosion on highly grammaticalized morphemes such as affixes. Therefore, one may get the impression that parsimony of storage is only an indirect effect of a functionally motivated change, which is subject to Nagel’s (1979: 310-311) proviso that “functional ascriptions presuppose that, and are appropriate only if, the system under consideration [...] is ‘directly organized’ or ‘goal-directed’”, in the very specific sense that “the function ascribed to an item contributes to the realization or maintenance of some goal for which the system is directly organized”. Given the undoubted erosive effect of phonological change on morphemes, I am not sure whether parsimony of storage may be invoked as functional ascription to homonymy. At any rate, the following section will settle the role of economy in language.


7 It is not by accident that NM always preferred an item-and-process model for inflectional morphology over the other two other classical models of item-and-arrangement and word-and-paradigm as surveyed by a champion of structuralism like Hockett (1954).

8 Cf. Bybee (2001: 115): “frequency is probably a more basic factor, and [...] the structural
relations posited by Jakobson are describing phenomena conditioned by frequency [...] frequency is a main determinant of markedness”.

9 A similar objection can also be raised against Fenk-Oczlon (1991: 368), who suggests replacing “unmarked” by “frequent” to get “easy explanations” of markedness: “Im Singular äußern wir uns häufiger, weil wir aufgrund der Organisation unserer Wahrnehmung und Vorstellung (Gestaltprinzipien, Invarianzbildungen) dazu tendieren, eine Gestalt (z.B. Bewegungsgruppe) wahrzunehmen oder uns vorzustellen: eine Beere, ein Schaf, einen Fisch; und eine Traube, eine Schafherde, einen Fischschwarm”. However, one wonders what is the egg and what the chicken. In fact, one may easily interpret singulars as unmarked because of our Gestalt-based perceptual ability in spite of the frequent pluralities we are confronted with in the world.

10 To conclude this discussion with a witty remark, Bailey (1996: 91) is worth quoting, who observes that “[t]he distinction between nature and normal can be illustrated with the statement that it is normal to catch colds every winter, but this is not natural because it impedes the fulfilling of our nature”.

11 Similar observations hold true for the English suffix -ment, which shows a high type frequency but is not synchronically productive (cf. Bauer 2001: 48). For this reason, Bauer treats type frequency as “the result of past productivity rather than an indication of present productivity”, adopting Corbin’s (1987: 177) term rentabilité for defining this property. It must be added that transparency is a necessary but not sufficient condition for productivity given that -ment is fairly transparent (cf. on this point Bauer 2001: 54).

12 The universal tendency may be said to indicate the path of possible development within a single language to increase consistency with respect to the general type. It is not by accident that English has also developed a NA order, however only under specific circumstances; how far such language changes will be implemented must be however kept outside the domain of scientific research, which can only make predictions about the possible evolution of a system without becoming deterministic (cf. Wurzel 1988a).


14 This claim must be taken cum grano salis: grammaticalization may have as a side-effect the increase of markedness in the inflectional system as shown by Haspelmath (1998) on the basis of old present forms reanalysed as subjunctive in languages where new present forms arose as a consequence of grammaticalization. Synchronically, the effects are unusual markedness values, because the new present turns out to be more marked than the usually marked subjunctive. The prediction here is that this state of affairs will not remain stable, but will evolve towards better markedness values. There are indeed some clues showing that this prediction is borne out by what is actually going on in these languages, although the caveat against strict determinism in language change must be recalled.

15 A noteworthy exception among generativist circles is constituted by Optimality Theory, which directly incorporates markedness as a universal constraint; in OT-based morphological analyses, markedness conveys the basic idea of constructional iconism (cf. Wegener 2004). On the basis of what has been said above, it does not strike us as a surprise to observe that several OT-theorists mention Natural Phonology as an important ancestor (see Ritt 2001 for a discussion).
The iconism of phoneme–allophone relationship in essence also underlies Blevins’ (2004: 152) Feature-to-segment Mapping Principle: “In the learning algorithm which allows listeners to interpret the phonetic string as a sequence of segments, a phonetic feature, $F_p$, whose domain overlaps with other segment-defining phonetic features, is assumed to have a unique featural source /SF/ in the phonological representation (where F may be a feature or feature-complex).”

Address of the author

Livio Gaeta
Dipartimento di Filologia Moderna “S. Battaglia”
Università di Napoli “Federico II”
via Porta di Massa 1
80133 Napoli
Italy
e-mail: livio.gaeta@unina.it

References


