LEXICAL INTEGRITY AS A FORMAL UNIVERSAL: A CONSTRUCTIONIST VIEW

Geert Booij (University of Leiden)

Keywords: construction grammar, inflection, lexical integrity, morphology-syntax interface, morphological universals, word formation

Abstract:

This paper deals with an important formal universal with respect to the interface of morphology and syntax, the Lexical Integrity Principle. This principle encompasses both non-interruptability and non-accessibility of word-internal structure. Non-interruptability is a defining property of canonical wordhood, and this part of Lexical Integrity is therefore almost never violated. Non-accessibility of word-internal structure should be rejected on empirical grounds, In a constructionist view of morphology, the possibility of syntax and semantics having access to word-internal structure is to be expected.

1.Introduction: morphological universals

Morphology has always played an important role in language typology, that is in the systematic characterisation of variation between languages, and the constraints on that variation. The best known classical form of morphological typology is the ranking of languages by means of two indices, the index of synthesis and the index of fusion (Comrie 2001). In more recent work on word formation, in particular compounding that has been inspired by the Principles-and-Parameters framework of Chomsky (1981), the idea of a head-parameter (a language has either right-headed or left-headed compounds) has shown to be a fruitful typological perspective (cf. Scalise (ed., 1992)). However, it is hard to find uncontested substantive morphological universals of an absolute nature, certainly not in the domain of word formation. This may be a reflection of the fact that the building blocks of word formation patterns often derive from lexical items, through the process of grammaticalization. Hence, word formation patterns will reflect the language-specificity of the lexicon.

The discussion of morphological universals in Greenberg (1963), the publication that forms the historical background for the studies in this book, focuses on two issues. The first is that of affix order. Greenberg showed that derivational morphemes tend to be closer to the root than inflectional morphemes. The tradition of investigating the order of derivation and inflection, and regularities in affix ordering has been continued in the work of Bybee, for instance Bybee (1985) and Bybee et al. (1994).

The second issue broached by Greenberg is that of the morphological marking of inflectional categories. His work focused on the morphological asymmetries in the relation between form and meaning. For instance, Greenberg observed such an asymmetry for the category of number: if in languages with two values for number, singular and plural, only one of the values is formally marked, this is usually the plural. Such generalizations are expressed in the form of

implicational universals. An excellent survey of such universals can be found in The Universals Archive (http://typo.uni-konstanz.de/archive) developed by Frans Plank and his co-workers.

Greenberg's generalizations on the order of inflectional categories can be accounted for by making a distinction between contextual and inherent inflection. Contextual inflection is the kind of inflection that is required to be present by syntactic contexts, such as person and number marking on finite verbs, gender and number marking on adjectives, and the marking of structural case on nouns. Inherent inflection is inflection that is determined by semantic considerations, such as the marking of number and semantic case on nouns. Arguments for this distinction are presented in Booij (1994, 1996). The generalization then is that contextual inflection is peripheral to inherent inflection, whereas inherent inflection is peripheral to derivational morphological constituents. For instance, the marking of structural case on nouns is peripheral to that for number, and both are peripheral to derivational morphemes in a complex noun. The following example from Hungarian illustrates the ordering of inherent inflection and contextual inflection: the accusative suffix *–t* (an instance of contextual inflection) is preceded by two suffixes that express categories of inherent inflection: the number and the possessor of the noun:

(1) gyereke-i-nke-t
child-PL-1PL-ACC
'our children'(acc.)

Whereas absolute universals are hard to find in the domain of word formation, there is certainly evidence for implicational universals. Recent work on morphological universals of the implicational type can be found in Haspelmath (2006, 2007). An example of such an implicational universal in the domain of word formation that deals with the formation of causative verbs is the following: "If a language has causative verbs derived from transitive bases, then it also has causatives derived from intransitive bases", a generalization that Haspelmath (2007) ascribes to the Russian linguists Nedjalkov and Sil'nickij.

At a more fundamental level the issue is how such universals can be explained. There are a number of types of explanation that have been invoked: cognitive and pragmatic factors such as Relevance, Iconicity, and Economy (as in the work of Bybee and the theory of Natural Morphology), processing factors (Cutler and Hawkins 1988), and universal mechanisms and pathways of linguistic change (Anderson 2004, Harris 2004, Haspelmath 2006, Bybee, ms).

As observed by Helmbrecht (2004: 1248), "cross-linguistic generalizations in contemporary morphology are largely generalizations over form-function relations in morphological units" of the sort proposed by Greenberg, such as the generalization about the formal expression of singular and plural mentioned above. Helmbrecht also notes that "[t]here are practically no substantive, i.e. absolute universals with regard to morphology" (Helmbrecht 2004: 1250). This suggests that we have to look for formal universals, if we want to find morphological universals at all. Universals serve to define the notion 'possible natural language' and hence they may have the form of constraints. We may distinguish the following two (related) types of formal universal constraints in the domain of morphology: (i) constraints on the kind of relations that are possible between syntax and word structure, and (ii) constraints on the accessibility of the internal structure of complex words for

modules of the grammar such as the syntax and semantics. Such constraints will be discussed in the next section.

2. Constraints on the interaction between syntax and word formation

There are two well-known constraints on the interaction between syntax and word formation in the literature, the No Phrase Constraint and the Lexical Integrity Constraint.

As to the No-Phrase Constraint, a good summary of the discussion of this constraint can be found in Lieber & Scalise (2006). They show that this constraint is incorrect: certain types of phrases can form part of complex words, as has been shown in many publications referred to in their article. Such facts receive a straightforward interpretation in a modular and constraint-based grammar. The morphological module specifies well-formedness constraints on complex words. This may include the occurrence of certain types of phrases such as $[AN]_{NP}$ in the non-head position of complex words, as illustrated by the Dutch compound $[[[oude]_A\ [mannen]_N]_{NP}\ [huis]_N]_N$ 'old men's home'. It is the morphological module that defines which kind of phrases can appear within complex words. The syntactic module in its turn defines the well-formedness of those word-internal phrases. Hence, these two modules have to operate in a parallel fashion. Thus, such facts pose a problem to the view of the grammar as a set of ordered components.

The Lexical Integrity Hypothesis is a constraint on the interface between rules/constraints of the grammar and the internal structure of complex words (for recent discussions see Lieber and Scalise 2006). Anderson is one of the morphologists who proposed that syntax has no access to word-internal morphological structure:

(2) *Principle of Lexical Integrity*
"The syntax neither manipulates nor has access to the internal structure of words" (Anderson 1992: 84)

In recent work on the interface between syntax and morphology by Ackema and Neeleman (2005), essentially the same position is defended. In their model of the grammar, sentence grammar and word grammar are different parts of the grammar that only touch each other at the level of lexical insertion where the features of syntactic nodes have to match those of (simplex or complex words). For example, it is necessary for a proper account of agreement phenomena for syntax to have access to the feature specification of a noun for the morpho-syntactic category Number. However, it is not relevant how this feature is expressed morphologically. For instance, for the purpose of number agreement, it does not matter whether the plural suffix of Dutch nouns is *-s* or *-en*; we only need to know if a noun is singular or plural.

The principle of Lexical Integrity as formulated in (1) excludes two kinds of syntax-morphology interaction: (i) manipulation of parts of word-internal structure, and (ii) access to word-internal structure. Under manipulation I subsume the syntactic movement of a word constituent, and the splitting up of words by intermediate constituents. I consider the impossibility of syntactic movement of the constituents of a linguistic unit as a necessary condition for that linguistic unit to be a word. As to the possibility of splitting up a complex word, we will see below that this is a very rare phenomenon.

We need the prohibition on the movement of word constituents for explaining why in Dutch and German the rule of Verb Second that places finite forms of verbs in second position in root clauses cannot strand the prefix of a complex verb such as *doordénken* 'to think through completely', whereas the particle in particle verbs such as *dóordenken*'to continue thinking' can be stranded:[1]

(3) Jan door-dacht het problem / *Jan dacht het probleem door
'John thought about the problem thoroughly'

(4) Jan dacht door over het probleem 'John continued thinking about the problem'

Thus, this part of the Lexical Integrity principle may serves as a basic test to find out if a sequence of morphemes is a word or a phrasal lexical unit (cf. Bresnan and Mchombo 1995). Particle verbs in Dutch, though clearly lexical units, cannot be words since the particle can be separated from the verb as illustrated above. Hence, we have to conclude that they are not words, but phrasal units. A phrasal analysis of the Dutch particle verbs is argued for in Booij (2002) and Blom (2005).

This also holds for the Hungarian pseudo-verbal compounds such as *tévét nez* 'be engaged in television watching' discussed in Kiefer (1992). The two parts of this lexical unit can be split in certain syntactic contexts, for instance by the negative word *nem* 'not'. We take the separability of these units as evidence for them not being words. This is confirmed by the fact that the noun constituent *tévét* in this example is marked with accusative case (the suffix *-t*). This assignment of structural case to the word *tévé* shows that it must be an independent word. Given the principle of Lexical Integrity, one does not expect structural case assignment to a subconstituent of a word.

In his study of the word in Eastern / Central Arrernte (a language spoken in the Alice Springs area, Australia), Henderson observes that in complex predicates of this language, "non-verbal morphemes can intervene between the two parts" (Henderson 2002: 114), as illustrated by the following example (in 5a, the complex predicate *arrernelheme* is split by the word *akewele* 'supposedly'):

(5) (a) arrerne akwele lh+eme
place SUPPO REFL+PRES
'supposedly sit down'

(b) arrern+elh+eme akwele
place+REFL+PRES SUPPO

Again, the complex predicates can be interpreted as forming lexical units that consist of two grammatical words, and hence they can be split without violating Lexical Integrity. Henderson (2002: 119) remarks that such combinations of a root and suffixes must sometimes be assumed to form one grammatical word because the root and the suffix part are co-dependent, as in the following example:

(6) apan+erle=arteke re ap+em+ele
feel+GO.ALONG$_1$=SEMBL 3SG:ERG GO.ALONG$_2$+PRES+SAME SUBJ
'like going along continuously feeling (its way)'

Henderson observes that "while the *–ap* part can be analyzed as a verb root, it does not occur productively as the sole root of a non-compound verb" (Henderson 2002:

119). This cannot to be taken as conclusive evidence for such complex words being one grammatical word. In Dutch, for instance, we find many particle verbs of which the verb does not occur on its own (e.g. *opkalefateren* 'to recover', of which the verb *kalefateren* does not occur on its own). Yet, we know for certain that such particle verbs form two grammatical words on the basis of both syntactic and morphological evidence (cf. Booij 2002). Note that there also syntactic idioms containing words that do not occur by themselves, but only in these idioms.

The phenomenon of word splitting is traditionally referred to as tmesis, and occurs in Ancient Greek and Classical Latin texts. However, the Greek and Latin verbs that are said to undergo tmesis can be analysed as cases of particle verb combinations as well, like those in Dutch discussed above. Hence, what is referred to as tmesis does not necessarily form counterevidence to the hypothesis that words cannot be split by syntactic rules.

Another phenomenon that is sometimes considered a case of tmesis is the insertion of a word between the syllables of another word, as in *abso-fuckin-lutely*. However, this phenomenon is a special kind of word formation that makes use of the prosodic structure of its base words (prosodic morphology), and hence does not count against the claims that words cannot be split in the syntax.

2.1. Apparent counterexamples to the prohibition on manipulation

Manipulation as referred to in the definition of Lexical Integrity in (2) above may have various forms: movement of parts of words to another position in the sentence (should be excluded, see the discussion of the examples (3 and 4) above), or the assignment or checking of morpho-syntactic properties of part of a word from outside, i.e. by a syntactic rule or constraint. Number agreement is an example of this latter kind of syntactic manipulation. The Principle of Lexical Integrity predicts that we will not find cases where the number of a word constituent is checked by number agreement. A well known form of number agreement in many languages is that between quantifier and head noun. If a quantifier expresses plural number, its head noun may have to appear with plural number as well. This is, for instance, the case for Spanish. However, it appears that both parts of Spanish copulative compounds have to be plural, as is the case for the copulative compound *poeta-pintor* 'poet-painter': its plural form is *poetas-pintores* (for example, *dos poetas-pintores* 'two poet-painters'.[2] This does not imply, however, that a syntactic rule of agreement of Spanish manipulates parts of words. Instead, this regularity can be interpreted as a morphological phenomenon: if we want to pluralize a Spanish copulative compound, both constituents must be marked for plural. Since Number agreement with *dos* requires a plural word, both parts of the compound are marked as plural.

The Hungarian example *tévét nez* 'be engaged in television watching' is also relevant here. As we saw above, the noun *tévé* is marked here as an accusative form. Since this word is not a word part, but a separate word that is part of a phrasal lexical unit, there is no manipulation of word-internal structure at stake here.

Another apparent counterexample to the 'no manipulation' part of the Lexical Integrity Principle is the phenomenon of gapping of parts of words. Gapping of parts of word is a clear case of manipulation of internal word structure. However, gapping is not necessarily a form of syntactic manipulation, and there is convincing evidence that it is prosodic in nature. Some examples from English and European Portuguese respectively are (Vigario 2003: 251):

(7) a. mono- and polysyllabic
inter- and intranational uses
homo- and heterosexual relations

b. pré- e pós-guerra 'pre- and post-war'
segura- mas lentamente [= seguramente mas lentamente]
'surely but slowly'

(cf. Booij 1985 for a prosodic analysis of gapping in Dutch and German compounds). These are cases of prosodic gapping rather than syntactic gapping: a prosodic word is deleted under identity with another phonological word in the same phonological phrase. That is, this process of gapping does not refer to the morphological structure of words, but to their prosodic constituency, and operates at the level of prosodic structure. This is confirmed by the observation that this type of gapping is not restricted to structures with coordination. In the following Dutch and German examples, there is no coordination:

(8) a. Hij verwisselde de dagblad- voor de weekbladjournalistiek
He exchanged the newspaper- for the weekly-journalism
'He changed from newspaper journalism to weekly journalism'

b. Sie ersetzten Ofen- durch Zentralheizung
They exchanged stove- by central heating
'They changed from stove heating to central heating'

Therefore, this kind of gapping does not violate the principle of Lexical Integrity since this principle pertains to words in the syntactic sense, not to prosodic words.

The occurrence of inflected word forms as parts of complex words also deserves some discussion in this context, since at first sight the occurrence of such word-internal inflectional markings suggest that syntax can manipulate the internal structure of words. First, the inflection might be due to the fact that the relevant part of the complex word is a phrase, as is the case for the Dutch compound *oudemannenhuis* 'old men's home' discussed above. The well-formedness of phrases including their inflectional markings is determined by the syntactic module, even when they form parts of words, and hence this kind of manipulation does not count as counterevidence. Furthermore, we find inflected forms such as plural forms and case-marked forms of nouns in the non-head position of compounds (Booij 1994, 1996). As pointed out in these articles, the kind of inflection that we find there is inherent inflection that is determined by the semantics, and not by syntactic rules such as agreement or structural case marking that require the inflectional marking of word parts by a linguistic unit outside of that word. Hence, they form no counterevidence for the prohibition on syntactic manipulation of word-internal structure. [3]

2.2. Evidence for accessibility

An example that suggests that access of syntactic or semantic rules to word-internal structure cannot be completely ruled out comes from Georgian. In Georgian we find expressions such as the following (Harris 2006a):

(9) sam tit-moč'r-il-i (k'aci)
Three.OBL finger-cut.off-PTCP-NOM man.NOM
'(a man) with three fingers cut off'

The first word *sam* has to appear in the oblique form, because it modifies the word *tit* 'finger' which is part of the second word. That is, for the purpose of both case assignment (to the independent word *sam* only) and semantic interpretation, *sam* and *tit* form a unit. As Harris argues, the word *sam* cannot be considered a part of the next word (even though its form is indeterminate and could also be interpreted as a stem form), because recursive modification is not allowed within Georgian compounds. Hence, it should be interpreted as the oblique form of an independent word. This case assignment thus requires access to the internal morphological structure of the second word in (9). The construction in (9) may be compared to that in (10) where the first word bears nominative case, and you get a different interpretation:

(10) sam-i tit-moč'r-il-i
Three-NOM finger-cut.off-PTCP-NOM
'three (men, people, statues) with fingers cut off'

In (10), the word form *sami* agrees in case marking with the second word as a whole, and hence it is a modifier of the whole word. Note that the word *tit*, being part of a compound, does not receive case marking itself.

This example reminds us of Corbett (1987), who also noted that the internal structure of complex words must sometimes be accessible for elements outside of that word. His prototypical example is the following Upper Sorbian phrase (Corbett 1987: 300):

(11) mojeho (gen. sg. masc.) bratrowe (nom.pl.) dźěći (nom.pl)
My brother's children

The selection of a gen.sg.masc. form of the possessive pronoun requires the pronoun selection process (agreement) to have access to the nominal base *bratr* 'brother' of the adjective *bratrowe*. The genitive case of *mojeho* depends for its semantic interpretation on the fact that the possessive pronoun and the nominal base *bratr* form a semantic unit. As in the Georgian example discussed above, it is not the case that a part of a complex word is manipulated by the syntax (in the sense that case is assigned to a word-internal constituent). Instead, the selection of a specific word form requires access to the internal structure of the complex word that it modifies. This once more underlines that manipulation and accessibility are different notions, and should be distinguished when we investigate the viability of the principle of Lexical Integrity.

As far as the semantics is concerned, the Georgian example is similar to the English phrase *transformational grammarian* in which the adjective *transformational* modifies the constituent *grammar* of the complex word *grammarian*. This is a well known example of a bracketing paradox: semantically the adjective *transformational* forms a constituent with *grammar*, but syntactically it forms a unit with *grammarian*. A similar bracketing paradox can be observed for the English phrase *a hard worker* and its Dutch counterpart *een harde werker*. The word *hard* receives a specific adverbial interpretation 'with great intensity' which is

dependent on the presence of a verb that it can modify. This verb is available as part of the deverbal noun *werk-er*. The Dutch version is even more interesting than the English one, because the Dutch word *hard-e* is overtly inflected as an adjective witness the presence of an inflectional ending *-e*.

The additional property of the Georgian construction is that the semantic scope of the numeral modifier includes the semantic information expressed by the case marking. Note once more that there is no case marking of the relevant word part itself. Hence, the possibility of manipulation of word-internal structure by the syntax can still be excluded, but access to the word-internal structure is necessary in order to give a proper interpretation to the semantic role of the word part *tit* and the oblique case of the numeral *sam*. We might therefore consider this example as showing that rules of semantic interpretation may need access to word-internal structure, a topic to be dealt with in more detail in the next section.

Another case of syntax requiring access to word-internal structure is the phenomenon of construction-dependent morphology. This is the situation in which a syntactic construction requires or allows a particular morphological form of words in that construction.

Normally, the syntax may require particular morphological properties (case, number, person, finiteness, etc) to be present in words, as is the case for phenomena such as agreement and government, and the choice between finite or non-finite forms of verbs.

There are also construction-specific requirements of this sort. For instance, the Dutch progressive construction has the shape [*aan het* V-INF]$_{PP}$, as in *Ik ben aan het fiets-en* 'I am cycling' (Booij in press). In that case, the syntax specifies information about the verb as a whole. It does not matter whether the infinitival form is realized by the suffix *-en*, the default suffix, or by *-n* (as is the case for a small set of Dutch verbs such as *doe-n* 'to do'). Moreover, the infinitival form of verbs is not tied to this construction, and is used in other constructions as well.

The form of construction-dependent morphology that is of special relevance for the discussion on lexical integrity is the situation in which words with a specific affix occur in a specific syntactic construction only. An example of that situation is the case of the inflection of Dutch numerals discussed in Booij (2005a). The use of the inflected forms of most numerals is restricted to a number of specific constructions that are exemplified in (12):

(12) a. *collective adverbialI*
met ons / jullie / hun drie-en
with us / you / their three-*en*
'the three of us /you / them together'

b. *collective adverbial II*
met zijn drie-en
with his three-*en*
'the three of us / you / them'

c. *temporal expressions*
bij zess-en
at six-*en*
'at about six o'clock'

na zeven-en
after seven-*en*
'after seven o'clock'

voor en-en
before one-*en*
'before one o'clock'

d. *number of parts*
Het schip brak in drie-en
The ship broke into three-*en*
'The ship broke into three pieces'

e. *appositive collective*
wij / ons drie-en
we / us three-*en*
'the three of us (SUBJ. / OBJ.)'

In present-day Dutch, the ending *-en* is used as one of the two plural suffixes for nouns (the other suffix is *-s*). Historically, however, the ending *-en* in the constructions (12) is a case ending, for instance the inflectional ending required by a preposition, as in (12a-c). The temporal expression *voor en-en* 'before one o 'clock' (12c) makes it quite clear that *-en* cannot have been a plural suffix originally, since it would be semantically odd to use a plural form for the notion 'one hour'. The 1sg. plural pronoun *ons* in (12a) was originally the oblique form of a personal pronoun ('us') as required by the governing preposition, but could be reinterpreted as the possessive pronoun *ons* ('our'). Hence, the numerals in *-en* could be reinterpreted as plural endings in the collective constructions, and this is what happened (Van Loey 1959: 154). This is proven by (12e) where we find the numeral ending in- *en* preceded by the subject, non-oblique form of the 1st person plural pronoun *wij* 'we'. This reinterpretation of the ending *-en* as a plural form is also confirmed by the change of the original form *twee-n* (the inflected form of *twee* 'two') into the plural form *twee-en*. This reinterpretation also led to the collective adverbial construction exemplified in (12b), with the fixed 3rd person singular possessive pronoun *zijn* 'his'. When this construction is used, there is no agreement for person and number with the subject, as illustrated by the following sentence:

(13) Wij gaan met zijn drie-en naar het feest
We go with his three-*en* to the party
'The three of us will go to the party'

In this sentence a 1.pl. subject is combined with the 3.sg possessive pronoun *zijn* 'his'. So there are two different collective adverbial constructions that are identical except that the possessive pronoun can be either a variable (and thus subject to the normal agreement constraints for possessive pronouns), or a fixed possessive pronoun *zijn* 'his'.

The reader will have noticed that in the glosses for the examples in (12) I did not make use of the morpho-syntactic feature PLURAL, but I mentioned the concrete Dutch suffix *-en* instead. There are two reasons for this. First, in (12c) an interpretation of *-en* as a plural suffix does not make sense semantically. Secondly, the specific suffix *-en* that is required by this construction cannot be equated with

the abstract morpho-syntactic feature PLURAL, because this feature is expressed by either *-s* or *-en* depending on the prosodic make up of the stem (*-s* after an unstressed syllable, otherwise *-en*, cf, Booij 2002). For instance, the plural for the number 7 is *zeven-s* /zevəns/, as is the case when we talk about grading (*Jan kreeg twee zeven-s* 'John got two grades 7'). Yet, in the use of the word *zeven* shown in (12), the form of *zeven* is *zevenen* /zevənən/. The same applies to the number *negen* '9' /neɣən/: its plural is normally *negen-s*, but in these constructions it should be *negen-en*.

These observations imply that we have to specify the presence of a specific suffix *-en* in the constructions exemplified in (12). For instance, the constructional idiom for phrases like that in (12b) is:

(14) $[[\text{met}]_P\ [\text{z'n}\ [[\text{x}]_{\text{Numeral}}\ \text{-en}]_N]_{NP}]_{PP}$
with his x-en
'the x of us / you / them'

This analysis implies that the principle of Lexical Integrity as formulated in (2) is too strong, and that the syntax may require access to the internal morphological structure of words. This admittedly special situation is the effect of the rise of syntactic constructions in which specific morphological information is 'frozen'.

In short, syntactic constructions may require the presence of specific morphemes to be present in words, and hence the visibility of word-internal structure to syntax cannot be excluded completely.

3. Accessibility of word-internal structure: semantic scope phenomena

The discussion in the preceding section implies that, although syntax may not be allowed to manipulate the internal structure of word, there are cases in which syntax does require access to the internal structure of words. This means that the interface between morphology and syntax is such that the syntax may have to see word-internal morphological structure.

Word-internal structure needs to be visible to phonology as well. There is abundant evidence that the computation of the correct phonetic form of complex words may require information about morphological structure, This holds in particular for the computation of prosodic forms of words (syllabification and stress patterns) (Booij 2005c).

A-morphous morphology, as defended in Anderson (1992) for the domain of derivation – not for compounding which he considers as being syntax-like - , has also to be rejected for morphological reasons. There is massive evidence that the internal structure of complex words has to be accessible to morphological processes (cf. Carstairs-McCarthy 1994, Booij 2002).

This position is corroborated by psycholinguistic findings. Psycholinguistic research has provided clear evidence for the existence of word families, i.e. families of words that share one or more constituents (Schreuder and Baayen 1997). The notion 'word family' presupposes the accessibility of word-internal structure as well, because a family is to be defined as a set of words that share one or more morphological constituents. Hence, the size of a family can only be determined if the word-internal structure of words is accessible.

In this section I will discuss some data on the accessibility of word-internal structure of complex word to another level of the grammar, that of semantics. In particular, in Dutch NPs of the form [(Det) A + N]$_{NP}$, the adjective may have scope over the first constituent of the complex word only, rather than over the complex word as a whole. The occurrence of this kind of restricted scope is pervasive in texts on issues of government policy. My source of data is Bijker and Peperkamp (2002), a Dutch science policy document with a title that translates as *Involved humanities. Perspectives on cultural changes in an area of digitalization.* The following phrases can be found in this document:

(15) [A [NN]$_N$]$_{NP}$

a. visuele informatie-verwerking
visual information processing
'processing of visual information'

b. intellectuele eigendoms-rechten
intellectual property rights
'rights of intellectual property'

c. taalpolitieke beleids-makers
language political policy makers
'language policy makers'

d. electronische reproductie-rechten
electronic reproduction rights
'rights of electronic reproduction'

e. digitale kennis-omgevingen
digital knowledge environments
'environments of digital knowledge'

f. wetenschappelijke kennis-cyclus
scientific knowledge cycle
'cycle of scientific knowledge'

g. publieke oordeels-vorming
public opinion formation
'formation of public opinion'

h. interactieve gebruiksmogelijkheden
interactive use possibilities
'possibilities for interactive use'

(16) [A [N-suffix]$_N$]$_{NP}$

wetenschappelijke onderzoek-er
scientific research-er
'person who does scientific research'

(17) [A [A-suffix]$_N$]$_{NP}$

a. wetenschappelijke deskundig-heid
scientific expert-ness
'scientific expertise'

b. digitale vaardig-heid
digital competent-ness
'digital competence'

In these examples, the first word, an inflected adjective, has semantic scope over the first part of the second (complex) word. For instance, *visuele informatieverwerking* 'visual information processing' is the processing of visual information, not the visual processing of information. Similarly, in the phrase *wetenschappelijke deskundigheid*, the adjective *wetenschappelijke* 'scientific' modifies the part *deskundig* 'competent': it refers to the property of being a scientific expert, not to the scientific property of being an expert. All these examples are therefore 'semantic bracketing paradoxes' of the same type as *transformational grammarian* (cf. Spencer 1988 and Beard 1991 for discussion).

It is obvious that these are not cases where the first two words form a phrasal constituent that is embedded in a compound, given the inflectional rules for Dutch adjectives. For instance, if the phrase *intellectuele eigendomsrechten* had the structure [[*intellectuele eigendoms*]$_{NP}$ *rechten*]$_{N}$, the presence of the inflectional ending *–e* at the end of *intellectuele* could not be explained, since the correct phrase is *intellectueel eigendom*, without the final inflectional *–e*, because *eigendom* is a neuter noun that requires *intellectueel* as the attributive form of the adjective.

In Spencer's (1988) analysis, the restricted scope interpretation of *transformational grammarian* is related to the existence of a lexical unit *transformational grammar* and the word pair *grammar-grammarian*, and therefore analysed as an analogical formation.. However, restricted scope also occurs in cases where such an explanation cannot be invoked. In the examples if restricted semantic scope given above such NPs are not available. For instance, there is no well formed phrase *digitale vaardig* that can be invoked as a basis for the forming of *digitale vaardig-heid*, since the relevant AP should have the form *digitaal vaardig* 'digitally competent', without the first word being inflected: it is an adverb, not an adjective in pre-adjectival position. Similarly, there is not a well-formed phrase *wetenschappelijke onderzoek* 'scientific research' that can be related to *wetenschappelijke onderzoeker* 'scientific researcher' since the correct phrase is *wetenschappelijk onderzoek,* without a final schwa on the adjective.

These facts clearly show that word-internal structure must be visible to rules of semantic interpretation. Hence, the Principle of Lexical Integrity should be phrased in such a way that it does not forbid rules for the semantic interpretation of phrasal constituents to have access to word-internal structure.

The accessibility of word-internal structure for reasons of semantic interpetation is also important at the word level, for the proper scope assignment of bound modifiers such as prefixes within complex words. This is illustrated by the following two Dutch examples:

(18) pro-Pakistaan-s-e extremisten 'pro-Pakistan extremists'
tussen-gemeente-lijk-e oplossingen 'inter-council solutions'

In the first example the part *Pakistaans* is an adjective derived from the noun *Pakistan* (the *-e* is an inflectional ending). The semantic scope of *pro* is the nominal base *Pakistan* as shown by the interpretation specified in the gloss. In the second example we see a denominal adjective *gemeente-lijk* derived from the noun *gemeente* 'council'. Clearly, the prefix *tussen* 'between' must have scope over the nominal part *gemeente* only, given the meaning 'between councils' of the adjective *tussengemeentelijk*. This observation, however, is not so much an argument concerning Lexical Integrity as rather one against a-morphous morphology.

4. Lexical integrity and construction morphology

The restriction of Lexical Integrity to a prohibition on movement, splitting, deletion of parts of words and the assignment by the syntax of morpho-syntactic properties to word parts, receives a natural explanation in the framework of Construction Grammar and the constructionist view of morphology defended in Booij (2005b) for word formation; (cf. Blevins 2006, and Gurevich 2006 for an application of the insights of Construction Grammar to the domain of inflection).

> "In Construction Grammar, the grammar represents an inventory of form-meaning-function complexes, in which words are distinguished from grammatical constructions only with regard to their internal complexity. The inventory of constructions is not unstructured; it is more like a map than a shopping list. Elements in this inventory are related through inheritance hierarchies, containing more or less general patterns." (Michaelis and Lambrecht 1996: 216)

Word can be seen as constructions on the word level. Grammaticalization is an essential factor in the historical development of morphology from syntactic structures, and hence it should come as no surprise that complex words have constructional properties: an internal structure that is visible to rules of the grammar. One of the best known examples in this respect is the rise of compounds from phrasal patterns. Therefore. it has often been stated in the literature that compounds are still syntax-like, and not always easy to demarcate from phrases (cf. Dahl 2004: Chapter 10 for extensive illustration and discussion of the continuum syntactic construction – compound word).

The Dutch examples of construction-dependent morphology discussed in Booij (2005a) are all cases where a specific construction preserved morphological structure that is no longer regular from a synchronic point of view and has therefore to be specified as part of the construction. Such forms of accessibility of morphological information to syntax are to be expected if the notion 'construction' has a central role in the grammar of natural languages.

The main reason why we consider a sequence of morphemes a word is that that sequence behaves as a cohesive unit with respect to syntactic processes. In other words, cohesiveness is the defining criterion for canonical wordhood, whereas other properties such as being a listeme (a conventional expression) are clearly not to be seen as defining properties for wordhood. Hence, if we take the notion word seriously, we might say that its defining property is cohesiveness or non-interruptability. In other words, the 'no manipulation' part of the principle of Lexical Integrity is the proper interpretation of word cohesiveness.

The constructionist view of morphology does not imply that morphology can be equated with the syntax of morphemes. The word remains an essential unit

for stating regularities. Both words and phrasal constructions are domains over which certain generalizations can be stated, and hence the domains of 'word' and 'phrase' are both essential for the analysis of natural languages (Blevins 2006).

5. Conclusions and discussion

The conclusion of this paper is that the principle of Lexical Integrity has to be formulated in such a way that it does not exclude the accessibility of word-internal structure. The question then remains whether we can formulate this principle as an absolute universal that forbids syntactic manipulation.

In this restricted form, this principle is a further interpretation of the universal principle that all languages distinguish between words and phrases. The distinction between complex words and phrases makes only sense if the word exhibits a higher degree of cohesiveness than the phrase, and hence we need the 'no syntactic manipulation' constraint to give substance to the distinction between words and phrases.

Although this form of lexical integrity seems to be the default situation for natural languages and thus defines the canonical notion 'word', there are exceptional cases in which even this restricted form of lexical integrity appears to be a violable constraint.

An example is the behaviour of endoclitics in Udi, a North-East Caucasian language. The relevant facts are discussed in Harris (2000): clitics that function as person markers can appear word-internally, that is as endoclitics, in between two morphemes in complex verbs (and morpheme-internally in simplex verbs).[4, 5]

Another relevant case is that of Arrernte, a language discussed above in section 2. Henderson reports that in this language we may have 'initial separation': "the first two, or rarely three syllables of a verb can optionally be separated from the remainder of the verb. Intervening material seems to be limited to particles, clitics, pronouns, and simple NPs" Henderson (2002: 121). The most telling example is that in which a lexical root 'to cough' is split into two parts due to the presence of an intervening word:

(19) ateke akwele tn+eme
cough$_1$ SUPPO cough$_2$+PRES
'(she's) supposedly coughing'

Hence, there appear to be cases of real violation of lexical integrity, although this is cross-linguistically a marginal phenomenon.

Finally, it should be mentioned that there are also languages in which the internal structure of complex words is accessible to rules of anaphora (Harris 2006b).

In conclusion, this paper has shown that the principle of Lexical Integrity should be formulated in such a way as not to exclude the different modules of the grammar from ever having access to word-internal structure. Moreover, Lexical Integrity as the prohibition on syntactic manipulation of word-internal constituents is not an absolute universal, but rather the default situation.

Notes

* I would like to thank Alice Harris and the anonymous reviewers for their constructive comments on a previous draft of this paper.

1. The rules of Dutch orthography require the particle verb to be spelled as one word, without internal spacing. This reflects the status of 'lexical unit' of particle verbs, but obscures the fact that a particle verb is not one grammatical word.

2. I thank Franz Rainer for bringing this case to my attention.

3. According to Bauer (2001: 704), "Booij's generalization [only inherent inflection feeds word formation] appears to be true, but it is probably not an absolute universal", because Bauer observed a number of cases in which finite verb forms are embedded in compounds, and a few cases of accusative marking of nouns within nominal compounds. Some of the examples that Bauer discusses, are not clear cases of compounding. This applies to the Hungarian lexical unit *tévé-t néz* discussed above in which the first part is marked as an accusative by means of the suffix *-t* (Bauer 2001: 704). As I mentioned above, there is clear evidence that such N V combinations are not compounds but phrasal in nature since the two parts can be split by other words such as the negative word *nem* 'not' (Kiefer 1992: 76).

4. A similar case of clitic intrusion is reported for Sorani Kurdish in Samvelian (2006).

5. Therefore, Anderson (2005: 161-165) concluded that Lexical Integrity is a violable constraint in the sense of Optimality Theory. In Anderson's analysis, this constraint is normally undominated, but in Udi, the positional constraints on person marker clitics are ranked higher, and hence, such clitics can appear word-internally, thus behaving as endoclitics.

References

Ackema, Peter, and Ad Neeleman. 2005. *Beyond Morphology. Interface Conditions on word formation*. Oxford: Oxford University Press.
Anderson, Stephen R. 1992. *A-morphous Morphology*. Cambridge: Cambridge University Press.
Anderson, Stephen R. 2004. Morphological universals and diachrony. In *Yearbook of Morphology 2004*, eds. Geert Booij, and Jaap van Marle 1-18. Dordrecht: Springer.
Anderson, Stephen R. 2005. *Aspects of a Theory of Clitics*. Cambridge: Cambridge University Press.
Bauer, Laurie 2001. Compounding. In *Language Typology and Language Universals*, eds. Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, Vol 1, 695-707. Berlin: De Gruyter.
Beard, Robert 1991. Decompositional composition: The semantics of scope ambiguities and "bracketing paradoxes". *Natural Language and Linguistic Theory* 9: 195-229.
Bijker, Wiebe and Ben Peperkamp 2002. *Geëngageerde geesteswetenschappen. Perspectieven op cultuurveranderingen in een digitaliserend tijdperk*. Den Haag: Adviesraad voor Wetenschaps- en Technologiebeleid.
Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42: 531-573.
Blom, Corrien 2005. *Complex predicates in Dutch, synchrony and diachrony*. Ph. D. Diss. Vrije Universiteit Amsterdam. Utrecht: LOT.
Booij, Geert 1985. Conjunction reduction in complex words: a case for prosodic phonology. In *Advances in non-linear phonology*, eds. Harry van der Hulst, and Norval Smith, 143-160. Dordrecht: Foris.
Booij, Geert 1994. Against split morphology. In *Yearbook of Morphology 1993*, eds. Geert Booij and Jaap van Marle, 27-50. Dordrecht: Kluwer.
Booij, Geert 1996. Inherent versus contextual inflection and the Split Morphology hypothesis. In *Yearbook of Morphology 1995*, eds. Geert Booij, and Jaap van Marle, 1-16. Dordrecht: Kluwer.
Booij, Geert 2002. *The Morphology of Dutch*. Oxford: Oxford University Press.
Booij , Geert 2005a. Construction-dependent morphology. *Lingue e Linguaggio* 5: 163-178.
Booij, Geert 2005b. The demarcation of derivation and compounding: evidence for Construction Morphology. In *Morphology and its Boundaries*, eds. Wolfgang U. Dressler, Dieter Kastovsky, Oskar Pfeiffer, and Franz Rainer, 111-132. Amsterdam / Philadelphia: John Benjamins.
Booij, Geert 2005c. The interface between morphology and phonology. *Skase Journal of Theoretical Linguistics* 2: 17-25.
Booij, Geert. In press. Constructional idioms as products of linguistic change: the *aan het* + INFINTIVE construction in Dutch. In *Construction Grammar and Language Change*, eds. Alexander Bergs, and Gabriele Diewald, 79-104. Berlin: Mouton de Gruyter.
Bresnan, Joan, and Sam A. Mchombo 1995. The Lexical Integrity Principle: Evidence from Bantu. *Natural Language and Linguistic Theory* 13: 181-254.
Bybee, Joan 1985. *Morphology. A Study of the Relation between Meaning and Form.* Amsterdam / Philadelphia: John Benjamins.
Bybee, Joan. Ms. Mechanisms of change as universals of language.
Bybee, Joan L., Revere Perkins,and William Pagliuca. 1994. The evolution of grammar. Tense, aspect, and modality in the lamguages of the world. Chicago: ChicagoUuniversity Press.
Carstairs-McCarthy, Andrew 1994. Morphology without word-internal constituents A review of Anderson (1992). In *Yearbook of Morphology 1993,* eds. Geert Booij, and Jaap van Marle, 209-234. Dordrecht: Kluwer.
Chomsky, Noam 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
Comrie, Bernard. 2001. Different views of language typology. In *Language Typology and Language Universals,* eds. Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, Vol 1, 25-39 Berlin: De Gruyter.
Corbett, Greville G. 1987. The morphology/syntax interface. *Language* 63: 299-345.

Dahl, Östen 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam / Philadelphia: John Benjamins.
Greenberg, Joseph 1963. Some universals of grammar with particular reference to the order of meaningful elements. In *Universals of Language*, ed. Joseph Greenberg, 73-113. Cambridge Mass. MIT Pres [1966[2]].
Gurevich, Olga I. 2006. *Constructional Morphology. The Georgian Version.* Ph. D. diss, Univ. of California at Berkeley.
Hawkins, John A., and Anne Cutler 1988. Psycholinguistic factors in morphological asymmetry. In *Explaining Language Universals,* ed. John Hawkins, 280-305. Oxford: Blackwell.
Harris, Alice 2000. Where in the word is the Udi clitic? *Language* 76: 593-616.
Harris, Alice 2004. The challenge of typological unusual structures. In *Morphology and Linguistic Typology. On-line Proceedings of the Fourth Mediterranean Morphology Meeting, Catania, 21-23 September 2004,* eds. Geert Booij, Emiliano Guevara, Angela Ralli, and Salvatore Sgroi. http://mmm.lingue.unibo.it/proc-mmm4.php
Harris, Alice 2006a. In other words: External modifiers in Georgian. *Morphology* 16: 205-229.
Harris, Alice 2006b. Revisiting anaphoric islands. *Language* 82: 114-130.
Haspelmath, Martin 2006. Creating economical morphosyntactic patterns in language change. To appear in *Language Universals and Language Change*, ed. Jeff Good. Oxford: Oxford University Press.
Haspelmath, Martin 2007. Explaining some universals of causative verb formation. Lecture, LSA meeting Anaheim.
Helmbrecht, Johannes 2004. Cross-linguistic generalizations and their explanations. In *Morphology. An International Handbook of Inflection and Word Formation. Vol 2.* eds. Geert Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopeteas, 1247-1254. Berlin : De Gruyter.
Henderson, John 2002. The word in Eastern / Central Arrernte. In *Word. A Cross-linguistic Typology*, eds. R. M. W. Dixon and Alexandra Y. Aikhenvald, 100-124. Cambridge: Cambridge University Press.
Kiefer, Ferenc 1992. Compounding in Hungarian. *Rivista di Linguistica* 4: 61-78.
Lieber, Rochelle, and Sergio Scalise 2006. The Lexical Integrity Hypothesis in a new theoretical universe. *Lingue e Linguaggio* 6: 7-32.
Loey, A. van 1959. *Schönfelds Historische Grammatica van het Nederlands.* Zutphen: Thieme.
Michaelis, Laura A., and Knud Lambrecht 1996. Toward a construction-based theory of language function: The case of nominal extraposition. *Language* 72: 215-247.
Samvelian, P. 2006. Pronominal clitics in Sorani (Central) Kurdish Dialects. Paper presented at the 5th Décembrettes Morphology Meeting, Université de Toulouse, 7-8 December 2006.
Scalise, Sergio (ed.) 1992. *The Morphology of Compounding*. *Rivista di Linguistica* 4: 1-251.
Schreuder, Robert and Harald Baayen 1997. How complex simplex words can be. *Journal of Memory and Language* 37: 118-139.
Spencer, Andrew 1988. "Bracketing paradoxes" and the English lexicon. *Language* 64: 663-682.
Vigario, Marina 2003. *The Prosodic Word in European Portuguese*. Berlin: Mouton de Gruyter.