Indirect measurement in morphological typology

Bernhard Wälchli

1. Introduction

Languages differ in how much information is packed in a word.\(^1\) Consider example (1) where English uses four words for what is expressed in one word in Finnish. Finnish is a more synthetic language than English. In maximally analytic word structure, a word expresses only one lexical or one grammatical meaning at a time. Accordingly, analytic languages have many function words, words expressing grammatical meaning only, such as prepositions and pronouns. In synthetic word structure, a word expresses several grammatical meanings along with its lexical meaning. Grammatical morphemes such as case and possessive affixes abound as in (1b) where there is an illative case marker and a possessive suffix for the third person.

1.\(^\text{English as a more analytic and Finnish as a more synthetic language}\)
   a. into his own country
   b. koti-kaupunki-i-nsa

Morphological typology is a traditional field within linguistic typology concerned with assessing the degree of cross-linguistic variation in morphology – the internal structure of words. As the name “typology” suggests, the original idea was that languages can be classified into a small set of neatly discrete types. Accordingly, there would be, for instance, two types: synthetic and analytic languages, or three types: polysynthetic, synthetic and analytic languages.

A major progress in morphological typology was made in 1954 when Joseph Greenberg – notably in a contribution to a festschrift – suggested that synthesis and other features of morphology should be measured on a continuous scale as synthesis is a matter of degree rather than a set of discrete types. The method he proposed for synthesis was to count the number of morphemes per number of word-forms in a continuous text. A morpheme is the smallest meaning-bearing unit of a word. In (1b) the Finnish word-form is segmented into four morphemes which are separated by hyphens both in the form and in the analysis as it is common in interlinear
glossing. However, morphologists do not agree whether the morpheme is an ideal solution to the analysis of word structure in all cases, especially not for inflectional morphology (see Stump 2001: 1–3). Take the example of the French word /ɛ/ <ai> ‘(I) have’ which consists of a single phoneme whose form can be hardly further segmented into parts. However, it is grammatically highly complex since it is a carrier of the grammatical meanings ‘1st person’, ‘singular’, ‘present tense’, ‘indicative mood’ and ‘active voice’ along with its lexical meaning ‘have’ (if not used as an auxiliary). If the problem of morphological segmentation is left aside – an alternative is to count the number of grammatical meanings expressed per word-form – a major disadvantage of Greenberg’s method of measuring synthesis is that it requires a high proficiency on the side of the person applying it: it can be applied only by language experts and is very labor-intensive. Not astonishingly, therefore, Greenberg’s method did not meet with much practical response.

If it is not possible to measure what one wants to measure directly, it can be attempted to measure something else which is easier to measure and where there is good reason to assume that there is a strong correlation with what has been aimed at originally. This is indirect measurement. Indirect measurement is practiced throughout science as a method to make the impossible possible and to make science less costly. Dendrochronologists measure time by counting rings in wood, assuming that every ring is a year and that each year features characteristic properties in the corresponding ring. Trigonometry allows us to measure the height of mountains without having to climb them. Mainstream astronomers hold that the red shift in the spectrum of a galaxy correlates with its distance.

Indirect methods of measurement are also widespread in most different branches of linguistics. In psycholinguistics, for instance, reaction time is considered to be an index of speed of processing (see also Vorwerg, this volume). Formal semanticists approach meaning indirectly by considering exclusively under which conditions utterances are true or false. In morphology, the best-known approach to indirect measurement is probably Harald Baayen’s proposal to measure productivity of derivational affixes in terms of hapax legomena (words occurring only once) per token frequency of the words generated by a derivational process. Neologisms are assumed to occur only once, established “old” words are assumed to be more frequent on average (see, for instance, Baayen and Lieber 1991 and Bauer 2001: 145–162 for a critical discussion). A widespread indirect approach in typology is the semantic map approach where semantic similarity is ac-
This chapter explores to what extent indirect methods of measurement might be useful in morphological typology. For measuring the degree of synthesis it is profitable to take samples from different languages with highly similar content and highly similar length written in the same kind of register and style in the same genre. This is approximated in parallel texts (Cysouw and Wälchli 2007). A parallel text of substantial size which is available electronically in many languages from all continents is the New Testament, from which example (1) is taken (Mark 6:1). In this context it is important to note that it is impossible to address language directly. All what can be considered in empirical linguistics are *doculects*, documented language varieties, be it in form of secondary data (such as reference grammars) or primary data (such as texts, as used in this paper).

The same content in a more analytic language is packed into more word-form tokens than in a synthetic language. (Word-form tokens are all instances of word-forms in a text; types are the set of unique word-forms in a text. The first sentence of this paragraph contains two tokens of the type *in.*) In the short example in (1) there are four word-forms in English as opposed to one in Finnish. However, more compact synthetic word-forms are less likely to be recurrent in a text than analytical function words. The word-form type *kotikaupunkiinsa* occurs only four times in the Finnish N.T., whereas *into* occurs 604 times in the American Standard translation. In parallel texts, therefore, more word-form tokens mean a higher degree of analysis and more word-form types mean a higher degree of synthesis. Thus, a good measure for the degree of synthesis is the *type-token ratio*. This will be further discussed in Section 2 below.

What is important to note here is that for measuring degree of synthesis it is not necessary to segment any single word-form into morphemes, the indirect approach has thus the further advantage that it is compatible also with approaches to morphological theory rejecting the notion of inflectional morpheme (Stump 2001; Aronoff 1994). Furthermore, the method is cheap (once there are parallel texts that are freely available). It can be applied without the help of any language experts.

This paper is organized as follows. Section 2 sketches a dynamic model of morphology with processing chains with increasing complexity of the representation of inner word structure based on the notion of procedural universals. Section 3 considers the indirect measurement of synthesis in parallel texts and some problems associated with it. Sections 4 and 5 ad-
dress the question as to how indirect methods can be combined with partial automatic analysis as outlined in Section 2 to become more effective. Finally Section 6 concludes this paper.

2. The text-to-device approach, algorithmic morphology, procedural universals and three levels of analysis

In this and the following sections three levels of morphological analysis will be addressed. On level zero, word-forms are considered indivisible units and their internal structure is only approached indirectly by considering their distribution across texts. On level one, only continuous segments are recognized (morphemes as occurring in a chain of strings: prefixes, stems, and suffixes) which do not exhaust the diversity of known morphological processes. Level two additionally recognizes non-concatenative processes, such as infix and ablaut. A text-to-device approach is applied, i.e. the word-form is taken as the basic unit and the internal structure is deduced by subsequent processes, rather than a device-to-text approach. The notion “device” is used here as a theory-neutral term for any kind of paradigmatic organization of language, such as “langue”, “grammar”, “competence”, “mental representation”, or “the language system”.

The text-to-device approach has the following advantages:

(a) It is accessible for computational modeling cross-linguistically, since the starting point – text – is of the same kind for all languages and easily available without previous analysis.

(b) It is more useful for descriptive linguistics and typology, since no language-specific entities must be posited to start with and the same process of analysis can be applied in the same way to material from all languages (called “man-from-Mars attitude” by Nida 1949: 1 which simply means that the linguist or the device performing the analysis does not know anything about the language to start with).

(c) It is empirical since it proceeds from given to inferred units.

The basic assumption is that the internal structure of words can be inferred from any text of sufficient length in any language by a universal algorithm, given relevant semantic cues. The approach is thus necessarily massively cross-linguistic. Only by applying the method to corpora of different languages can it be assured that the model is not biased to languages of a certain type.
Different from other approaches it is not assumed that there is a single correct morphological analysis, but rather that analysis is a processing chain from less to more abstract representations. On level zero word-forms are considered as wholes without further subdividing them into parts. Next, the first level is concatenative morphology which only recognizes continuous morphemes of three kinds: stems, prefixes and suffixes. Higher levels, then, add more complex morphological processes, such as ablaut, infixes, and partial reduplication, which are more difficult to identify and less widespread cross-linguistically. One reason for the multi-layered model is that higher-level processes are reminiscent of lower-level processes. Ablaut can more easily be addressed in terms of sets of stems where affixes have been stripped away rather than in word-forms. Infixes tend to be edge-oriented (Yu 2007: 3), which makes them similar to affixes, and partial reduplication is usually reminiscent of prefixing or suffixing.

While the underlying assumption of structural universals is that there is a certain constant structure in all languages, procedural universals are universally applicable algorithms which extract highly different structures when confronted with different input. Algorithms may serve both for the acquisition of linguistic categories from corpora (learning) and for measuring cross-linguistic variation in texts (typology).

A simple example of a universal in morphological typology is that any language with non-concatenative morphology (e.g., infixes, ablaut) also has concatenative morphology. Translated into an algorithm, this means that the first step is to identify stems, prefixes and suffixes and, if there are any such structures, a second step may further proceed to look for internal inflection, stem alternations, etc. First all forms likely to have the same lexical meaning on the basis of parallel distribution are extracted (a primitive universally applicable lemmatizer). Next stems defined as shared sequences are isolated and affixes are whatever is left over. Once this simple analysis has been performed, one can proceed to look for more complex morphological processes, such as internal inflection.

3. **Level zero: word-forms, and degree of synthesis**

Degree of synthesis can be assessed by various measures deriving from counting types and tokens in parallel texts. It is not the aim of this section to identify the best measure but rather to illustrate what happens when syn-
thesis is accounted for in terms of types and tokens, and what kinds of practical problems may arise when doing so.

A good starting point is a type-token diagram of parallel text doculects. Figure 1 displays the number of word-form types and tokens in the Gospel according to Mark in a strongly biased convenience sample of 168 languages from 46 language families from all continents. Each dot is a doculect. It is expected that synthesis will correspond to high type frequency and low token frequency (bottom right) and analysis to low type and high token frequency (top left). If first some languages from Greenberg’s morpheme per word ratio counts are considered – Greenlandic [kal] 3.72, Swahili [swa] 2.55, German [deu] 1.92, English [eng] 1.68, and Vietnamese [vie] 1.06 – these are found to be arranged in a scale from bottom-right to top-left in Figure 1 as expected.

However, a general observation is that most doculects strive toward a medium degree of synthesis while highly synthetic and analytic languages are the exception (see also Bickel and Nichols 2005). This might be a general problem in ranking two languages on a synthesis scale. While it is clear beyond any doubt that Vietnamese is more analytic than Greenlandic, for many other language pairs – such as Cherokee [chr] and Turkish [trk] or German and Italian [ita] – the difference might just be too subtle to be significant. It is thus easier to treat synthesis in a small sample containing extreme languages such as Vietnamese and Greenlandic than in a larger convenience sample.

No linear regression analysis need be applied to Figure 1 to see that there is a strong inverse correlation between type and token frequency as expected. However, the correlation is not as good as it could be. Ideally, all doculects would be arranged on one line; and this line could then be interpreted as the constant amount of information in the single text translated to different languages with different morphological types.

A clear outlier is STR which is lemmatized Classical Greek (Strong’s numbers, see Dahl 2007). Lemmatizing means stripping away all inflectional information. All word-forms are replaced by citation forms of their lexemes or by numbers pointing to a list of lexemes. Accordingly, STR has lost all of the grammatical information contained in Greek inflectional morphology. It has the token frequency of a synthetic language (actually exactly the same as Classical Greek [grc]), but at the same time the low type frequency of an analytical language. However, STR is not the only data point disturbing the inverse correlation. Basically there are four major
kinds of reasons for the deviations. They are briefly introduced here and further discussed below:

(i) What is measured is not languages, but languages written in particular orthographies. Deviations result from the fact that these orthographies do not follow the same principles ("word" viewed as the string between two spaces is not the same kind of unit in all texts).

(ii) Translation does not lead to full equivalence of texts. There are more elaborate translations containing "more information" and these are closer to the top-right corner.

(iii) It is wrong to believe that the same text in different languages contains exactly the same amount of overt structural information. Situations
where “the same” grammatical information is always there — such as case and possession in function words as in (1a) and in morphemes as in (1b) — is the exception rather than the rule. Languages differ in the frequency and kind of grammatical meanings that have to be marked overtly.

(iv) Considering type frequency in a parallel text as a measure of degree of synthesis rests on the assumption that the number of lexemes is cross-linguistically constant whereas languages only differ in the number and kind of inflectional categories. However, the lexicon is structured differently across languages and this disturbs the picture.

There is clear evidence for all four sets of deviations. Put differently, types and tokens do measure degree of synthesis, but at the same time they measure a bunch of other things as well. Let us discuss some examples.

(i) Orthography: Southeastern Tepehuan [stp] and Northern Tepehuan [ntp] (Uto-Aztecan; Mexico) are closely related, but the latter is considerably more synthetic (Bascom 1982). However, Figure 1 grossly distorts the difference because in the SE Tepehuan translation inflectional verbal prefixes are written as words which clearly deviates from the practice of virtually all other orthographies (and from the orthography in the grammar by Willett 1991).

Vietnamese [vie] is known as a language without any inflectional morphology, thus the question arises as to why there are a couple of doculects with less types in Figure 1. Of these Lahu [lhu] (Tibeto-Burman; Myanmar) is the most extreme case. Lahu actually has a small number of suffixes (Matisoff 1982: 16; e.g., -ô ‘change of state’ and -e ‘directional’), so it is in a way more synthetic than Vietnamese. But it is written in a writing system where every syllable is a word (and there is good evidence that this is actually the phonological word in Lahu). Thus many English concepts correspond to long sequences of Lahu orthographic words (e.g., là’-še ʒ qhɔ lo chí à’ ‘hand shaft NPREF inside LOC rise deliver > deliver’; note that the dictionary by Matisoff 1988 considers là’-še, ʒ-qhɔ and chí-à as compounds, reducing the number of words in this example from seven to four). Add to this that Lahu has a very restricted phonotactic inventory of possible syllables. The label CAQ in Figure 1 stands for Car Nicobarese [caq] divided into syllables. In Car Nicobarese orthography all syllables are marked by hyphens. Thus, Car syllables may serve as a baseline for localizing syllables instead of words in the diagram. Figure 1 shows that Lahu is even below that baseline.

In Maltese the definite article is separated by a hyphen from following nouns and, similarly as in Italian and French, fuses with some preceding
prepositions. Maltese articles and prepositions are clitics (syntactic words prosodically dependent on a host). Clitics are intermediate between word-forms and morphemes and are treated very differently across orthographies in different languages. Since the Maltese definite article is frequent and fuses with several frequent prepositions, it makes a big difference for synthesis whether articles are considered words [MLT] or morphemes [mlt], especially in terms of ranking order, since there are many doculects between MLT and mlt in Figure 1. If the Maltese hyphen is replaced by space and articles thus are counted as word-forms of their own, Maltese is much more analytic than in standard orthography.

Tepehuan, Lahu, and Maltese are just three examples for different kinds of problems with orthography. Now it might be argued that all these problems would disappear if all texts were written in a consistent way with grammatical words separated as in spoken languages. However, there is nothing separating words in spoken language, so that for every language many decisions must be taken by linguists, and especially for clitics and compounds there is no agreement how to treat them. As Haspelmath summarizes (2002: 162) the relevant discussion in a morphology textbook “Clearly, morphology and syntax are different, but the question of whether the difference is minor or gradual or major and sharp will probably be debated for a long time to come.” Strictly speaking, measuring synthesis presupposes that this debate has already been solved. Nevertheless, measuring synthesis can contribute to that debate by showing where the big differences are when different solutions are adopted.

(ii) Translationese: As pointed out by de Vries (2007: 154) many Bible translations for minority languages after the Second World War (especially in New Guinea, Australia and South America) – with their missionary purpose – are meant as stand-alone texts. Accordingly they contain many explicative elements which make the texts longer. An example for a particularly elaborate translation is Yanesha' [ame] (Arawakan; Peru). In Figure 1 all South American [S], Papuan [P], Australian [A], and Mesoamerican [M] doculects appear with one letter labels if not explicitly labeled with three-letter codes. From that it becomes clear that the inverse correlation of type and token frequency would be considerably stronger if the Oceanian and South American doculects were removed. Whether the effect is due entirely to translation style or partly derives from continental macro-areal structural differences cannot be determined here. Note that one of the major advantages of the Bible translation is that languages from these continents can be included at all in such a parallel text study.
(iii) **Different grammatical categories**: Example (1) suggests that Finnish and English express the same grammatical categories case and possession, which are only realized differently, i.e. either morphologically or syntactically. However, languages greatly differ in the kind and frequency of grammatical categories expressed. This effect can be best exemplified with articles which have been extended in use to general noun phrase marker. In the Mayan languages Central Cakchiquel [cak], Eastern Cakchiquel [cke] and Western Tzutujil [tzt] the article *ri, ri, ja* accounts for a large number of tokens (3815, 3758, 3040, respectively, as compared to 875 tokens for *the* in English). This is not the whole explanation why these Mayan doculects have an exceptionally high token frequency, but it explains part of the deviation. Tagalog [tgl] (Austronesian; Philippines), for instance, has highly grammaticalized noun phrase markers as well (with 2967 tokens for *sa* ‘oblique’, *ang* ‘topic’, *nay* ‘non-topic’, *si* ‘person name topic’, and *ni* ‘person name non-topic’ in total), but is much more in line with European doculects than the Mayan languages. A further example is Wik Mungkan [wim] with a set of frequent discourse clitics without any obvious counterpart in other languages. On the other hand, there are East and South East Asian languages, such as Mandarin Chinese [cmn] and Vietnamese [vie], which are known for an exceptionally low frequency of their function words given their analytic character. This moves those doculects toward the bottom-left corner of Figure 1.

(iv) **Inventory of lexemes**: Lemmatized Classical Greek [STR: Strong’s numbers] can be considered a baseline for the average number of lemmata to be expected. Interestingly, some doculects have fewer word-forms than Greek has lexemes, and not only Lahu [lhu] with syllable-type orthography. A restricted lexical inventory has to be expected especially for creoles such as Haitian Creole [hat]. Another potential reason for exceptionally small inventories is extensive homonymy in languages with small phoneme inventories as it is characteristic for Oceanic languages, such as Hawaiian [haw] and Maori [mri]. The effect is enforced by orthography where vowel length and glottal stop are not usually marked.

Finally, it is easily understandable from Figure 1 that the type-token ratio is the better measure for degree of synthesis than either type or token frequency in isolation. Token frequency is highly sensitive to the frequency of the most frequent type(s) and this is highly variable cross-linguistically. However, in considering type-token ratio it is taken for granted that there is a strict inverse correlation between number of types and number of tokens in a parallel text, which holds true only as a tendency.
Of course, there are many more methods of indirect measurement of degree of synthesis (none of them working well anywhere but in parallel texts). Juola (2008) uses the ratio between zipped and uncompressed text size and Popescu et al. (2009) use trigonometry in type-token diagrams.

4. Level one: Prefixing and suffixing, and refracting synthesis

4.1. Concatenative morphology

Considering morphemes as basic units of internal word structure presupposes that there is a one-to-one relationship between lexical and grammatical meaning components of word-forms and strings of phonemes. However, even though it is not always possible to segment word-forms into lexical and inflectional substrings, lexical and grammatical meanings are certainly not randomly distributed across the phonemes of word-forms: there is a high degree of iconicity in all inflectional languages in the sense that different segments within words express different components of meaning.

Concatenative morphology is a simplifying model of word structure assuming full isomorphism between form and function. Each phoneme is either part of a lexical or of an inflectional string. Suprasegmental phonology is disregarded. Further, it is assumed that there is always exactly one lexical string in each word-form (the stem) and that grammatical strings (if any) precede (prefix) or follow (suffix) the stem. In this model, synthesis can now be refracted into two components: prefixation and suffixation.

Even though it is not fully accurate to account for word structure, concatenative morphology has the advantage that it underlies heavy constraints. It can more easily be cracked by “men-from-Mars”, put differently, it is useful for indirect measurement – and possibly as a first step of analysis in an algorithmic processing chain to analyze morphology in all of its relevant features. An advantage is that if it is known which part is lexical (the stem), the inflectional parts (affixes) are obtained automatically by subtracting the stem from the word-form (and vice versa, if the affixes are known, the stem can be obtained for free).

In compiling indices for affixation there are two options: start identifying grammatical affixes directly (as done in Dryer’s 2005a approach based on reference grammars) or first identify the stems. Proceeding by way of the lexical strings has several advantages: (i) they are subject to stronger constraints: each word-form can be expected to contain exactly one stem;
(ii) they can more easily be accessed by way of distribution across languages since the number of corresponding grammatical forms in a grammar domain is subject to much more variation on average than the number of lexemes per lexeme domain; and (iii) stems are subject to less formal variation (complete allomorphy and suppletion) than grammatical markers, and can thus more easily be identified by simply finding identical strings in all word-forms used to express a particular category. What will be discussed in the following section is a general algorithm how prefixation and suffixation can be measured in parallel texts if the lexeme domains in at least one language are known.

4.2. Global indices for prefixation and suffixation

Affixation can be measured as a by-product of automatic lemmatization in parallel texts. This section gives a summary of the four subsequent steps in the procedure of measuring prefixing and suffixing indirectly. The mathematical details of the concrete algorithm implemented in the Python program extracting the affixes are glossed over. The four major steps are the following: (i) extraction of forms in a lexeme domain, (ii) grouping of extracted forms into sets with a shared stem, (iii) detection of recurrent prefixes and suffixes, and (iv) calculating typological indices.

A lexeme domain is the set of contexts which are expected to be expressed by forms of a lexeme (and hence differing only in grammatical morphology) in any language. A practical problem is that lexeme domains are not congruent cross-linguistically. However, an astonishingly good approximation to lexeme domains is the distribution of lexemes in one particular language used as trigger for identifying lexeme domains of other languages in parallel texts. The effect of the trigger language will be discussed below.

An easy approach to implement step (i) goes as follows: (a) Assume that lexical domains are always expressed by one word-form only. This is mostly the case, but some phrasal expressions and cases of multiple exponence are lost, such as French ne...pas for negation (the algorithm will find either pas or ne), (b) Find the best equivalent for the lexeme domain as defined in terms of the set of contexts where it occurs by means of a collocation measure, such as Dice, t-score or log-likelihood (see, e.g., Manning and Schütze 1999: 151–189; Dahl 2007; Wälchli 2011) and add this form to the set of extracted word-form types, (c) Subtract the contexts accounted for by the extracted form from the lexeme domain and repeat (b) with that smaller
distribution while prioritizing forms which are phonologically similar to those already in the set of extracted forms. Apply (c) recursively until the collocation measure falls under a certain threshold where it is unlikely that the best form should be considered a reliable member of the set.

Table 1 lists the forms that are extracted in four doculects in the lexical domain defined by the Classical Greek lemma ἄνθρωπος by recursively finding the best equivalent as long as the measure does not fall below a certain threshold. In Table 1 step (ii) has already been applied: the forms are grouped according to the most likely candidates for stems (italics). Candidates for prefixes and suffixes are then all strings left if the stems are subtracted (not in italics in Table 1) and these are counted as affixes if they are recurrent, i.e. occur with more than one stem across all the lexeme domains surveyed. This step can be shown in Table 1 only with English -’s which is a recurrent suffix with the two stems man and men.

Table 1. Forms extracted with the Greek lemma ἄνθρωπος (Strong’s number 444, 553 tokens in N.T.)

<table>
<thead>
<tr>
<th>ENGLISH</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>man: man</td>
<td>336</td>
<td>[tokens]</td>
<td>man’s 9</td>
</tr>
<tr>
<td>men: men</td>
<td>186</td>
<td>men’s 3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HUNGARIAN</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ember: ember</td>
<td>170</td>
<td>embernek 104</td>
<td>emberek 85</td>
</tr>
<tr>
<td></td>
<td>embert 40</td>
<td>emberekkel 5</td>
<td>embereknél 5</td>
</tr>
<tr>
<td></td>
<td>emberböl 6</td>
<td>emberekhez 3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ZULU</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ntu: umuntu</td>
<td>114</td>
<td>yomuntu 89</td>
<td>muntu 64</td>
</tr>
<tr>
<td></td>
<td>abantu 62</td>
<td>kubantu 38</td>
<td>kwa-bantu 20</td>
</tr>
<tr>
<td></td>
<td>bantu 11</td>
<td>yabantu 9</td>
<td>ngabantu 12</td>
</tr>
<tr>
<td></td>
<td>kumuntu 11</td>
<td>nomuntu 9</td>
<td>bantu 7</td>
</tr>
<tr>
<td></td>
<td>okwabantu 4</td>
<td>kunabantu 5</td>
<td>zabantu 8</td>
</tr>
<tr>
<td></td>
<td>ngokwabantu 3</td>
<td>ngu-muntu 5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAORI</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>tangata: tangata</td>
<td>532</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There is no single unequivocal way to tell what the best stem candidates are given a set of forms with the same lexical meaning, especially since it is not known whether the forms all contain one stem or should rather be grouped into several stem sets. What is clear is that the extracted forms are highly similar in their lexical meaning. They need not belong to one lexeme though, but may rather belong to two near-synonyms. If making stem set grouping greedier than in Table 1 one could also obtain for English a single stem {m} or {n}, and if it were less greedy two stems {muntu} and {bantu} would be obtained in Zulu. In the present approach stem extraction is set at moderately greedy.11
The method for affix extraction sketched here is indirect as far as the function of the affixes is concerned. There is no way to find out with this method that Hungarian -t is accusative, -nek is dative, and -ek is plural and that -eknek is a complex suffix consisting of plural and dative. But it can be assumed for this lexeme domain – without consulting a Hungarian grammar – that Hungarian has a large number of inflectional suffixes, Zulu a large number of inflectional prefixes, English some very restricted suffixation, and Maori no inflection at all in this lexeme domain.

The simplest approach to calculate prefix and suffix indices is to count the number of recurrent left and right leftovers after stem subtraction in a set of the most frequent lexeme domains in a parallel text. However, it is preferable to make some effort to first exclude domains which are rather grammatical than lexical in character, i.e., function words. For Greek 175 Strong’s numbers have been selected that are lexical, occur at least five times in Mark, and yield good extraction on average. For Vietnamese simply the most frequent 200 word-forms are taken irrespective of whether or not they are function words (which is rather hard to determine in a language such as Vietnamese).

Unlike Greenberg the present algorithm does not count the number of affixes per word in continuous text. There is no way our simple method can determine the number of affixes per form, it can only determine whether or not a form has a prefix and/or a suffix. For calculating the suffix index, the
extracted forms are divided in three groups: $N$ or non-suffixed forms, $R$ or forms with recurrent suffixes (i.e., suffixes occurring at least with two stems), and $U$ unique suffixes. The suffix ratio is number of $R$ types divided by $R+N$ types while $U$ is ignored. The prefix index is calculated analogously.

The major difference to Greenberg’s indices is that languages with many forms with several suffixes or prefixes per word yield lower values, and that it is important how many unaffixed forms there are. In our approach, zero marking is counted as lack of affixing. This entails that languages with zero marked nominative singular, such as Turkish, get lower values for suffixing than inflectional Indo-European languages, such as Classical Greek [gre] and Lithuanian. By the same token, Italian gets higher values than Spanish because masculine singular is -o and not zero for most nouns.

Figure 2 (left) shows the results for Vietnamese as trigger and Figure 2 (right) connects the obtained values for Greek and Vietnamese by dotted lines with the labels added at the Greek side of the lines. Figure 2 (right) shows that using different triggers has consequences on the magnitude of affixes extracted, but the degree of prefixing as opposed to suffixes remains remarkably stable and is thus little dependent on the trigger language. The most heavily prefixing doculects in the sample are three Bantu languages from South Africa (Zulu, Xhosa and Ndebele). With Greek trigger, some Indo-European doculects yield the highest values for suffixing (Latin, Greek, Polish), mainly due to better extraction with a trigger with a similar lexical structure. However, these languages get high values even with Vietnamese as trigger. The top scorer for suffixing with Vietnamese trigger is Oromo [hae] (Afro-Asiatic, Eastern Cushitic; Ethiopia).

Finally, the type-token ratio from Section 3 is compared to the number of non-affixed forms in level one where prefixed and suffixed forms have been subtracted. For this purpose the proportion of non-affixed word-form tokens with Greek and Vietnamese trigger are taken which are shown on the $y$-axis in Figure 3 connected by lines with the language labels added at the Greek side of the lines. The $x$-axis is the type-token ratio from Section 2. Figure 3 shows that the measures strongly correlate and are all variations on one theme: degree of synthesis. The level one measure has the advantage that there is an absolute boundary point for no inflection at 1.0. This is where Vietnamese is located both with Vietnamese and Greek as trigger (for Vietnamese trigger 1.0 by definition).
Figure 3. Synthesis measured as type-token ratio (x-axis) and as proportion of non-affixed tokens with Greek and Vietnamese as trigger (y-axis)

4.3. Case

Accessing morphological structure through the prism of lexeme domains allows us to focus on lexical subdomains where only some types of grammatical categories are expected to be expressed in morphology. In nouns the major category types to be expected are number, case, possession and definiteness. However, in proper names, the only major category type according to which word-forms are expected to vary is case. (Most texts do not contrast my Peter to your Peter, one Peter to the Peter or Peter to Peters.)\(^{12}\) Proper names have the further advantage that their distribution in parallel texts tends to be more uniform across languages than that of appellatives. Extracting the correspondences for Greek Ἰωάννης ‘John’ in Mark in Hungarian with the method presented in 4.2 yields the following forms: jános\(^{13}\) 14 [tokens] jánost 5 jánosnak 3 jánostól 1 jánossal 1 jánosról 1 jánosra 1 suggesting that Hungarian has at least six overtly marked cases. For Swahili, only one form yohane 26 is extracted, suggesting that there are no cases marked on nouns in Swahili.\(^{14}\) Map 1 displays the different degrees of overtly case-marked tokens in proper names for a large convenience sample measured fully automatically in parallel texts. What is measured is not the number of cases, but the proportion of tokens with affixes across twelve typical proper name lexeme domains in Mark. Measured in this way, degree of case in Hungarian is relatively limited, since, in the
example above, 14 of 26 tokens lack case suffixes. The zero marking in the nominative János is counted as no case.

However, the result largely meets the typologist’s expectations. Case is almost always suffixing, there are very few languages with preposed case (mainly Bantu languages in Southern Africa; see also Dryer 2005b). Case mainly occurs in languages of Eastern Europe, Central Eurasia, Australia, and Peru. There is a large degree of overlap with the WALS map by Iggesen (2005) based on materials from reference grammars and counting number of cases. For a few doculects, however, things go wrong. Northern Tepehuan (Northern Mexico, dark square) is one of the rare instances where definiteness is marked on proper names as prefix and not consistently. In this language there are also complementizer clitic prefixes (that can occur on words of various word classes). The Celtic doculects behave inconsistently due to variable principles of treating onset consonant mutation in orthography (see 5 below).

Map 1. Number of case affix tokens (darkness) measured in proper names*
*circle: suffixes, triangle: prefixes, square: mixed pre- and suffixes, diamond: no affixes

5. Level two: Proceeding to non-concatenative morphology

Here three processes of non-concatenative morphology will be discussed: infixes, partial reduplication, and consonant mutation (for internal inflection and ablaut see Wälchli 2010). Non-concatenative processes have in common that they are more sensitive to phonology than prefixes and suffixes. However, in most cases, there is a simple phonological distinction that helps us go a very long way: the one between consonants and vowels, and there are many different approaches to determine the set of vowels and
consonants in texts automatically, the simplest one being Sukhotin’s (1962) algorithm.

**Infixes**, such as `<um>` in Tagalog `p<um>asok` ‘<ACT:TOP.PV>enter’’, are first extracted as parts of affixes (actually mostly prefixes as in this example). A concatenative stem `asok` is obtained as recurrent continuous sequence in a set such as `{pumasok, magsipasok, pagkapasok, nangag-sipasok, papasok, pinasok, makapasok, makapapapasok}` extracted in the lexeme domain `eisérchomai` ‘enter’. Next recurrent sequences are searched in all prefixes (and in all suffixes). They are likely to belong to the stem if they are recurrent in all prefixes, especially if they are (or contain) consonants.¹⁵ In the Tagalog form set, the sequence `p` recurs in all forms. It is therefore likely that this must be a discontinuous stem `p.asok` (p. stands for the sequence `p` that is located most closely to `asok` if the sequence occurs more than once in a first order prefix). The sequences `<um>` and `<in>` are therefore likely candidates for infixes. Of course, this procedure will extract only inflectional infixes.

**Inflectional partial reduplication** can be illustrated with the same set of examples. Like infixes, inflectional reduplication will first be treated as an affix on level one. Repeated consonants separated by vowels within stems look as if they were reduplication, but if there is no opposition to any other form in a lexical domain, this catchy sequence does not mean anything. It is just part of a funny looking stem (as in Latin `totus` ‘all’ or Classical Greek `laleo` ‘speak’). The following algorithm is proposed for the identification of partial reduplication:

- In all extracted forms look for sequences of `C_iV_x(V_x(V_x))C_i` where `C_i` is a consonant of the same type and `V_x` is any vowel.
- If a form contains such a sequence, check whether the sequence is fully contained in the stem. If yes, this is no inflectional reduplication.
- If no, check whether the prefix or suffix can be accounted for by a true prefix or suffix, one that occurs with all kinds of stems, also stems not sharing a consonant with the “affix”. This subtracts candidates, such as the frequent `t-et` and `n-en` in German with third singular and plural on `t-` and `n-ending verbs stems.`
- If not, this is a good candidate for partial reduplication.

In the Tagalog form set this extracts first `{papasok, makapapasok}` containing the sequence `pap`. This sequence is not contained in the first order stem `asok`, and neither are `pap-` or `makapap-` frequently recurring affixes across all stems. Hence this must be inflectional reduplication.

Once it has been discovered that inflectional reduplication is pervasive in Tagalog one can look for more intricate examples of reduplication, such
as reduplication sequences with infixes in them like \( p^{in}a \sim p \) or \( p^{um}a \sim p \). This can again be formulated as an implicative universal. If a language has reduplication sequences with infixes in them it also has reduplication sequences without infixes in them. As a procedural universal this means: start looking for discontinuous reduplication sequences only if there is good evidence that there are cases of continuous reduplication sequences.

The procedure for *consonant mutation*, as it is frequent in Celtic languages, is similar to discontinuous stems but without any intervening affixes. Taking the four Breton forms extracted in the lexical domain *adelphos* ‘brother’ \{breur, vreur, vreudeur, vreudeur\} yields a first level stem *reu*. The prefixes, however, are likely to contain some stem information, if they all end on (or entirely consist of) a consonant or consonant sequence. This applies especially if the consonant alternation – here *b:v* – is recurrent in different lexeme domains (also *bras:vras* ‘much’; *m:v* recurs in *mamm:vamm* ‘mother’, *mat:vat* ‘beautiful’, *menez:venez* ‘mountain’; in all these cases the first level stem has no initial consonant: *amm, at, enez*), and especially if these first order prefixes do not tend to occur with any other stems where they are not part of a consonant alternation.

These are just three examples of how one might further proceed to extract more complex morphological processes by starting out from a simplistic model of concatenative morphology. Since non-concatenative morphology is more dependent on phonological cues than first level morphology, the performance of the automatic analysis is increasingly dependent on phonological orthographies.

6. Conclusions

As shown already by Greenberg, it is not sufficient in morphological typology to classify languages into a small number of discrete types. Rather typology should engage in measuring features on continuous scales in texts. However, when done all manually, text typology is very costly in large samples of languages. Therefore advancement in morphological typology will be highly dependent on establishing indirect methods of measurement. These have the advantages that they can be applied fully automatically and that measurement is replicable.

It has been shown in this paper that morphological typology provides much potential for methods of indirect measurement. It has further been argued that indirect measurement in morphological typology is particularly
useful if it uses semi-analytic approaches where part of the word structure is analyzed automatically. Partial analysis is indispensable because it is difficult to interpret measurements otherwise, and it is indispensable to refract morphological complexity into its underlying subcomponents.

Analysis can be done more easily if there are some semantic cues based on distribution of lexeme domains. It has been shown that such distributional cues are available in parallel texts where at least one text is lemmatized (at the cost of some ethnocentrism entailing better extraction with languages structurally similar to the trigger language). Typology can thus be viewed as a by-product of morphological analysis in a text-to-device approach.

However, there are also many caveats that must dampen our enthusiasm for indirect measurement in morphological typology. The method is highly dependent on orthography (and an invitation to study the interaction of morphological modeling and writing systems more carefully). Parallel texts are an easy place to start, but in the future one should also start thinking about possibilities to use parallel texts as keys to original texts.

Indirect methods of measurement are certainly not the solution for all problems. But they are an important tool for different approaches to empirical linguistics. However, it is important to remain explicitly aware of their indirect character, which always entails that there is an underlying set of assumptions about strict correlations between the thing measured and the thing that cannot be measured. The results of indirect measurement only hold to the extent these correlations exist. The investigation of expected and unexpected correlations, in turn is highly revealing for a better understanding of the constraints restricting the diversity of human languages.

Notes

1. I am grateful to Andrea Ender, Adrian Leemann, Thomas Mayer and two anonymous reviewers for many useful suggestions. While writing this paper I was funded by the Swiss National Science Foundation (PP001-114840).
2. The examples are from the Gospel according to Mark 6:1. This is why there is country in the English text (American Standard translation) and ‘city’ in the Finnish text. ACT:TOP actor topic, 3 3rd person, ILL illative, M masculine, NPREF noun prefix, LOC locative, POSS possession, PV perfective, SG singular
3. For a recent renaissance of Greenberg-inspired measures in morphological typology see, for instance, Kortmann and Szmrecsanyi (2011).
4. The sample consists of the following doculects: Adyghe [ady], Afrikaans [afr], Albanian, Amele [aey], Ampeeli, Amuzgo (Guerrero), Apalai, Apurina, Armenian, Bana, Barai, Basque, Bicolano, Bimoba, Bora [boa], Breton [bre], Buki-yip, Burarra, Cacua, Cakchiquel (Central [cak] and Eastern [cke]), Car Nicobarese [caq], Cebuano [ceb], Chamorro, Chechen, Cherokee [chr], Chinantec (Lealao, Ozumacin, Quiotepec, and Sochiapan), Croatian, Czech, Danish, Dinka [dik], Duruma, Dutch, English (Early Modern [eng] and Middle), Estonian, Ewe, Finnish [fin], French [fra], Ful (Adamawa) [fub], Gagauz, Georgian, German (Standard [deu], Swabian, and Bernese), Greek (Classical) [grc], Greenlandic (Western) [kal], Guaraní, Guarani (Bolivian), Haitian Creole [hat], Hausa, Hawaiian [haw], Hindi, Hmong, Huave [huv], Huichol [hch], Hungarian [hun], Icelandic [isl], Indonesian, Italian [ita], Jacalteca, Jamaican [jam], Javanese, Jivaro, Kabyle, Kadi, Karo Batak, Khasi [kha], Korean [kor], Kriol, Kuna, Lahu, Latin [lat], Latvian, Lhaovo, Lithuanian [lit], Liv, Longuda, Low Saxon, Luo [lou], Ma'anyan, Makassar, Malayalam [mal], Maltese [mlt], Mam (Central) [mcm], Mandarin [cmn], Manx, Maori [mri], Mapudungun, Mataco, Mazatec [vym], Mentawai, Mixe (Coatlan), Mordvin (Erzya) [myv], Mouk-Aria, Nahua (Northern Puebla and Tetelcingo), Nakanai, Nalca, Ndebele [nbl], Ngalum, Norwegian, Oromo [hae], Ossetic, Otomi (Mezquital, Queretaro, and Tenango), Pamona, Papago [ood], Pintupi [piu], Polish [pol], Portuguese, Quechua (Cajamarca) [qvc], Romani, Romanian, Russian, Saami (Lule and Northern), Scots Gaelic, Secoya, Shipibo [shp], Slovene, Somali [som], Sougb, Spanish [spa], Sundanese [sun], Sutsilvan Romansh, Swhahili [swa], Swedish [swe], Tagalog [tgl], Tanggo, Tapiete, Tepehuan (Northern [ntp] and Southwestern [stp]), Ticuna, Tidjik, Timorese, Tobelo, Tol, Turkish [trk], Tuva, Tzutujil (Eastern [tzj] and Western [tzt]), Ukrainian, Uma, Ura-rina, Usarufa, Uzbek, Vietnamese [vie], Wayuu, Welsh [cym], Wix Mungen [wim], Wolof [wol], Xhosa [xho], Yagua, Yaminahua, Yanesh' [ame], Yine [pib], Zanniat, Zapotec (Ozoltepec and Quioquitani Quieri), Zarma, Zoque [zos], Zotung, Zulu [zul].

As can be seen from this list, the sample is strongly biased toward Europe, but also languages from Mexico and Peru are overrepresented. Languages from Africa, North America, Australia, and Russia are underrepresented.

5. All three letter codes are from the ISO 639-3 standard.

6. Greenberg’s index has one clear advantage over the type-token values: there is an absolute lower limit at 1.00 (there cannot be more word-forms than morphemes) which means “completely analytic”. There is no type-token value that means “completely analytic”.

7. A further problem not addressed in this contribution is that different parts of language structure can exhibit different degrees of synthesis. German has a more synthetic noun whereas Italian has a more synthetic verb.

8. These numbers are annotations added to a Bible text following a system devised by James Strong in the 19th century. Every number corresponds to a Greek lemma in the N.T. (Hebrew in the O.T.).
9. There are more obvious problems, such as French [fra] which makes much more morphological distinctions in orthography than in the spoken language.

10. A possibility without indices is selective measurement. Bickel and Nichols (2005) measure the inflectional synthesis of verbs by counting the categories in the maximally inflected verb form.

11. The best stem candidates are extracted recursively until all forms have been assigned to a stem. In the current version implemented here, the best candidate has the highest value for \( t \cdot l \) where \( t \) is the sum of all token frequencies plus one of all word-forms containing the form string of the candidate and \( l \) is the number of letters of the form-string. If \( t \cdot l \) is replaced by \( l \), length becomes a much more important factor and longer stem candidates are preferred, which makes the algorithm less greedy.

12. It may happen, however, that proper names, especially borrowed proper names, have a deviant behavior concerning case. An example is Greek *Mariam* which is not inflected (does not happen to occur in the Gospel according to Mark). However, in practice this effect has no major influence on the results.

13. These forms are not capitalized since the program treats all letters as lower case. Otherwise the same word-form at the beginning of a sentence would be a different type.

14. Actually Swahili has three local cases marked only on attributes. The method is too crude to discover those. Neither can case be detected if it is distinguished only in articles or any other separate word-forms.

15. Vowels are not distinctive enough; it may be mere coincidence if a vowel happen to recur in all prefixes.

References

Aronoff, Mark  

Baayen, Harald, and Rochelle Lieber  

Bascom, Burton  

Bauer, Laurie  
Bickel, Balthasar, and Johanna Nichols

Cysouw, Michael, and Bernhard Wälchli

Dahl, Östen

Dryer, Matthew

Greenberg, Joseph H.

Haiman, John

Haspelmath, Martin
2002 Understanding Morphology. London: Arnold

Iggesen, Oliver A.

Juola, Patrick
Kortmann, Bernd, and Benedikt Szmrecsanyi

Manning, Christopher D., and Hinrich Schütze

Matisoff, James A.

Nida, Eugene A.

Popescu, Ioan-Iovitz and Ján Mačutek, and Gabriel Altmann
2009 Aspects of Word Frequencies. Lüdenscheid: RAM.

Stump, Gregory

Sukhotin, Boris V.

Vorwerk, Constanze
this vol. Experimental methods in psycholinguistics.

Vries, Lourens de

Wälchli, Bernhard

Wälchli, Bernhard, and Michael Cysouw
2012 Lexical typology through similarity semantics: Toward a semantic map of motion verbs. Linguistics.

Willett, Thomas Leslie

Yu, Alan C. L.