1. Introduction: what is morphological complexity?

The paper is a prospectus for an European Research Council funded project\(^1\) (Corbett, Baerman and Brown), which aims to chart the limits of linguistic complexity by focusing on inflectional morphology. Quite apart from its inherent interest to linguists, we believe that it may present particular challenges for psycholinguistic and neurolinguistic investigation, and we will mention these in passing as we go through.

Our first purpose is to explain what we mean by morphological complexity. When most people hear the term *morphological complexity*, they will probably think of examples like the Turkish word in (1).

\[(1)\text{ alyorduysam} \]
\[
\text{al-iyor-du-isa-m} \]
\[
\text{take-CONTINUOUS-PST-CONDITIONAL-1SG} \]
\[
\text{‘if I was taking’} \]

This consists of a root plus four inflectional suffixes. If complexity is measured in terms of internal structure, then this word form is complex. People might also think of examples such as the Sanksrit noun in Figure 1:

```
NOM    dev-as
VOC    dev-a
ACC    dev-am
INS    dev-ena
DAT    dev-āya
ABL    dev-āt
GEN    dev-asya
LOC    dev-e
```

**Figure 1:** Sanskrit a-stem noun ‘god’, singular forms (Whitney 1889: 114-15)

Here the case paradigm is made up of eight distinct forms. If complexity is measured in the size of the paradigm, then this Sanskrit example is complex. But in both cases, the complexity of the morphology is a direct reflection of the complexity of the functions they express. The Turkish word form express functions that in another language might be expressed by separate words such as auxiliary verbs and pronouns, while the Sanskrit noun uses case endings to express functions which in another language might involve adpositions and word order. If we take these functions as a given (for these languages, that is), the morphology itself is actually quite simple; the complexity, such as it is, is a property of the linguistic system as a whole.

But there is another way to construe the term *morphological complexity*, and that is as the additional structure imposed by inflectional morphology, above and beyond its dedicated task of expressing syntactic and semantic distinctions.\(^2\) For example, take the Italian present

\(^1\) ‘Morphological Complexity’, grant number ERC-2008-AdG-230268 MORPHOLOGY.
\(^2\) Dressler (1999) construes complexity in a related but nevertheless distinct fashion He contrasts morphological *richness* with *complexity*, where richness is assessed over all the morphological patterns of the language, both
indicative forms in Figure 2. The paradigm distinguishes six morphosyntactic values: first singular, second singular and so on through third plural. If Italian inflectional morphology were merely an expression of these values, we would expect six morphological markers, and these would serve for every verb in the language. But of course this is not the case; for example, the endings of ‘talk’ and ‘sleep’ are different in the 3SG, 2PL and 3PL. The verb ‘finish’ has the same endings as ‘sleep’, but with the addition of a stem extension (orthographically -isc-) whose presence or absence divides the paradigm into two chunks, singular and 3PL versus 1PL and 2PL.

<table>
<thead>
<tr>
<th></th>
<th>‘talk’</th>
<th>‘sleep’</th>
<th>‘finish’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG</td>
<td>parl-o</td>
<td>dorm-o</td>
<td>fin-isc-o</td>
</tr>
<tr>
<td>2SG</td>
<td>parl-i</td>
<td>dorm-i</td>
<td>fin-isc-i</td>
</tr>
<tr>
<td>3SG</td>
<td>parl-a</td>
<td>dorm-e</td>
<td>fin-isc-e</td>
</tr>
<tr>
<td>1PL</td>
<td>parl-iamo</td>
<td>dorm-iamo</td>
<td>fin-iamo</td>
</tr>
<tr>
<td>2PL</td>
<td>parl-ate</td>
<td>dorm-ite</td>
<td>fin-ite</td>
</tr>
<tr>
<td>3PL</td>
<td>parl-ano</td>
<td>dorm-ono</td>
<td>fin-isc-ono</td>
</tr>
</tbody>
</table>

Figure 2: Some Italian verbs of different classes

While such patterns might be familiar, they should continue to surprise us. These patterns have no obvious correspondence in meaning or function: that verbal endings have varying shapes, and that the 1PL and 2PL behave differently from the rest of the paradigm, are facts which the morphological system introduces in its own right: gratuitous complexity. Nonetheless they are central organizing principles in the morphologies of the languages which manifest them (see Maiden 2003). This poses a particular challenge for our models of language. Other aspects of language, such as syntax, semantics or pragmatics, are commonly thought of as universal, in that they are relevant to all languages and subject to shared constraints (though of course the explanation for those constraints varies!), and these constraints reveal something about cognition, biology and social interaction. The position of purely morphological patterning in this larger picture is far from clear, which all too often leads to its being ignored or downplayed.

We would argue that autonomous morphological complexity - however idiosyncratic, language specific and bizarre it might be - is an element of language worthy of investigation in its own right. It is produced and learned by the human mind and governs the shape of many a language. This has implications for the architecture of grammar (where exactly in the system is morphological complexity housed and what, if anything, constrains it?), psycholinguistics (how can it be learned, and how does it affect production and processing?) and computational linguistics (how can automated processes deal with it, and what in turn does this tell us about human cognitive processes?). Below in §2 we outline a typology of the sorts of phenomena we believe constitute autonomous morphological complexity, and in §3 lay out some of the areas in which morphological complexity may be a fruitful topic of investigation.

2. Typology of morphological complexity

We start then from the baseline comparison where we have items which have simple expression of both lexical and grammatical meaning. The lexical meaning will differ for each item (“dog” and “cow” will be expressed differently), and the grammatical meaning could be productive and unproductive, while complexity is measured only in terms of the unproductive patterns. Our use of the term complexity makes no direct reference to productivity: while the presence of unproductive patterns implies complexity, we take no a priori stance on the reverse implication.
realised in a consistent way. Suppose there is a case marker to indicate possession, say -a, we could find dog-a ‘of the dog’ and cow-a ‘of the cow’. Such a system would be perfect in functional terms, achieving the communicative requirement with minimal phonological material.

This simple situation may be complicated in different ways. A helpful way to think of it is in terms of where the additional, complicating information must be stated. In broad terms we may need two types of additional specification (a) inflectional classes, and (b) form classes. We explain and exemplify these in turn.

2.1 Inflectional classes

Staying with our simple example, suppose that instead of all items marking the genitive in -a there are two possibilities, namely -a and -i. At first this appears to be a small increase in complexity of the morphological possibilities. The real increase, however, lies elsewhere. We now have to indicate which items take -a and which take -i. This gives us two inflectional classes, consisting of the items which take -a and those which take -i.

From the point of view of the lexical item, the choice of affix may be motivated or unmotivated. We may find semantic motivation, for example: nouns denoting humans take -a and others take -i. We might find phonological motivation: nouns ending in plosives take -a and others take -i. The more complex situation is that in which we find no such extra-morphological motivation, and that we have to stipulate, item by item, which affix is appropriate. In this situation the additional complexity appears to contribute nothing.

Returning to our rudimentary morphological system, let us expand it to include a further opposition. Let us introduce a dative, marked by two alternative affixes, -e and -o. The choice might be motivated, and in the simplest case according to the same factors as with -a and -i.

But what if we have a system in which there are two choices, two sets of allomorphs, namely -a versus -i and -e versus -o, and there is no semantic or morphological motivation? The smallest additional complication would then be the situation in which the choice of one allomorph was sufficient to predict the choice of the other. For instance, those items which are listed as taking the genitive in -a are exactly those which take the dative in -e. There are then just two possibilities:

<table>
<thead>
<tr>
<th></th>
<th>class A</th>
<th>class B</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENITIVE</td>
<td>-a</td>
<td>-i</td>
</tr>
<tr>
<td>DATIVE</td>
<td>-e</td>
<td>-o</td>
</tr>
</tbody>
</table>

Figure 3: Two hypothetical paradigms

We thus divide the lexical items (in this instance, nouns) into two classes, A and B, each taking a different set of inflections. Within those two classes there is consistent inflectional behaviour. Thinking of it another way, for a given lexical item, the choice for one specification predicts the choice for the other. It is time to illustrate with a real example.

Consider these data from Burmeso, a language of the Mamberamo River area of Western New Guinea, from Donohue (2001). Burmeso was previously considered an isolate, but Ross (2005) suggests a possible affiliation on the basis of the pronouns. Among several interesting characteristics, Burmeso shows two classifications of nouns. We consider just the first classification, which involves six gender values.3 Evidence for the classification comes

---

3 Donohue calls these ‘noun classes’, reserving ‘gender’ for the partly cross-cutting distinctions marked on adjectives
from object agreement on the verb (Donohue 2001: 99).4 In Figure 4, the outline criteria according to which nouns are assigned to the genders are given under ‘assignment’.

<table>
<thead>
<tr>
<th>assignment</th>
<th>inflectional class A</th>
<th>inflectional class B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>e.g. -ihi- ‘see’</td>
<td>e.g. -akwa- ‘bite’</td>
</tr>
<tr>
<td>I male</td>
<td>j-</td>
<td>s-</td>
</tr>
<tr>
<td>II female, animate</td>
<td>s-</td>
<td>b-</td>
</tr>
<tr>
<td>III miscellaneous</td>
<td>g-</td>
<td>n-</td>
</tr>
<tr>
<td>IV mass nouns</td>
<td>j-</td>
<td>j-</td>
</tr>
<tr>
<td>V banana, sago tree</td>
<td>s-</td>
<td>b-</td>
</tr>
<tr>
<td>VI arrows, coconuts</td>
<td>g-</td>
<td>n-</td>
</tr>
</tbody>
</table>

**Figure 4:** Verbal inflectional classes in Burmeso (Donohue 2001: 100, 102)

The different forms of -ihi- ‘see’ distinguish gender and number. They utilize a small inventory in drawing a good number of distinctions. In addition to verbs like -ihi- ‘see’, there is another set like -akwa- ‘bite’, also given in Figure 4. These have exactly the same paradigm shape, but the phonological realizations are entirely different (there are some syncretisms but even these are identical across the classes). Burmeso thus has two inflectional classes of verbs. In terms of membership in the classes, Donohue (2001: 101) states explicitly: ‘… there are no obvious semantic correlations for verbs which take the different sets of prefixes, and both sets of verbs are of approximately equal size.’

The degree of allomorphy that characterizes inflectional classes may vary widely. Compare the Czech verbal paradigms in Figure 5 with the paradigms in Figure 6 from Atsugewi (an extinct language of northern California). In both examples the contrasting inflectional classes are distinct for each cell in the paradigm. But whereas the Czech classes are minimally distinct from each other (e.g. 1SG -ám vs -ím), the contrasting inflectional classes in Atsugewi bear almost no resemblance to each other.

**Figure 5:** Czech present tense

<table>
<thead>
<tr>
<th>class I ‘do’</th>
<th>class II ‘ask’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1SG</td>
<td>děl-ám</td>
</tr>
<tr>
<td>2SG</td>
<td>děl-áš</td>
</tr>
<tr>
<td>3SG</td>
<td>děl-á</td>
</tr>
<tr>
<td>1PL</td>
<td>děl-áme</td>
</tr>
<tr>
<td>2PL</td>
<td>děl-áte</td>
</tr>
<tr>
<td>3PL</td>
<td>děl-ají</td>
</tr>
</tbody>
</table>

**Figure 6:** Atsugewi present actual (Olmsted 1961: 96, 98f, 112f)5

4 At least gender V is ‘inquorate’ in that it has only two members (Donohue 2001: 102) and its agreements are a combination of agreement from other genders; the two nouns could therefore be marked as lexical exceptions.
The degree of differentiation between may also differ in another respect. In the examples we have reviewed so far, the classes are distinct in all the cells, but more typically the inflectional classes are distinct only for a portion of the paradigm. For example, in Russian (Figure 7), we can distinguish four inflectional classes (at least), but in most instances the classes are in fact conflated, according to different patterns for the different morphosyntactic values.

\begin{center}
\begin{tabular}{|c|c|c|c|c|}
\hline
 & ‘thing’ & ‘factory’ & ‘country’ & ‘bone’ \\
\hline
NOM SG & del-o & zavod & stran-a & kost´ \\
\hline
ACC SG & del-o & zavod & stran-u & kost´ \\
\hline
GEN SG & del-a & zavod-a & stran-i & kost´-i \\
\hline
DAT SG & del-u & zavod-u & stran-e & kost´-i \\
\hline
LOC SG & del-e & zavod-e & stran-e & kost´-i \\
\hline
INS SG & del-om & zavod-om & stran-oj & kost´-ju \\
\hline
NOM PL & del-a & zavod-i & stran-i & kost´-i \\
\hline
ACC PL & del-a & zavod-i & stran-i & kost´-i \\
\hline
GEN PL & del & zavod-ov & stran & kost´-ej \\
\hline
DAT PL & del-am & zavod-am & stran-am & kost´-am \\
\hline
LOC PL & del-ax & zavod-ax & stran-ax & kost´-ax \\
\hline
INS PL & del-am´i & zavod-am´i & stran-am´i & kost´-am´i \\
\hline
\end{tabular}
\end{center}

\textbf{Figure 7:} Russian inflectional classes (phonological transcription)

The nested and interlocking relationships of the Russian inflectional classes arguably make them more complex than those of Burmeso, Czech or Atsugewi. One helpful way of visualizing this is in terms of ‘principal parts’. The traditional notion of the principal part, familiar to many from language instruction, is a fixed list of those forms from which one can deduce the rest of the paradigm, functioning as a concrete representation of inflectional class membership; for example, in Latin, the nominative and genitive singular are the traditional principal parts of the noun. Recent approaches (Finkel and Stump 2007, Ackerman, Blevins and Malouf 2009; see also Wurzel 1984 for an earlier approach in this mode) have reevaluated the notion in information-theoretic terms: principal parts are the most informative forms, in terms of narrowing down the options for the remaining forms in the paradigm. Seen in these terms, the richer the principal part information, and the more principal parts one needs, the more complex the inflectional class is. Thus a language such as Burmeso has a quite simple system: a lexical entry need only contain a single principal part – and one could choose any of the forms, and not even bother to identify its morphosyntactic value – in order to unambiguously identify its inflectional class. The ambiguity of the Russian classes means that it is more constrained. One could get by with a single principal part for each class, but only by carefully picking the right form for each, and possibly annotating it with

---

5 Olmsted (1961) gives affixes and roots separately; the paradigms in Figure 6 have been compiled on the basis of his description.

6 Note that, if zero exponence is included, ‘factory’ and ‘bone’ are grouped together in the nominative/accusative singular, while ‘thing’ and ‘room’ are grouped together in the genitive plural. Since there is no way to bring all the conflated exponents together in a two-dimensional chart, these have not been noted in Figure 7.
morphosyntactic information (e.g. a form in -a only identifies class II if it is labelled as ‘NOM SG’.)

In other cases, though, multiple principal parts will be needed, producing quite a dense lexical entry. Consider Nuer (a West Nilotic language spoken in Sudan) in the variety described by Frank (1999). Nouns inflect for three cases (nominative, genitive and locative) in the singular and plural. The set of case endings is quite restricted: for most nouns case inflection involves the bare stem and/or -kä in the non-nominative cases of the singular, and the bare stem and/or -ni in the plural (in addition, there are possible stem alternations which are largely independent of the choice of case ending). However, while the repertoire of case endings is small, their distribution is highly variable from lexeme to lexeme. Figure 8 gives some examples.

<table>
<thead>
<tr>
<th>‘bear’</th>
<th>‘ant’</th>
<th>‘lion’</th>
<th>‘fat’</th>
<th>‘egret’</th>
<th>‘monkey’</th>
<th>‘child’</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOM SG</td>
<td>let</td>
<td>ṇiec</td>
<td>lony</td>
<td>lieth</td>
<td>böön</td>
<td>gɔɔk</td>
</tr>
<tr>
<td>GEN SG</td>
<td>let</td>
<td>ṇiec-kä</td>
<td>lony</td>
<td>lieth-kä</td>
<td>böön-kä</td>
<td>gɔɔk-kä</td>
</tr>
<tr>
<td>LOC SG</td>
<td>let</td>
<td>ṇiec-kä</td>
<td>lony</td>
<td>lieth</td>
<td>böön-kä</td>
<td>gɔɔk</td>
</tr>
<tr>
<td>NOM PL</td>
<td>leet</td>
<td>ṇiec</td>
<td>luony</td>
<td>līth</td>
<td>böɔŋ-ni</td>
<td>gɔɔk-ni</td>
</tr>
<tr>
<td>GEG PL</td>
<td>leet-ni ṇiec-ni</td>
<td>luony-ni</td>
<td>līth-ni</td>
<td>böɔŋ-ni</td>
<td>gɔɔk-ni</td>
<td>gaaat</td>
</tr>
<tr>
<td>LOC PL</td>
<td>leet-ni ṇiec-ni</td>
<td>luony</td>
<td>līth-ni</td>
<td>böɔŋ-ni</td>
<td>gɔɔak-ni</td>
<td>gaaat</td>
</tr>
</tbody>
</table>

**Figure 8**: Varieties of Nuer noun inflection (Frank 1999)

Within the singular and plural nearly every logical possible pattern is found in Frank’s corpus, yielding four singular patterns and six plural patterns.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>2</td>
<td>Ø</td>
<td>-kä</td>
<td>-kä</td>
<td>Ø</td>
</tr>
<tr>
<td>3</td>
<td>-kä</td>
<td>-kä</td>
<td>Ø</td>
<td>Ø</td>
</tr>
</tbody>
</table>

**Figure 9**: Singular noun inflection in Nuer (Frank 1999)

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ø</td>
<td>-nį</td>
<td>Ø</td>
<td>Ø</td>
<td>-nį</td>
<td>Ø</td>
</tr>
<tr>
<td>2</td>
<td>Ø</td>
<td>-nį</td>
<td>-nį</td>
<td>-nį</td>
<td>Ø</td>
<td>Ø</td>
</tr>
<tr>
<td>3</td>
<td>-nį</td>
<td>-nį</td>
<td>-nį</td>
<td>Ø</td>
<td>-nį</td>
<td>Ø</td>
</tr>
</tbody>
</table>

**Figure 10**: Plural noun inflection in Nuer (Frank 1999)

Looking first at the singular and plural patterns individually, one sees that because of the small number of elements that constitute the system, and the freedom with which they are combined, in none of the classes is one form by itself sufficient to predict the rest of the paradigm. One needs two or even three forms in order to unambiguously identify the pattern.

And what is more, singular and plural patterns combine promiscuously with each other, yielding sixteen paradigm types:

---

7 This is in fact a simplified version of the system: (i) there is an additional NOM and GEN SG allomorph -dį, whose distribution matches that of -kä, with which it may also be combined in the same paradigm, and (ii) there are a few additional aberrant patterns involving the affixes -kä and -nį.
Thus, even though the paradigm for any lexeme is small, its lexical entry on a principal parts analysis will be rich; for some classes, the better part of the paradigm must be stored.

A further factor which compounds the complexity of inflectional classes is distributed exponence; that is, when inflectional information is distributed across various positions or zones of the word form, which may then behave independently of each other. Consider the verb paradigm in Chiquitlan Mazatec, an Otomanguean language of Oaxaca, Mexico (Jamieson 1982). Subject person and number is realized in three positions in the word: the final vowel, the prefixed initial syllable, and the tone (represented by superscript numerals, with ‘1’ the highest and ‘4’ the lowest tone). These three systems vary independently of each other, with each one falling into distinct inflectional classes in their own right. Figure 12 illustrates these multiple intersecting inflectional classes, taking the verb ‘gather’ as point of reference, and comparing it to three other verbs. In terms of its final vowel inflection, ‘gather’ forms a class with ‘return’, but not with the others. Its prefixal inflection groups it with ‘pull out’, while its tonal inflection matches that of ‘take out’. That is, the verb ‘gather’ simultaneously belongs to three inflectional classes, depending on which subsystem one is looking at. And the verbs ‘return’, ‘pull out’ and ‘take out’ in turn pattern with still other verbs, forming a network of interlocking inflectional classes. (There are approximately half a dozen final vowel and tonal classes, and two dozen prefixal classes.)

Figure 11: Singular ~ plural pattern mapping in Nuer (based on Frank 1999)
2.2 Form classes

Our first type of class involved stipulating the differences between different lexemes. The second type of class stipulates identities between cells in the paradigm. Recall the stem alternation shown in the Italian paradigm in Figure 2, where the augment -isc- is introduced in all singular forms, and in the third person plural, but not in the first or second person plural. To characterize this we need to state which cells in the paradigm take the augment. It is often tempting to explain away such form classes by seeking a semantic or syntactic motivation for the apparent disjunction of functions, or to treat it as accidental homophony. While this may be a matter open to discussion in individual cases, examples that defy such analyses are not rare. Take the paradigms illustrated in Figure 13, from the Cushitic language Dhaasanac, spoken in Ethiopia. The person-number-gender paradigm of verbs contains at most two forms, whose distribution cuts across any semantically coherent distinctions (second person + third person feminine singular + first person exclusive plural versus first person singular + third person singular masculine + first person inclusive plural + third person plural). Morphologically, the opposition is expressed in diverse ways: vowel alternations, consonant alternations, suffixation, prefixation; just some of the many possibilities are illustrated here. If these classes are not semantically coherent wholes, clearly they are also not morphological accidents. Therefore we suggest that the paradigmatic pattern itself – the form class – is encoded in the grammar (see Baerman, Brown and Corbett 2005: 183-186).

**Figure 12:** Chiquitlan Mazatec (Jamieson 1982: 152, 166-7)
As was shown above for inflectional classes in Chiquitin Mazatec, the existence of multiple zones of inflectional exponence within the word form potentially leads to the simultaneous expression of overlapping form classes. Chichimec, another Otomanguean language, offers a good example. The full paradigms are given in Figure 14, though the picture will be clearer if we look just at the base forms, namely the first, second and third person singular, and the third person plural (since the remaining forms of the dual and plural are formed through regular suffixation to the singular forms).

The prefixes and suffixes both conflate person-number values, but according to different patterns. In the recent past the prefix conflates first and third person, while the stem conflates second and third. In the potential mood, the prefix conflates second person and third person plural, while the stem conflates first and third person (the latter in the singular and dual only). (It is also interesting to note that the stem alternants are the same in both paradigms, only the pattern of alternation differs.) This is example is like Dhaasanac in that classes of forms must be stipulated, but with the added complication that the prefixal form classes and the stem form classes are distinct from each other.

### 3 Implications of morphological complexity

Morphological complexity as we have characterized it has long remained on the sidelines of linguistic research. In particular, because morphological patterns of the sort we have described above are by definition language-specific, it is a major challenge simply to extract any cross-linguistic generalizations, let alone to incorporate any of these generalizations into
any model of language. Yet such patterns are a central feature of the languages that have them, and to ignore them is to ignore a whole dimension of possible linguistic structure. Below we outline some of implications that we believe morphological complexity – in the sense described in §2 – has for different areas of research into language.

3.1 Challenge for linguistic theory
The two phenomena outlined in §2 - inflectional classes and morphological form classes - clearly will have some effect on grammatical representation, as they require something more than a simple one-to-one mapping of form and function. Some take this as evidence for the recognition of morphology as a distinct grammatical component, while even those approaches that explicitly reject the notion of the autonomy of morphology (Distributed Morphology being an obvious example) are nevertheless compelled to enrich their descriptive language through devices such as inflectional class features. So a theory of grammar that takes morphological complexity into account will look different from one that does not.

But of course, theory and formal representation are only as interesting as the insights they lend and the empirical predictions they make. And here the question that looms large is: are there any inherent limits on morphological complexity? Various proposals have been to define both possible inflectional classes and possible form classes, as well as to constrain multiple exponence (which can lead to multiple layers of morphological complexity). All these can have profound effects on the architecture of the grammar.

3.2 Challenge for computational approaches
At one level, morphological complexity is an implementational problem for computational applications. Issues such as stemming and lemmatisation are treated as purely technical problems which coincide with linguistic questions. Nevertheless, there may well be insights to be drawn from practical solutions to the problems that real languages create. But at another level, computational modelling may be used for purely experimental purposes. For example, Goldsmith and O’Brien (2006) use a feed-forward connectionist model in which a hidden layer encodes the information associated with the complexity we discuss in our typology. We can also measure the similarity of the paradigms of lexical items by using compression algorithms as a proxy for Kolmogorov Complexity (which is not in itself computable). Simply put, if we take the full paradigm of one lexeme and use this information to compress the paradigm of another lexeme, the more alike the two lexemes, the more effective the compression will be. As an illustration of this, we applied this technique to eighty high frequency Russian nouns. We compared their full paradigms using Cilibrasi and Vitányi’s (2005) CompLearn tool.

---

8 Compression-based techniques have a wide application. In terms of linguistic applications they have been used for determining genealogical relatedness of languages and also for comparing the various types of similarity in EU languages (see Cilibrasi & Vitányi 2005, and Kettunen et al 2006). We compared the full paradigms of the Russian nouns, together with overlaid stress pattern, derived from the Network Morphology analyses of Corbett and Fraser (1993), Brown et al (1996) and Brown (1998). As this was implemented in DATR (Evans and Gazdar 1996) the theory can output the fully inflected forms for each lexeme. Furthermore, the lexicon of 1500 or so noun lexemes had been created on the basis of frequency information from Zasorina (1977) and so could be presented in frequency order. Cilibrasi and Vitányi’s Complearn (available from www.complearn.org) calculates a normalized compression distance (NCD) to determine similarities. The output can be visualized in a variety of ways. In figure 1, we see the result of the analysis of the first 80 lexemes presented as an unrooted tree.
As we can see, class II nouns (strana etc.) in Figure 16 now occupy the left of the graph. Similar grouping occurs for the other classes, but with some interesting intermediate cases. The technique used is not domain-specific and figure 2 indicates that inflectional classes are in a certain sense learnable. However, it should be borne in mind that the data are not presented as full paradigms when someone acquires the language. Pirrelli (2007: 133) points out that the natural input is underspecified, whereas here the comparison could be seen as based on 'fully specified lexical representations' (Pirrelli 2007: 134).

3.3 Challenge for psycholinguistics

It is clear that linguistic theory and computational methodology must take some account of morphological complexity. But the underlying question remains: do the morphological objects that we have described have any psychological reality? The more complex the data become, the greater the temptation to believe that we are dealing with unanalyzed units. Psycholinguistic studies have long suggested a difference between productive and stored grammatical expression, with disagreement about whether this is categorical or gradient. Typically this difference is correlated with the contrast between regular concatenative affixation and irregular non-concatenative alternations. But more recent work has begun to look into the psycholinguistic effects of morphological complexity (e.g. Milin et al. 2009, Veríssimo and Clahsen 2009), showing that inflectional classes and stem alternations have observable psychological correlates.
The question of the acquisition of morphological complexity is both challenging and important. Anecdotally, complex morphology is learned easily by children, and only with considerable difficulty by adults. In practical terms our ability to make useful observations is limited, but any work in this area will be valuable.

References
Finkel, Raphael and Gregory Stump. 2007. Principal parts and morphological typology Morphology 17.39-75.
Milin, Peter, Dušica Filipović Đurđević, and Fermín Moscoso del Prado Martín. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: evidence from Serbian. *Journal of Memory and Language* 60.50-64.


