Combining formal and frequency based approches to morphology[1]

Dunstan Brown
Surrey Morphology Group, University of Surrey

1 Introduction

In corpus-based approaches it is typically considered an advantage to use large corpora. In contrast with this, we wish to determine how readily the inflectional classes recognized by linguists can be inferred by an unsupervised learning method when it is presented with the paradigms of a small number of Russian noun lexemes (namely 80) which are the most frequently occurring in a corpus (Zasorina 1977). Inflectional classes constitute a particular challenge, because they constitute a kind of morphological complexity whereby one and the same grammatical distinction can be expressed in a number of different ways. This is additional structure which is not relevant from the point of view syntax. In other words, it is complexity associated with autonomous morphology in the sense of Aronoff (1994). The ability to cluster items into their correct inflectional class is a prerequisite for inferring the other surface forms of a lexeme. So correct identification of inflectional classes represents part of the solution to what Ackerman *et al.* (2009) call the Paradigm Cell Filling Problem (PCFP):

> "What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?" (Ackerman *et al.* 2009: 54)

Given some of the inflectional forms alone, we cannot reliably infer all of the exponents for the noun in question. For example, knowing the dative, locative or instrumental plural in Russian is of no help in inferring the other forms of the paradigm of a given noun, as these are the same across all classes. In contrast, knowing the nominative singular is more helpful, although it will not always guarantee success. Ackerman *et al.* (2009) claim that the tractability of this problem is guaranteed by the fact that inflectional classes are constrained to reduce entropy, so that not all instances of particular inflectional exponents are equally probable. Finkel and Stump (2007) appeal to the traditional notion of principal parts, as these are the most informative as to the class membership of the lexical item.

This talk will provide some preliminary results and discussion of how useful this method is in helping us understand formal notions such as principal part. As we have used a formal theory to generate the full paradigms of the nouns, we can then contrast the outcome involving full paradigm information with the performance in clustering the same high frequency Russian noun lexemes into their appropriate inflectional classes when particular types of information have been removed from their paradigms. When we remove default information, shared across classes, we expect there to be little effect on the clustering relative to the base set with all information included. This contrasts with principal part information where we expect there to be a detrimental effect on clustering. Our results do indeed show that removal of forms classified as principal parts has a more detrimental effect on the clustering than removal of default information. However, we also find that there are differences within the defaults and principal parts. Furthermore, removal of some information may actually improve the clustering in comparison with the base set. This method provides external validation of the classes recognized by linguists, and the systematic removal of default and principal part information prepares the ground for the next logical step, namely to check the inference of inflectional classes on the basis of the actual token frequencies with associated paradigmatic gaps to be found in actual usage.

2 Method for investigating defaults and principal parts

The data for our machine-learning experiment are full paradigm listings of the first 80 most frequent nouns from Zasorina's (1977) frequency dictionary of Russian. An example of such a listing, for the noun *strana* (country) is given in (1).

```
(1)
mor sg nom = stran ^ a @".
mor sg acc = stran ^ u @".
mor sg gen = stran ^ i @".
mor sg dat = stran ^ e @".
mor sg inst = stran ^ o @" ^ j ( u ).
```

[1] The work reported on here has been carried out jointly with Roger Evans (University of Brighton).

```
mor sg prep = stran ^ e @".
mor sg prep loc = stran ^ e @".
mor pl nom = stran ^ i.
mor pl acc = stran ^ i.
mor pl gen = stran.
mor pl dat = stran ^ a ^ m.
mor pl inst = stran ^ a ^ m'i.
mor pl prep = stran ^ a ^ x.
```

These listings include morphological feature information, as well as the forms themselves in phonological transcription. The caret (^) marks concatenation and the symbol combination @" marks stress. These paradigms were generated from a DATR theory of Russian nouns, and the use of the inheritance based representation, DATR (Evans and Gazdar 1996) means that we can check the default status of the exponents against the underlying theory to see how the clustering method performs when default information is removed. We can contrast this with principal part information.

2.1 Compression-based machine learning

The machine-learning paradigm that we use is the compression-based approach described in Cilibrasi and Vitányi (2005) and Cilibrasi (2007), and implemented in the CompLearn tools.[2] This approach has two main components: (a) the use of compression (in the sense of standard compression tools such as zip, bzip etc.) as the basis of a metric for comparing data objects (Normalized Compression Distance or NCD) and (b) a heuristic clustering method, which groups objects together according to their similarity using this metric. Together, these components provide a general purpose unsupervised method for clustering arbitrary digital data objects. Cilibrasi (2007) provides examples of its application to fields as diverse as genetics in mammals and viruses, music, literature, and language relatedness. From the distance matrix containing the NCDs for pairs of data objects, CompLearn creates an unordered tree representing clustering relationships implicit in the distance matrix. For our purposes a further step is added. This is to generate clusters using several clustering heuristics (balance number of leaves, balance maximum NCD of leaves, balance average NCD of leaves) and compare the outcome with the expected result for inflectional class membership.

3 Results and Conclusion

Our initial results have proved to be interesting. As expected, removal of default information, such as the oblique plural forms (dative plural, instrumental plural and locative plural) has a small effect on the correct classification in comparison with the base set. In contrast, removal of forms which play a key role in identifying classes, such as the instrumental singular, impairs the accuracy of the classification to a greater extent. The next logical step from this is then to compare the results where we have removed elements of the paradigm systematically with actual token frequencies in corpora for the high frequency set, and also with lower frequency items. This approach complements formal work on principal parts.

4 References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*. Oxford University Press, Oxford, UK, pages 54-82.

Mark Aronoff. 1994. *Morphology by itself: Stems and Inflectional Classes*. The M.I.T. Press, Cambridge, Mass.

Rudi Cilibrasi and Paul M. Vitányi. 2005. Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523-1545. xxx

Rudi Cilibrasi. 2007. *Statistical Inference Through Data Compression*. Ph.D. Institute for Logic, Language and Computation, University of Amsterdam.

Roger Evans and Gerald Gazdar. 1996. DATR: a Language for Lexical Knowledge Representation. *Computational Linguistics* , 22(2), pp. 167-216.

Raphael Finkel and Gregory Stump. 2007. Principal parts and morphological typology. *Morphology*, 17: 39–75.

L. N. Zasorina. 1977. *Častotnyj slovar' russkogo jazyka*. Russkij jazyk, Moscow. xx

[2] http://www.complearn.org