*Estonian Morphology: What is Learned and How?**

Authors omitted for review

Abstract

This paper is about some theoretical considerations that one should be aware of when applying quantificational methods for describing Estonian morphology. Estonian is a morphologically rich language, and being available for researchers in the form of various large and tagged text corpora, it should be a good test-bed to investigations into the nature of morphology. The paper is based on evidence from text corpora, although it does not rely on very sophisticated statistics. It proposes an alternative view to explain the data that has inspired theories about the dual and single mechanisms for morphological processing, and namely that during learning the morphological rules, the learners over-generalize from the regularities they see in naturally occurring communication. The paper serves as a preliminary to a lot more counting and calculating that could be done on Estonian text corpora.

Keywords—Estonian, morphology, dual mechanism, single mechanism

I. Introduction

This paper is about some theoretical considerations that one should be aware of when applying quantificational methods for describing Estonian morphology. One can access various text corpora of Estonian, e.g. a 270 million token corpus collected from the Web etTenTen, searcable via www.keeleveeb.ee, a 200 million token Estonian Reference corpus, downloadable from www.cl.ut.ee and also searchable via www.keeleveeb.ee, a 0.5 million token morphologically hand-tagged corpus, downloadable from www.cl.ut.ee/korpused/morfkorpus/ , the Estonian CHILDES corpus etc. The paper is based on evidence from these corpora, although it does not rely on very sophisticated statistics. So the paper serves as a preliminary to a lot more counting and calculating that could be done on this material.

II. Dual and single mechanism views

There is an ongoing debate about the nature of morphological regularities and exceptions, as exemplified by papers on the English, German and Dutch plural inflection (for references, see [1]). According to the dual mechanism view, there is a default way of forming the plural, requiring no awareness of any specific properties the word has; and there are exceptions that depend so heavily on the properties of the word that they have to be remembered one by one. E.g. in English, the default is adding s to any stem, regardless of its phonological form (*spouse-spouses*, *house-houses*), the exceptions *mouse-mice* and *louse-lice* simply have to be remembered.

According to the single mechanism view, a speaker is always aware of the word's phonological, semantic etc. properties. For choosing the right inflectional affix for plural, e.g. *-en* or *-s* for German or Dutch, he first has to classify the word according to its properties. The default is simply the most numerous class of words after the classification has been performed on the vocabulary of the language.

A crucial argument in this debate is related to the inflection of rare and new words in text corpora and to the inflection of nonsense words in experiments: they should reveal the speakers' language intuitions, because their affiliation to certain inflectional classes could not have been learned beforehand and cannot be retrieved directly from memory.

[2] argue that if these intuitions are the result of a single memorizing and generalizing process, then it should be possible to observe how they change if the input data changes during the learning phase. However, in case of a dual mechanism, the intuitions would remain the same, because the learning phase affects only the exceptions which have to be memorized.

III. Estonian singular genitive formation

Depending on the phonological form of the singular nominative, there are various ways of forming the singular genitive form of the word in Estonian, much like there are several ways of forming the German plural or the Russian singular genitive.

(Singular genitive serves as the base for almost half of the case forms in the 28 slot paradigm of a declinable word (noun or adjective) : 11 singular case forms and the plural nominative. Incidentally, instead of the singular genitive, we might as well speak about Estonian plural inflection as a test case for morphological processes, on par with the English, German or Dutch one, because the plural nominative is always formed simply by adding *d* to the singular genitive form; there are no exceptions to this rule.)

The singular genitive form has to end with a vowel. According to the description of productive inflectional morphology of monomorphemic words in [3: 434-435)], if the singular nominative already ends in a vowel, then the singular genitive is exactly the same as the nominative; however, if the singular nominative ends in a consonant, then either *a*, *e*, *u* or *i* has to be attached as a stem vowel. The choice of the vowel depends on the phonological form and the “degree of wordiness” which indicates how tensely the word is related to the vocabulary of the language, accounting for exceptional inflection for acronyms and citations.

For example, if a disyllabic common noun starts with a short syllable and ends in -C*in*, then the vowel is normally *a* (*plugin – plugina* plug-in); if a polysyllabic word ends in -CV*s*, then the vowel is always *e* (*Sokrates – Sokratese* ; *ISIS – ISISe* 'Acronym for Islamic State'); if a polysyllabic word ends in *-ik*, then the vowel is normally *u* (*Dubrovnik – Dubrovniku*; *sputnik – sputniku* 'satellite'); in other cases, the vowel is *i* (*Bing – Bingi*, *London – Londoni, Camelot - Cameloti*).

However, it is quite usual that a word is inflected differently from the productive pattern, although it has the phonological (or other, extrinsic to morphology) features that should designate it to the productive inflectional class, much like mouse and louse do not belong to the house and spouse class.

A telling example is presented by Estonian monosyllabic consonant-ending words, e.g. *näpp* 'finger', *käpp* 'paw' and *täpp* 'point', the genitive singular of which is formed by adding a different vowel to the stem: *näpu*, *käpa* and *täpi*. Rare words and new loans are inflected exclusively with *i*, e.g. *äpp* - *äpi* 'app, application'. Moreover, old words with *a*- or *u*-genitive tend to move into *i*-genitive class.

When describing German plural, [4] points that among the words with similar extra-morphological properties, one can distinguish exactly one stable inflectional class and possibly several unstable classes. A stable class is productive, is considered to be the default, normal way of inflecting such words, and contains much more words than unstable classes (including all the rare words). In our example, it is the class with the genitive with *i*. An unstable class is not productive, is considered to be exceptional, and contains only a small number of words, which are never the infrequent ones. In our example, these are the classes with the genitive with *a* and *u*. Notice that while a and u are unproductive stem vowels here, they are productive for other words with different sets of extra-morphological properties.

How does this state of affairs come into being? One explanation is that when a speaker sees a new word, he will first designate it to several inflectional classes, in accordance with its extra-morphological features (e.g. *äpp – äpa/äpi/äpu*), and later, after mutual communication, the speakers arrive at the agreement about the single acceptable, correct inflectional class (*äpi*). There is only one problem with this explanation: evidence does not support it.

When confronted with a rare or previously unseen word, speakers of Estonian immediately exhibit remarkable consensus about what is the generally accepted (i.e. normal, correct) way of forming its inflectional forms (e.g. no-one tries *äpa* or *äpu*). To put it differently, the speakers somehow manage to classify the word in a similar way, designating it to the same inflectional class. Their lack of disagreement is really noteworthy, because for an unseen word, the speakers could not have been negotiating its inflectional class beforehand.

Moreover, it is noteworthy that instances of actual negotiations about the inflectional class of a rare or new word (like a foreign name) are virtually absent in everyday communication.

Evidence from a 270 million token corpus collected from the Web etTenTen show that there is actually no need for such negotiations: there is almost no variety in the choice of the inflectional class for a new word; typing errors account for a far larger amount in variation in word forms than misclassifications into alternative inflectional classes.

To sum up, it looks like Estonian singular genitive inflection for monomorphemic consonant-ending words is best described as containing four default classes, each with its own exceptions, so that we see a fourfold dual mechanism at work.

IV. *Learning as Over-Generalization?*

However, at closer inspection we may observe a few telling instances when Estonians do have problems in deciding what the correct inflectional class of a word is. Those instances fall into two scenarios. In the first scenario, the choice is between an exceptional, old, unproductive inflectional class versus a regular, productive one, and involves cases when an old word has become rare and thus its exceptional inflectional behaviour cannot be remembered by everyone, or when an Estonian family name coincides with a common noun belonging to an exceptional inflectional class (like in English: the plural of the family name *Foreman* is the *Foremans*, not the *Foremen*). In the second scenario, the choice is between two productive classes, and happens when a new word or proper noun has extra-morphological properties that belong to orthogonal categories (e.g. phonetic properties and wordiness), predicting a different productive class membership. E. g. *Breivik* is a foreign name that appeared in Estonian texts only recently. Being disyllabic and ending in *-ik*, it should phonetically belong to the class of *u*-ending singular genitives (*Breiviku*). Being a new and foreign word, it looks very non-wordy, and thus should belong to the class of i-ending singular genitives (*Breiviki*). According to etTenTen, 75% of the 400 mentions are *Breiviki*, 25% are *Breiviku*. The other possible stem vowels, *a* and *e*, are never used.

The first observed scenario provides confirming evidence for the dual mechanism at work and the second provides evidence for the single mechanism. In order to re-conciliate the evidence with the two alternative mechanisms, one should first question the notion of "mechanism": what is the exact meaning of this metaphor in this context? Intuitively, "mechanism" is a known system of operations (e.g. a clockwork) or transmissions (e.g. from the engine to the wheels). Applying this metaphor to mental processes seems highly speculative. If we discard this metaphor altogether as potentially misleading, we may propose another explanation for the evidence we see in Estonian: what we observe as the default behaviour in classifying the words, is simply the manifestation of previous over-generalization during deducing the morphological rules from the regularities one sees in naturally occurring communication; the rules are used for connecting extra-morphological features with the inflectional classes. It is very natural for humans to make decisions, based on incomplete evidence (in other words, jump to conclusions), and then revise their decisions, based on new data. These revisions may involve learning more detailed features for classification (e.g. that polysyllabic words ending in -CV*s* are different from other nouns), up to the degree of memorizing single exceptional words.

Given the inevitable variability of the input that different learners are exposed to and the conflicting examples it contains (like the Estonian consonant-ending monosyllabic words), how does it happen that learners acquire a uniform habit to classify a word into the single right, default, commonly accepted inflectional class, based on its extra-morphological properties? This is a question about the type and token frequency distributions in the input and the learning mechanism that (over)generalizes these distributions into morphological rules.

As a first stab, we should have a closer look at the frequency profiles of words that share the same extra-morphological properties. A telling example is the distribution of disyllabic words with a long first syllable and ending in -CV*s*. When it comes to singular genitive, these words belong to 2 inflectional classes: they either have *e* as the stem vowel, e.g. *pinnas – pinnase* 'soil', *boonus – boonuse* 'bonus', or they have *a* as the stem vowel and stem alternation, e.g. *kinnas – kinda* 'mitten, glove', *soodus – soodsa* 'beneficial'. In a 500,000 token morphologically hand-tagged corpus (www.cl.ut.ee/korpused/morfkorpus/) both classes are equally distributed in terms of tokens, but the *e*-genitive class has twice as many types as the *a*-genitive class. We know that all new words are inflected with e as the stem vowel. So it turns out that the learner simply has to be sensitive to the type frequency in order to learn the default rule for the words with similar extra-morphological properties. Notice, however, that we still do not know how it is learned in the first place that for inflection, there is a significant difference between CV*s*-ending versus other consonant-ending words.

V. Conclusion

Estonian is a morphologically rich language, and being available for researchers in the form of various large and tagged text corpora, it should be a good test-bed to investigations into the nature of morphology. It is also no wonder that Estonian poses some challenges to theories that have been formulated, based on data from languages like English, German or Dutch, which lack some features Estonian has. The paper proposed an alternative view to explain the data that has inspired theories about the dual and single mechanisms for morphological processing, and namely that during learning the morphological rules, the learners over-generalize from the regularities they see in naturally occurring communication. The exact nature of what is looked over during generalizing and how it is done remains yet to be found out.

References

[1] D. Nemeth, K. Janacsek, Z. Turi, A. Lukacs, D. Peckham, S. Szanka, et al., “The production of nominal and verbal inflection in an agglutinative language: evidence from Hungarian,” in PLoS ONE 10(3): e0119003., 2015 doi:10.1371/journal.pone.0119003

[2] E. Keuleers, D. Sandra, W. Daelemans, S. Gillis, G. Durieux and E. Martens. “Dutch plural inflection: The exception that proves the analogy,” in Cognitive Psychology, vol. 54, 2007, pp. 283-318.

[3] H.-J. Kaalep, “Eesti käänamissüsteemi seaduspärasused (The regularities of the Estonian declensional system),” in Keel ja kirjandus, vol. 6, 2012, pp. 418-449

[4] W. U. Wurzel, “System-dependent morphological naturalness in inflection,” in Leitmotifs in Natural Morphology. Studies in Language Companion Series, vol. 10, W.U. Dressler, W. Mayerthaler, O. Panagl, W.U. Wurzel, Eds. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1987, pp. 59-95.