MWU generation breaking down: unpredictable inflectional forms in Estonian noun phrase paradigms

(WG 1: Lexicon-Grammar Interface)

Heiki-Jaan Kaalep
University of Tartu
heiki-jaan.kaalep@ut.ee

A wide-spread view of the process of forming a natural language expression is the following: at some point the person knows what single word or noun phrase he wants to use, and that it must be in a certain case. The task is then to determine the corresponding inflectional form of the word or noun phrase. This task is filled by a morphology module, a (relatively) simple, autonomous, automatic subsystem in the complete workflow of natural language processing.

For many languages (including Estonian), tools for analysing and generating word forms are commonplace, and regularities governing the ways single words are combined into phrases, are well described. Thus it should be a simple task to generate all the inflectional forms of a noun phrase. This is a practical and common task in various fields. For example, a well known method for increasing the word form coverage for statistical machine translation is the following: add a bilingual dictionary to the parallel corpus during the training phase, so that the resulting phrase table will include the translations directly from the dictionary, in addition to the words and phrases, extracted from the corpus. A natural enhancement would be to generate all the possible inflectional word forms for the dictionary entries, and add these. Dictionaries usually also contain phrases, so the task of generating the inflectional forms involves them also.

Surprisingly, an attempt to automatically generate inflectional forms for Estonian noun phrases from an English-Estonian dictionary (http://www.eki.ee/dict/ies/) revealed a situation when certain forms for certain noun phrase types cannot be formed by simply inflecting individual words, but require conversion of the noun phrase into a compound word, prior to inflecting. An example noun phrase is *kott kartuleid* (sack (sg nominative) potato (pl partitive)), i.e. sack of potatoes. The whole paradigm is shown in Table 1. Notice that the plural genitive cannot be formed in a similar manner as the plural nominative and partitive, i.e. by simply inflecting the head of the phrase, which would result in **kottide kartuleid*. Instead, it is an inflectional form of a compound word *kartulikott* (potatosack).

case	singular	plural
nominative	kott kartuleid	kotid kartuleid
genitive	koti kartuleid	kartulikottide
partitive	kotti kartuleid	kotte kartuleid
illative	kotti kartuleisse	kartulikottidesse
inessive	kotis kartuleis	kartulikottides
elatiive	kotist kartuleist	kartulikottidest
allative	kotile kartuleile	kartulikottidele
adessive	kotil kartuleil	kartulikottidel
ablative	kotilt kartuleilt	kartulikottidelt
translative	kotiks kartuleik	kartulikottideks
terminative	koti kartuliteni	kartulikottideni
essive	koti kartulitena	kartulikottidena
abessive	koti kartuliteta	kartulikottideta
komitative	koti kartulitega	kartulikottidega

Table 1: Inflectional paradigm of *kott kartuleid* (sack of potatoes)

All the slots of the paradigm that are (in accordance to Estonian morphology) based

on the plural genitive, are also generated as forms of this compound word.

It is worth noting that this type of noun phrase – quantified noun phrase (EKG 1993: 144-146) – is rare, but not exceptional; one may come up with any similar one, e.g. *kast õlut* (rack of beer), *meeter lund* (meter of snow), *rida arve* (row of numbers). However, we notice that once a quantified noun phrase becomes rigid, lexicalises, then it also looses potential to be used in all possible case forms. E.g. *tükk aega* (piece (sg nominative) time (sg partitive), i.e. for a long time) has no plural forms.

It is also noteworthy that the original noun phrase in the singular nominative case is not strictly synonymous with the compound word, if the latter was in the singular nominative case; i.e. *kott kartuleid* ≠ *kartulikott* (sack of potatoes ≠ potato-sack), *kast õlut* ≠ *õllekast* (box of beer ≠ beer rack), *meeter lund* ≠ *lumemeeter* (meter of snow ≠ snowmeter). However, the plural inflectional forms are semantically perfectly fitting with the original semantics of the phrase.

This puzzling phenomenon, when the standard straightforward processing principles of Estonian noun phrases break down, might allow us to gain some understanding of the way grammar normally works. The present treatise merely attempts to bring it to the knowledge of researchers that there is a detour in the (hypothetical) sequence of language processing steps.

A similar phenomenon – impossibility to generate fillers for some slots in an inflectional paradigm of single words – has been observed before (Sims 2009). One would hope that similar instances from different languages could help to shed light on the mystery.

References

Andrea D. Sims. 2009. Why defective paradigms are, and aren't, the result of competing morphological patterns. *Proceedings of the 43rd Annual Meeting of the Chicago Linguistic Society* (http://www.academia.edu/3743377/Why_paradigmatic_gaps_are_and_arent_the_result_of_competing_morphological_patterns)

EKG 1993. *Eesti keele grammatika II*. (Estonian grammar II) Tallinn, ETA KKI.

.