7

Phonotactic Complexity of Finnish Nouns

FRED KARLSSON

7.1 Introduction

In the continuous list of publications on his homepage, Kimmo Koskenniemi gives an item from 1979 as the first one. But this is not strictly speaking his first publication. Here I shall elevate from international oblivion a report of Kimmo's from 1978 from which the following introductory prophecy is taken: "The computer might be an extremely useful tool for linguistic research. It is fast and precise and capable of treating even large materials" (Koskenniemi 1978: 5).

This published report is actually a printed version of Kimmo's Master's Thesis in general linguistics where he theoretically analyzed the possibilities of automatic lemmatization of Finnish texts, including a formalization of Finnish inflectional morphology. On the final pages of the report he estimates that the production rules he formulated may be formalized as analytic algorithms in several ways, that the machine lexicon might consist of some 200,000 (more or less truncated) stems, that there are some 4,000 inflectional elements, that all of these stems and elements can be accommodated on one magnetic tape or in direct-access memory, and that real-time computation could be 'very reasonable' (*varsin kohtuullista*) if the data were well organized and a reasonably big computer were available (ibid.: 52-53). I obviously am the happy owner of a bibliographical rarity because Kimmo's dedication of 1979 tells me that this is the next to the last copy.

This was five years before two-level morphology was launched in 1983 when Kimmo substantiated his 1978 exploratory work by presenting a full-blown theory of computational morphology and entered the international computational-linguistic scene where he has been a main character ever since.

Inquiries into Words, Constraints and Contexts.
Antti Arppe et al. (Eds.)

In the 1970s Kimmo worked at the Computing Centre of the University of Helsinki but he also studied general linguistics and engaged himself in linguistic computing, i.e. the computational treatment of corpora for purposes of linguistic research. When our collaboration started in 1980, he had obtained a copy of the magnetic tape of the Reverse Dictionary of Modern Standard Finnish (Tuomi 1972) of which he made various refined machine versions that were of great importance for our early theoretical and descriptive work in computational morphology.

My book *Suomen kielen äänne- ja muotorakenne* (Structure of Finnish Phonology and Morphology, Karlsson 1983) profited greatly from the computerized lexical and other corpora provided by Kimmo. In commemoration of Kimmo's work in linguistic computing in those early days I shall here present some observations on the phonotactic complexity of Finnish nouns using those same valuable data from around 1980 out of which not all potential scholarly juice has yet been squeezed.

If phonemically /long/ vowels and consonants are treated as combinations of identical phonemes (the standard solution), Finnish has eight vowel and thirteen consonant phonemes, /i e æ y œ u o a/ and /p t k d s h v j l r m n ŋ / for which I henceforth am going to use their standard orthographic equivalents <i e ä y ö u o a p t k d s h v j l r m n ng> (/ŋ/ is phonemic only when long, symbolized by the digraph <ng>).

Morpheme boundaries are indicated by '+', syllable boundaries by the period, '.'.

7.2 Number of nouns

How many nouns are there in Finnish (or in any language)? The question might seem silly because nouns are the prime example of an open-ended word class. But the question is relevant if rephrased to concern either (i) the size of the core vocabulary, i.e. the 'central words' presumed to be known by every normal native speaker, or (ii) the number of atomic free root morphemes, to the exclusion of derivatives and compounds. Here, I shall try to answer (ii) on the basis of the material in the *Reverse Dictionary of Modern Standard Finnish* (RDF; Tuomi 1972).

The starting point of RDF was the data comprising the original version of the standard Finnish reference dictionary *Nykysuomen sanakirja* (NS; 1952-1962) with 207,256 entries the majority of which were more or less productively formed compounds. RDF contains all basic words, derived words, basic components of compounds, compounds the basic parts of which occur only in compound words, and a handful of clitics (which are not words proper). In all, RDF comprises 72,711 entries which implies that the number of compounds in NS is 134,545. (The machine-readable version of RDF available

at the Department of General Linguistics, University of Helsinki, has 72,785 entries.)

Of the 72,785 entries in RDF 34,673 (47.6 %) have the code 'S', short for noun (including words like *suomalainen* 'Finn; Finnish (adj.)', which have homonymous nominal and adjectival readings). But among these there are huge numbers of fully productive derivatives like *pysäköi+nti* 'parking', *nuole+skel+u* '(habit of) licking', *ost+el+ija* 'one who habitually buys', *tilaa+ja* 'one who orders', *tanssi+ja+tar* 'female dancer', *dumppa+us* 'dumping', *suvaitse+vais+uus* 'tolerance', *riittä+mättöm+yys* 'insufficiency', *marksi+lainen* 'Marxist'. There are at least 16,000 – 17,000 fully transparent derivatives like these.

Among the remaining 18,000 – 19,000 nouns there are still many that are morphologically more or less complex. For example, there are 251 words ending in *-isti* and 285 ending in *-ismi* like *aktivisti* 'activist', *aktivismi* 'activism'. A conservative estimate is that 1,500 more nouns could be decomposed by careful morphological (and even etymological) analysis. Furthermore there are at least 5,000 clear borrowings, e.g. around 4,000 nouns containing the foreign letters <b d g z x f c š w q> (words with the sequence <ng> are not included among these, nor are the genuinely Finnish ones with <d>).

This would put the number of genuinely Finnish, morphologically atomic noun roots in the vicinity of 12,000, perhaps even lower, which I think is much less than popular beliefs would hold. The next section offers an analysis of the 18,500 nouns which contain no fully transparent productive derivatives.

7.3 Canonical patterns

Table 1 depicts the incidence of monosyllabic noun roots in the nominative singular case form that belong to the core vocabulary, i.e. words which are known to any normal speaker of Finnish. This means that e.g. musical terms such as *do, re, mi, es, ais, cis*, or obsolete and dialectal words and non-assimilated borrowings are not considered, e.g. *hie, huu, hyy, gnu, bei, boa, jen, yen, sir*. There are eight potential monosyllabic patterns:

The 29 monosyllabic nouns listed in Table 1 comprise just a fraction of one percent of the 12,000 atomic root nouns surmised above. The only significant monosyllabic nominal pattern is CVV. This is surprising because from the viewpoint of general markedness theory one would have predicted both that the phonotactically simpler CV-pattern would occur and that it would occur more frequently than CVV. But CV-nouns (and verbs) are effectively prohibited. The reason for the prevalence of CVV over CV cannot be morphological either because there are CV-pronouns that are regularly inflected: *tä+hän* 'into this one' (illative singular), *jo+hon* 'into which', *mi+hin* 'into which'.

Table 1. Monosyllabic noun patterns in Finnish.

Pattern	Number	Examples
V	-	
VV	1	yö
CV	-	
CVV	24	hai, hää-, jää, koi, kuu, kyy, luu, maa, pii, puu, pyy, pää, suo, suu, syy, sää, tee, tie, tiu, työ, täi,voi, vuo, vyö
CVC	-	
VC	-	
CVVC	4	mies, syys, hius, ruis
VVC	-	
sum	29	

The singleton VV-word *yö* 'night' stands out as an exception. Monosyllabic VVV-strings are prohibited; *aie* 'intention' is disyllabic, cf. *aikee+n* (genitive singular).

As for closed monosyllables, (C/V)VC-nouns are non-existing (*ien, oas, äes* are disyllabic). CVVC-nouns are extremely marginal, there are four of them, which all have marked inflection. *Syys* 'autumn' allows no inflection, the other three are all inflected in different ways: *mies* 'man'- *miehe+n*, *hius* 'hair' - *hiukse+n*, *ruis* 'rye' - *rukii+n*. The inflected stems point in the direction of 'deep' (etymological) bisyllabicity.

We proceed to the disyllabic nouns, first those with an open second syllable (Table 2), then those with a closed second syllable ('.' indicates the location of the syllable boundary).

The tendency to avoid long vowels and diphthongs in the second syllable of genuine underived noun roots is very strong. There are only a few handfuls of such words, fractions of one percent of 4,958. However, there are around 100 borrowings like *filee, revyy, turnee* and many bimorphemic derivatives like *takuu* 'guarantee' (from *takaa-*, inflectional stem, 'to guarantee').

The most striking fact of Table 2 is the prevalence of long (bimoraic) first syllables , i.e. the structures CVV.CV and especially CVC.CV which are much more frequent that than the theoretically simplest pattern CV.CV. The trimoraic pattern CVVC.CV is almost as frequent as monomoraic CV.CV, which must be considered very surprising. The share of monomoraic (C)V.CV is only 756 + 61 = 817, i.e. 16%. The four-moraic pattern (C)VVCC.CV is encountered only in a few borrowings. As is to be expected, V-initial first syllables are much more infrequent, by a factor of 10 – 20, than CV-initial first syllables, e.g. V.CV as compared to CV.CV. VV.V does not occur in underived words but the derivative *ai.e* 'intention' (*ai+e*, from *aiko-* 'intend')

Table 2. Disyllabic noun patterns in Finnish with an open second syllable.

Pattern	Number	Examples
CV.CV	756	kala, peto, maku
V.CV	61	aho, ele, äly
CVV.CV	950	jousi, laatu, määrä, tuoli
VV.CV	52	aamu, aika, ääni
VV.V	-	
CVC.CV	1795	hihna, kukko, pentu
VC.CV	143	ahma, olki, ämmä
CVVC.CV	628	haaska, juusto, lieska
VVC.CV	33	aalto, aitta, äänne
CVCC.CV	503	harppi, kalske, lamppu
VCC.CV	23	ankka, arkki, yrtti
(C)VVCC.CV	6	aortta, nyanssi, seanssi
X.CVV	7	ehtoo, harmaa, suklaa, vastuu, tienoo, Porvoo, vainaa
sum	4,958	

is a singleton example of this pattern. The number of underived bisyllabic nouns with a closed second syllable is around 800 of which some 650 have a bimoraic first syllable.

7.4 Conclusion

Finnish has only some 30 monosyllabic and less than 6,000 underived bisyllabic nouns. Somewhat surprisingly, we have demonstrated that the prototypical first syllable of mono- and disyllabic nouns is bimoraic rather than monomoraic. The latter would be expected on grounds relating to general phonological simplicity, i.e. the universal preference for optimal light CV-syllables. Trisyllabic and longer nouns have not been analyzed in detail here but a fast test shows that more than 75% of them too have bimoraic or even heavier first syllables. The same holds across the board for the vocabulary: 75% of the lexemes listed in RDF have at least a bimoraic first syllable (CVC. 40,378, CVV. 13,899, CV. 17,171).

Why are more complex phonotactic structures so clearly preferred over simpler ones? Three possible causes come to mind. First, languages with relatively few phonemes (e.g. Finnish with twenty-one) tend to have longer words than languages with more phonemes (Nettle 1999). Thus, in Nettle's sample of ten totally unrelated languages from different stocks, the mean word length of Khoisan !Kung with 147 phonemes in its inventory was 4.02 segments whereas that of Turkish with 28 phonemes was 6.44 segments, 'word' being defined as a random sample of fifty uninflected stems in a size-

able dictionary. Second, bimoraic (and longer) syllables amplify the effect of the word-stress fixed in Finnish to the first syllable. Third, for morphophonological reasons, new words and borrowings prefer quantitative over qualitative consonant gradation. Quantitative gradation is possible only in (at least) bimoraic syllables, e.g. *rokki* 'rock' - *roki+n* (genitive singular).

References

Karlsson, F. 1983. *Suomen kielen äänne- ja muotorakenne.* Porvoo Helsinki Juva: Werner Söderström osakeyhtiö.

Koskenniemi, K. 1978. Suomen kielen sananmuotojen perusmuodon automaattinen päättely. Mahdollisuuksien arviointia ja taivutuksen formalisointi. *Helsinki: Computing, Centre, University of Helsinki, Research Reports N:o 5.*

Koskenniemi, K. 1983. Two-level morphology: A General Computational Model for Word-Form Recognition and Production. *Helsinki: University of Helsinki, Department of General Linguistics, Publications No. 11.*

Nettle, D. 1999. *Linguistic Diversity.* Oxford: Oxford University Press.

Tuomi, T. 1972. Suomen kielen käänteissanakirja. Reverse Dictionary of Modern Standard Finnish. *Suomalaisen Kirjallisuuden Seuran Toimituksia. 274. osa.* Helsinki: Suomalaisen Kirjallisuuden Seura.