On the limits of generalizing from quantitative, corpus-based evidence in a morphologically rich language

Antti Arppe
University of Helsinki
antti.arppe@helsinki.fi

This paper aims to demonstrate how corpus-based results can be inherently deficient in the study of a morphologically rich language, when attempts at generalizations concerning the underlying linguistic system are made on the basis of quantitative data extracted from a corpus. It will be shown that even the use of relatively large and divergent corpora may provide only a limited subset of all possible inflected forms, even when confining to a limited core group of forms. Thus, this sparseness, which most probably is irrectifiable by whatever increase of corpus size, can be seen to set limits on the extendability of corpus-based results as linguistic evidence. On the other hand, the less frequent or unobserved phenomena in corpora can be seen as the very areas where other types of linguistic research methods and evidence, for instance experimentation, could be of greatest added value.

As a case example, differences in the usage of a set of four near-synonymous Finnish THINK verbs, namely *ajatella-miettiä-pohtia-harkita* 'think, reflect, ponder, consider', with respect to their inflected forms and constituent morphological features will be studied, using as evidence two corpora. These corpora represent two modes of written communication which have been selected with the expectation that they differ from each other fundamentally in terms of their level of formality and interactivity. The first corpus consists of approximately 2 million words of Finnish newspaper text, and can be characterized as public, unidirectional reporting of mainly past events, which can be argued to reflect the norms of the standard written Finnish of today. In contrast, the second corpus consists of Finnish Internet newsgroup discussion amounting to some 700,000 words, and it can be described as less public, informal interactive correspondence, closer to the norms of (spoken) colloquial Finnish, written in form though it may be.

In theory, each Finnish verb can have well over 20,000 inflected forms when counting in all the various participle and infinitive forms, but restricting to the so-called core finite forms, the number still adds up to 530, which are inflected in voice (2 categories), mood (4), person and number (6) and tense (2/4). Furthermore, all these forms can be negated with a complex construction. As a lexeme-feature association

concerning one category, e.g. person, can in principle be distributed among the other categories, e.g. mood, resulting in low frequencies for each individual feature combination, it is therefore practical even with the core forms to start off with individual morpho-syntactic features and their combinations as incorporated in the compositional morphological analysis of the verbs, in order to attain overall distributions large enough for statistical analysis.

Features	Singlet	IND	COND	IMP	POT	A:AFF	NEG	A:NEG
SG1	**41** **A>M,H**	**41** **A>M,H**	0	-	0	**41** **A>M,H**	0	-
A:SG1	**66** **A,M>H,P**	**44** **A>M,H**	0	-	0	**58** **A,M>H,P**	4 M,A	9 A,M,H
SG2	8 M,A,H	1 A	1 M	6 M,A,H	0	8 M,A,H	0	-
A:SG2	10 A,M,H	1 A	1 M	6 M,A,H	0	9 A,M,H	1 A	1 A
SG3	154 P,M,H,A	**150** **P,M>H,A**	4 H,P,A	0	0	154 P,M,H,A	0	-
A:SG3	328 M,H,P,A	154 P,M,H,A	4 H,P,A	0	0	299 M,P,H,A	7 H,A,M	**29** **H,A,M>(P)**

Table 1: Distribution of the co-occurrences of selected features among the studied THINK verbs in the newspaper corpus (E.g. the contents of the SG3+IND cell, 150: P,M>H,A, denote that the overall absolute frequency of the co-occurrence of the 3rd person singular (SG3) and indicative (IND) mood is 150 instances among the studied verbs ajatella (A), miettiä (M), pohtia (P) and harkita (H), which all have at least one occurrence and of which pohtia and miettiä are the most frequent, with a difference to next frequent, harkita and ajatella, which is both statistically significant and at least exponential (Zipfian); verb-chain-specific cases are preceded by the prefix A[nalytical])

Features	Singlet	IND	COND	IMP	POT	A:AFF	NEG	A:NEG
SG1	**134** **A>M,H,P**	**123** **A>M,P,H**	**11** **H>A,M**	-	0	**134** **A>** **H,M,P**	0	-
A:SG1	252 H,A,M,P	**136** **A>M,P,H**	**15** **H>P,A,M**	-	0	208 H,A,M,P	**20** **H,A>M**	**44** **H>** **A,M,P**
SG2	**108** **M,A>P,H**	37 A,M	7 H,A,M	**64** **M>A,P**	0	**108** **M,A>** **H,P**	0	-
A:SG2	167 M,A,H,P	39 A,M	7 H,A,M	**64** **M>A,P**	0	153 M,A,H,P	8 A,M	14 P,M,A
SG3	**115** **A>M,P,H**	**104** **A>M,P,H**	9 H,P,M,A	0	2 M,A	**115** **A>** **M,P,H**	0	0
A:SG3	392 M,P,H,A	**108** **A>M,P,H**	8 H,M,A	0	2 M,A	330 M,H,P,A	20 P,A,H,M	62 P,A,M,H

Table 2: Distribution of the co-occurrence of selected features among the studied THINK verbs in the Internet newsgroup corpus

Tables 1 and 2 provide the results of the distributions of the THINK verbs with respect to selected person, number and mood features and their combinations in both affirmative and negated forms in the two research corpora. In addition to the verb-

specific analyses, both Tables include also data on the verb chain constructions of which the studied verbs are a component. Thus, in addition to the case that some person feature, say 1st person singular (SG1), is a morphological component of the actual studied verb, e.g. *mietin* 'I think', or *miettisin* 'I would think', such cases are also included in which this person feature has been observed anywhere in the verb chain construction of which the verb forms part of, e.g. *haluan miettiä* 'I want to think', *en haluaisi miettiä* 'I wouldn't want to think'

As can be noticed in the Tables, the 1st person singular feature appears to be significantly associated with *ajatella* as compared with the other three verbs in both corpora. Furthermore, this association also remains quite intact in the Internet newspaper corpus. However, closer scrutiny of the pairings of person with mood and affirmation/negation reveal that in the case of the newspaper corpus this outcome is fully based on affirmative (AFF) indicative (IND) forms, corresponding to the full surface forms *ajattelen/ajattelin* 'I think/thought', as the 1st person singular feature does not occur even once in the conditional (COND) or the potential (POT) moods. In the case of the Internet newsgroup corpus, there are a couple of 1st person singular forms in the conditional mood and 3rd person singular forms in the potential mood, but here, too, an overwhelming majority of the instances are likewise in the indicative mood. Thus, one can very well question whether the corpus-based evidence on this association between 1st person singular and *ajatella* applies to all moods and also in their negated forms, or rather only to the indicative mood in its affirmative form. On a more general level, it can also be seen from both Tables that several other person-mood combinations have no or relatively very few observed instances in either corpus, especially the forms of potential mood, but also the forms of the conditional mood and the 3rd person singular form of the imperative (IMP) mood. This scarcity could not be rectified even in the largest publicly available Finnish corpus, the Text Bank of Finnish with roughly 170 million words, which contains altogether only 48 instances of potential forms of the studied four verbs, none of which are in the 1st person singular.

In conclusion, it has been shown in the case of one feature-lexeme association that the corpus-based evidence at hand is in fact derived from two inflected forms, and this evidence by itself cannot be validly used to support any more general assertions concerning the association in question. More generally, this is a clear reminder that great care should be taken when interpreting corpus-based results of this type in a morphologically rich language. Furthermore, if the two corpora are together in any sense indicative of the proportional occurrences of the observed feature combinations, it may be difficult or impractical to resolve with corpus-based methods, e.g. increasing corpus size, whether the observed association can be generalized or not. Thus, this is an example case where experimentational methods would be a solution, e.g. by evaluating constructed sentences containing the rarer or unobserved feature combinations in the corpora, for instance the 1st person singular in the potential or conditional moods and in negated forms.