The usage patterns and selectional preferences of synonyms in a morphologically rich language

Antti Arppe

Department of General Linguistics, University of Helsinki – P.O. Box 9 (Siltavuorenpenger 20 A) – FIN-00014 University of Helsinki – Finland

Abstract

This paper observes with the help of corpora and statistical methods the usage preferences of two synonymous Finnish verbs and their inflected forms (and the features that constitute these forms), an aspect in lexical relationships that has hitherto been little observed. On the basis of the analyses, I claim that synonyms can have purely inflectional selectional differences in a morphologically rich language such as Finnish. However, the reasons for this remain still open. This would and should have an impact on lexicographical work and the content of lexicographical products such as dictionaries for such languages.

Keywords: lexicography, lexical semantics, synonymy, inflectional morphology, Finnish

1. Introduction

In whatever way synonymy is defined in theory, determining it in practice for particular words is a considerably more difficult endeavour. Words may clearly seem to have exactly the same meaning and usage, when observed in the abstract isolation of synonym listings in a dictionary or lexical database, for instance. However, when these same synonymous words are observed in the context of their actual usage, for instance in corpora, after only a few occurrences it soon becomes apparent that each has its own semantic connotations, lexical-syntactic associations and pragmatic limitations that are specific to each such word, in addition to the core concept or meaning chain that unifies the entire synonym set. It consequently appears to be accepted knowledge that **absolute** synonymy, more or less corresponding to substitutability without change of meaning in all possible contexts, is at best very rare (e.g. Zgusta 1971: 89; Miller et al. 1990: 240; Cruse 2000: 157-158), and that synonymy is in practice regulated by a variety of constraints in the context. A "context" which is not necessarily obvious in morphologically poor languages such as English is indeed the various inflected forms and their constituent features of a word, something one could dub the **internal context** of a word. The aim of this paper is to direct attention to the influence of inflection on the usage patterns and selectional preferences of synonyms in a morphologically rich language such as Finnish.

In general, constraints on word usage seem to follow from the underlying linguistic theory or view of meaning and their application to the languages in question. A traditional, instrospective definition of the meaning of a word is given by e.g. Zgusta, for whom meaning is composed of the components of **designation**, **connotation** and **range of use**. Hence, a restriction to synonymy is a difference in at least one of the components (Zgusta 1971: 89-90). Differences in connotation are to be understood as stylistic differences or neutral/general vs. genre-specific oppositions, e.g. *r* vs. *b* , whereas differences in the designation concern conceptual specifications to the central designation e.g. *r* is a type of *r* (ibid: 91). Differences in range of use can either be interpreted as topical restrictions or associations, e.g. *p* is a remuneration paid to a teacher, while *r* is a similar compensation paid to an official (ibid: 42). The observation of synonymy along these lines seems quite abstract and generalizing, pointing out finesses in the semantic relationships among words, rather than addressing

synonymy from the practical view point of actual possibilities of lexical variation and choice in the real-time context of an utterance.

In contrast to the introspective viewpoint, Sinclair among others has demonstrated that large-scale corpora can be used efficiently to analyze on an empirical basis the meaning of words and to theorize about the structure and nature of meaning in general. Sinclair goes as far as to say that "most everyday words do not have an independent meaning, but are components of a rich repertoire of multi-word patterns that make up a text" (Sinclair 1991: 108). A corollary of this would be that even the various inflected forms of a single individual base form [lemma] may often have very different usage patterns (ibid: 8). Applying this to synonymy, Biber (1998: 95-100) has demonstrated that words traditionally judged as practically synonymous, e.g. *r* and *b e* , with similar meanings and identical valency potentials, are actually "strikingly" different when their association patterns, i.e. lexical contexts and syntactic usage in corpora, are taken into consideration. Within this empirical and contextual approach to language use, it seems that a non-compositional and scalable definition of synonymy would be desirable, which can be adapted from Cruse's (2000: 156-160) treatment of the concept. Hence, synonyms are in this study interpreted on the basis of empirical, contextual evidence as words 1) whose semantic similarities are more salient than their differences; 2) that do not primarily contrast with each other; and 3) whose permissible differences must in general be either minor, backgrounded, or both.

Many of the contextual association patterns and restrictions on synonymy observed by Biber and Sinclair are, however, specific to English, e.g. fixed word order, or they are substantially more limited in English than in other languages, e.g. the number of possible inflected word forms per base form. On the one hand, one can very well imagine that association pattern types observed in English can have corresponding association pattern types in other languages, even though the actual grammatical surface structure or mechanism is quite different. On the other hand, one could expect that typologically different languages would resort to grammatical association patterns typical to each language. Though Sinclair hints at different association patterns for individual word forms of the same lemma, what really has not been followed through, as far as goes for the selectional restrictions and preferences of synonymous words, is the role of inflection and the preferences of inflectional full forms or features in languages where it really could matter, i.e. languages with a rich morphological system such as Finnish.[1] It is interesting to note here that already Zgusta (1971: 123-127) suggested that variation in meaning within the paradigm is rather frequent in morphologically rich languages, though he did not have the benefit of large corpora from which to extract empirical evidence on the issue. Nevertheless, synonym dictionaries, or general dictionaries for that matter, rarely include information about inflectional usage preferences.

This is not to say that the interaction of inflection and usage of synonyms has been overlooked altogether. Some significant observations have been made by Kangasniemi (1992) concerning the behavior of Finnish modal verbs in this respect, though his main emphasis has been on describing the expression of modality in general in Finnish. Using one of the earliest electronic corpora for Finnish, Kangasniemi has demonstrated for instance that there can be a suppletive relationship in the usage of two Finnish modal verbs in some inflected forms, namely 'can' and 'may', when these are used to denote epistemic possibility (ibid: 318-319, 331). Furthermore, he observed in the usage of other pairs of modal verbs with similar meaning

[1] The number of morphologically constructible forms is often calculated as 1,872 for Finnish nouns (2 numbers X 13 cases X 6 possessives X 12 clitics) and over 20,000 for Finnish verbs, the latter figure depending on how participle forms are counted in the figure ([530 finite forms + 320 infinitives] X 12 clitics + 5 participles X 1,872). The number of so-called core forms, ignoring clitics, is considerably smaller. Of all of these forms, only a fraction can be observed in even very large corpora of millions of words (personal observations of the author in context of this and earlier work).

other types of contextual preferences which are connected with the inflection of these verbs, concerning negation (vs. and *a* 'can' vs. *p a* 'can') and the grade of animacy of their subjects (*p a* and *a*) (ibid: 35, 71). For instance, *a* is associated more strongly than *p a* with human subjects, which one could expect to show up as a higher relative frequency of inflected forms of *a* that are associated with human subjects, namely the first and second person and passive forms. Concerning other parts of speech, Jantunen has recently observed inflectional preferences in a pair of near-synonymous Finnish adjectives, namely *a r a* 'important' and *e e e* 'central' (Jantunen 2001).

The structure of this paper is as follows. After this introduction, the research aims of this paper are given in Section 2. The linguistic data, i.e. the corpora used in this study and the research methods are presented in Section 3. In Section 4, the results of this study are laid out, which are interpreted and discussed in Section 5. Finally, the conclusions and suggestions for further research are given in Section 6.

2. Research aims

The research questions in this study are the following:

1. Does some syn(g)9.9.5(y)39.9.2rp67 Tm 0e1(ima)5(c)-1 Tc 0.1.04 65.2aper i9a1 , the rmyn(g)9 whigr

definitions in terms of the suggested substitutions and the more they could be judged substitutable with each other in the typically several usage examples that were given in the dictionary entries, the better they were considered for the actual test of substitutability with respect to their inflected forms. In the end, out of the near synonym set *ɶ - ɶap h - hr -* 'think, ponder, consider, reflect', the pair *ɶap h* 'ponder, think' emerged as the most promising candidate (cf. Pajunen 1982: 169, 180-182).

The method for assessing the substitutability of the chosen pair was to evaluate this for all the individual occurrences of their inflected forms in a corpus, and thus require apply *r ɶ ɶ* , i.e. require that interchangability applies for all (or practically all) cases.[2] A computer program was written by myself for detecting an occurrence of either word in a morphosyntactically analyzed corpus, showing both the original word with a lexical context of surface forms of a chosen size, which in this study was 10 words to both the left and right[3], when possible within the same sentence. As an example, (1) below is an original sentence from the corpus with the original verb, i.e. *ɶ < ɶa* , and (2) shows the same context with the synonym in the corresponding form, i.e.*p h <p h* . A morphological analysis of the first clause is given in (3) and an English translation of the entire sentence in (4).

(1) *ɶ p r , h ɶɶ p .*

(2) *h p r , h ɶɶ p .*[5]

(2)	Mietin	muuttoa	pari	vuotta, ...
	consider	moving	pair/a few	year
	V:ACT:IND:PAST:SG1	N:SG:PTV	PRON:SG:NOM	N:SG:PTV

(4) 'I considered moving for a year or two, I counted together the plusses and the minuses.'

After each occurrence, I had to choose whether the suggested new sentence fragment with the substituted form was synonymous with the original, and also indicating whether the underlying linguistic analysis was incorrect for later manual inspection and correction, which information was added to the underlying analyzed form of the corpus. The corresponding morphological forms of the synonym pair had been automatically extracted from a corpus, with the missing forms being added manually. Prior to the assessment of synonymy, the corpus had been automatically morphologically and syntactically analyzed and disambiguated at the Department of General Linguistics at the University of Helsinki with an implementation of Functional Dependency Grammar[6] developed by Conexor <http://www.conexor.fi>. After the substitutability judgements, I manually inspected and disambiguated the analyses of all occurrences of the two lexemes in question.

[2] N.B. WordNet is based on a weaker notion of entailment, where interchangibility at least in one context suffices for synonymy (Fellbaum 1998: 77; Alonge et al. 1998: 21)

[3] This context seemed more than enough for determining interchangability, and one might note that in a study by Kaplan (1955) for English, mentioned by Leacock et al. (1998: 266), informants had been found capable of disambiguating word senses with a window of ±2 words. However, in the case of the underlying morphological analyses, there were a few ambiguous words which needed looking at several preceding paragraphs to choose the right interpretation (specifically choosing between the present of past tense analyses of the verb), and one case that would have required extratextual information relating the time of the publication of the article with the time of the depicted historical event.

[4] In article ID-number 7786 in the newspaper corpus

[5] Sentence (2) was deemed to mean the same as sentence (1), and thus at least in this case the forms were deemed to be synonymous and substitutable.

[6] A general description of the underlying formalism is given by Tapanainen and Järvinen (1997).

Two corpora were used. As the first one, a portion of the Finnish newspaper [illegible] <http://www.keskisuomalainen.fi/>, stored in the [illegible] 'Text Bank [of Finnish]' <http://www.csc.fi/tekstipankki/>, at the Center for Scientific Computing <http://www.csc.fi/> was chosen. The portion, of which headings, subheadings and lists were excluded, amounted to 2,054,413 words, consisting of 8,251 articles published between January 2 and April 31, 1994, representing all the sections of the newspaper except advertisements. In respect to these articles, one could identify 77 journalists who had written at least one article by themselves, adding up to 3,428 articles with a single author. Thus, 4,823 of the articles either had multiple or unspecified authors, typically having a national or international news agency as their source. A reason for choosing specifically this newspaper corpus was that it included extensively extra-linguistic information, i.e. each article was marked for its author (though partially as noted above), the section of the newspaper in which it was published, the date of publication plus other data, which could be used to ensure that the observed linguistic phenomena were not caused by other factors.

In this corpus, there were 410 occurrences of [illegible] representing 49 unique word forms, and 445 occurrences of *p h* , representing 45 unique word forms. Of the morphological analyses of these unique forms, 25 were common. The most frequent form for both was the active indicative present tense third person singular, namely 85 occurrences of [illegible] and 145 occurrences of *p h* . Either one of the words appeared at least once in 714 of the articles, at the most 5 times in an individual article and typically 1-2 times per article. The highest number of uses of either word by a single author was 25 times, by two distinct authors. Of the 77 uniquely identifiable authors in the corpus, 58 used either verb at least once, amounting to 371 occurrences. Thus, there were 484 occurrences for which a unique author could not be attributed. Among the occurrences with an identified author, there were 15 who used only *a* throughout the corpus, 14 who used *a* more than *p h* , 8 who used both verbs equally often, 13 who used *p h* more than *a* , and 8 who used only *p h* . This basically means that approximately 40% of the authors used only one of the two verbs through the entire corpus, though this represented only 68 of all the occurrences of the two verbs. Nevertheless, this could be of significance in interpreting the results.

Since the sum frequency of the two verbs in the first research corpus was not that high, at least for drawing strong statistical assumptions, an aggregated corpus of several Finnish newspapers (including the above-mentioned first corpus) was used as the second research corpus, available at the same site. This second corpus amounted to roughly 15.8 million words, but it lacked the extra-linguistic information present in the first corpus. Thus no information can be given on the number of authors or articles, though one could expect that these figures are roughly similar to those of the first research corpus. The second corpus contained all in all 4,545 instances of either word, being divided quite equally to 2,135 instances of *a* and 2,410 instances of *p h* . These instances represented 83 different unique inflected forms for *a* and 82 unique inflected forms for *p h* , of which forms 59 had a common morphological analysis. Since the first corpus gave a very strong indication that the two verbs were interchangeable in practically every case, the assumption was made that this would apply in the second corpus, as it did represent the same general genre.

After the first research corpus had been analyzed in terms of the substitutability of the two lexemes in the various occurrences in their different inflected forms and the underlying linguistic analysis had been verified, those linguistic analyses representing either lexeme under study were extracted from the corpus for closer statistical analysis. In addition to the purely linguistic analysis tags, each analysis was supplemented with an anonymized tag representing the author of the text (of the type META-BY_aaa), a running unique identification number for each article (e.g. META-ID_ks94_123), and the section the article had appeared in (e.g. META-DE_foreign 'foreign affairs' or META-DE_sports 'sports').

The reason for including the extra-linguistic information can be found in the two working hypotheses used in this study. The first of them was that any information available, whether linguistic or not, can be treated basically in a similar fashion and combined in any combination freely, e.g. some linguistic tag or tag combination corresponding to a linguistic feature or an entire inflected form with some or more of the extra-linguistic tags. A large majority of such combinations will turn out to be singular occurrences so that they will become redundant by any statistical test or cut-off point, but this full scale application of combinatorics allows for the possibility of the most common and possibly statistically most significant combinations or underlying sub-combinations of features, should one say abstractions of patterns in the corpus, to rise above the ocean of random combinations. On the other hand, having ready at hand all the possible combinations observed in a corpus, one can start from individual significant features or feature combinations at the top and work down the list to observe the network of their less frequent occurrence contexts. These aspects will be illustrated later in this paper. The second working hypothesis was that inflectional information (and other extra-linguistic data presented above), i.e. the tags produced by linguistic and other analysis, can be studied and treated in a similar fashion as collocate words, for in a sense Finnish words often correspond to a multiword expression or contruction in English, e.g. the subject-verb construction in English vs. personal inflection of verbs in Finnish. Thus, one could deem it justifiable to apply the same statistical measures and tools as are used in studying collocates.

In this study, a simplified version of t-score[7] as presented and defined by Church et al. (1991) was initially selected to highlight differences in usage-based preferences of individual features, partial feature combinations or entire feature sets between the two lexemes. The motivation for this was that the object of study appeared in principle to be very similar to the comparison of dissimilarities of lexical collocates of two English synonymous adjectives, namely *p erf* and *r* , which was the example given by Church et al. as a use for t-score. The t-score as defined by Church et al. should apparently not be used as statistical proof of the significance of associations as it is based on the assumption of normal distribution which does not accurately apply to word frequencies, but it combines in a practical way both the ratio of the associations of some particular feature in a dichotomous case and the number of cases on which this ratio is based on, and thus neatly orders preferences of this type observed in a corpus for further qualitative scrutiny and actual interpretation (cf. Stubbs 1995 for a thorough assessment of the benefits, limitations and motivated usage of t-score). Thus, in the case of similar ratios, those based on a higher number of cases receive a higher t-score and are consequently ranked higher. Other methods have been presented that might provide statistically more reliable results, e.g. Fisher's exact text (Pedersen et al 1996), and could naturally be used at later stages of this study. Nevertheless, t-score appears presently to be a standard tool in lexicographic work and available lexicographic software despite its shortcomings.

4. Results

The results for the first research corpus are presented first. All in all, 1,690,862 different combinations of tags representing various verbal morphological features were generated for *e a* and *p h* , based on the 855 occurrences of the two verbs. If one could allow oneself to interpret in a statistical sense the calculated t-scores, quite a few would be statistically significant, i.e. having a value of 1.645 or more, though one must note that the t-score test was firstly used as a tool for analysis and not for testing a specific predetermined hypothesis.

[6] $t = \frac{n(\text{base, feature}) - n(\text{base})*n(\text{feature})*n(\text{both bases})}{\sqrt{n(\text{base, feature})}}$, where n(x) is the absolute frequency of x

Nevertheless, there are some very interesting association tendencies; for instance, as many as 21 cases out of 22 of the first person singular feature (SG1) were associated with *e a* .

b e 1: A e e f h e r e f h e - r e f r h e p r e f e r e e f e a p h h r p h f e r e h e b f h e f r p (e p r e b : , h e r e e r e b 'D 2.0; ⇔ A B e e f f r h h h e - r e f r e e b , p e b r f r A r B)

#	t-score	$n_{tag(verb)}$/ $n_{tag(total)}$	Verb	Combination of tags/features
1	2.6544151	77/112	miettiä	INF1 ~ *e a* 'to ponder' ⇔ &-MV:V:ACT:**INF1** ~ *e a* 'to ponder'
2	2.3389739	22/23	miettiä	IND:SG1
3	2.3003402	199/320	pohtia	SG3
4	2.2804408	21/22	miettiä	SG1
5	2.2720630	198/319	pohtia	IND:SG3
6	2.2526106	32/37	pohtia	META-DE_foreign
7	2.1700721	201/355	miettiä	&-MV
8	2.1004193	145/230	pohtia	PRES:SG3 ⇔ &+MV:V:ACT:IND:**PRES**:**SG3** ~*p h* 'he ponders'
9	2.0711558	21/24	miettiä	META-BY_aaa
10	2.0627680	30/39	miettiä	&-MV:META-DE_sport
11	2.0157633	15/15	miettiä	&+MV:PAST:SG1 ⇔ **&+MV**:V:ACT:IND:**PAST**:**SG1** ~ *e* 'I pondered'
12	2.0103638	203/335	pohtia	&+MV:PRES
13	2.0025643	204/337	pohtia	IND:PRES
14	1.9871152	206/341	pohtia	PRES
15	1.9619883	16/17	miettiä	SG1:META-BY_unspec
16	1.9544431	286/486	pohtia	&+MV:IND
17	1.9242580	24/28	pohtia	ACT:META-DE_foreign
18	1.9121518	288/491	pohtia	IND
19	1.8786721	116/184	pohtia	SG3:META-BY_unspec
20	1.8690776	18/21	miettiä	ACT:META-BY_aaa
21	1.8192520	14/15	miettiä	PRES:META-BY_aaa
22	1.8090504	290/498	pohtia	&+MV
23	1.8074797	26/35	miettiä	&-MV:ACT:META-DE_sport
24	1.7435752	13/14	miettiä	&+MV:PRES:META-BY_aaa ⇔ &+MV:V:ACT:IND:**PRES**:SG3:**META:BY_aaa** ~ *e* 'he ponders'
25	1.6216452	165/277	pohtia	&+MV:ACT:PRES
26	1.6128154	61/93	pohtia	&+MV:PASS:IND
27	1.5904283	11/11	pohtia	IND:META-BY_bbb
28	1.5884516	15/17	pohtia	PRES:META-DE_foreign
29	1.5816121	11/12	miettiä	ACT:SG:META-DE_sport
30	1.5626323	20/25	pohtia	META-BY_bbb
31	1.5614035	9/9	miettiä	META-BY_ccc
32	1.5614035	9/9	miettiä	META-BY_ddd

LEGEND FOR MORPHOLOGICAL TAGS IN TABLE 1(in alphabetical order):

ACT	active voice	**&-MV**	non-finite verb forms, i.e. infinitives, participles and compound forms in tenses and negation
IND	indicative mood		
INF1	first infinitive forms		
META-BY_xxx	written by xxx	**PRES**	present tense
META-DE_xxx	newspaper section xxx	**PAST**	past tense
&+MV	finite verb forms	**SG1**	first person singular
		SG3	third person singular

A selection with the highest values of t-scores high-lighting these associations are provided in Table 1. If any tag combinations in the full list of possible combinations have both the same t-score values and frequencies, and furthermore contain one or more similar tags, only the tags common to the entire group, and thus the most informative ones, are provided, the only exception being tag combinations which define fully and exactly a unique word form. Furthermore, some redundancy in the morphological description of the linguistic analyser has been removed, for instance the part-of-speech tag for verbs (V). For example, the tag combinations &+MV:ACT:IND:PAST:SG1, &+MV:ACT:PAST:SG1, &+MV:IND:PAST:SG1, &+MV:PAST:SG1, &+MV:V:ACT:IND:PAST:SG1, &+MV:V:ACT:PAST:SG1, &+MV:V:IND:PAST:SG1, &+MV:V:PAST:SG1, all having a t-score of 2.0157633 when occurring together with *e a* , are represented simply by **&+MV:PAST:SG1**, being the morphological feature tags for finite verb form, past tense, and first person singular, respectively, and **&+MV:V:ACT:IND:PAST:SG1**, which is the exact and complete morphological analysis of the [finite] active indicative past tense first person singular word form of *e a* , i.e. *e* 'I pondered'.

From a purely linguistic viewpoint, there appears to be a clear preference for using *e a* in the first person singular forms (row 4 in the table) and more specifically in conjunction with the indicative mood (row 2). A full word form containing these three aforementioned features, i.e. *e* 'I pondered' follows quickly in the table (row 10). Furthermore, there is a preference for using *e a* in the first infinitive form (row 1). On the other hand,*p h* has some level of preference with the third person singular forms in its active indicative forms (rows 3 and 5), and more specifically in conjunction with the present tense, which features appear together in the full form*p h* 'he ponders' (row 8). On a more general level, usage with the present tense and the indicative mood in general, both separately and together, have a preference with*p h* (rows 13, 14 and 18). Finally, *e a* shows to have a slight tendency towards the non-finite forms whereas*p h* tilts towards the finite forms (row 7 versus 22), among which*p h* has further a preference for indicative forms (row 16).

From an extra-linguistic viewpoint, articles in general placed in the foreign affairs sections seem to use*p h* (row 6), and often in its active voice forms (row 17), whereas articles in the sports section have a predisposition for *e a* (row 10) apparently often in its non-finite and active forms (rows 10 and 23). Finally, there is one author ( ) who has a strong preference for *e a* (row 9), in its active voice and present tense forms (rows 20 and 21), and specifically for a full form containing all these features, *e* 'he ponders' (row 24). Two other authors prefer *e a* (and on rows 31 and 42), but on the other hand, author*bbb* prefers*p h* , in its indicative forms (rows 27 and 30). It is interesting to note that a feature which demonstrates one of the strongest preferences, namely first person singular with *e a* , is also strongly attached with unspecified authorship (row 15). In practice this means that we cannot connect this tendency further as belonging to an author's general predisposition for using *e a* , but neither can we rule out the possibility of such explanation without either knowledge about the authors and, preferably, considerably more examples on the usage of this particular feature.

b e 2: A e e f h e r e f h e - r e f r h e p r e f e r e e f e a p h h r p h f e r e h e b f h e f r p (e p r e b : , h e r e e r e b 'D 2.0; ⇔ A B e e f f r h h h e - r e f r e e b , p e b r f r A r B)

#	t-score	$n_{tag(verb)}$/ $n_{tag(total)}$	Verb	Combination of tags/features
1	5.02810160	417/670	miettiä	**INF1** &-MV:V:ACT:**INF1** ⇔ *e a* 'to ponder'
2	4.90857618	1141/2076	miettiä	&-MV
3	4.65436111	341/481	pohtia	PASS:IND
4	4.60405240	88/96	miettiä	**SG1**
5	4.60140472	332/468	pohtia	&+MV:PASS:IND
6	4.47394688	336/479	pohtia	&+MV:PASS
7	4.30643067	1474/2468	pohtia	&+MV
8	3.66672617	53/56	miettiä	**PAST**:**SG1** &+MV:V:ACT:IND:**PAST**:**SG1** ⇔ *e* 'I pondered'
9	3.42270008	173/242	pohtia	**IND**:**PL3**
10	3.39827308	932/1562	pohtia	**SG3**
11	3.38888039	111/142	pohtia	**PASS:PAST** &+MV:V:**PASS**:IND:**PAST** ⇔*p h* 'X was pondered'
12	3.37272932	924/1551	pohtia	**IND**:**SG3**
13	3.35194627	178/252	pohtia	**PL3**
14	3.29908324	224/330	pohtia	**PASS**:**PRES**
15	3.23807959	220/326	pohtia	**&+MV**:**PASS**: **PRES** **&+MV**:V:**PASS**:IND:**PRES** ⇔*p h* 'X is pondered'
16	2.87031644	289/453	pohtia	**PAST**:**SG3**
17	2.79006234	36/41	miettiä	**PRES**:**SG1** &+MV:V:ACT:IND:**PRES**:**SG1** ⇔ *e* 'I ponder'
18	2.53231288	122/178	pohtia	**PRES**:**PL3** &+MV:V:ACT:IND:**PRES**:**PL3** ⇔*p h* 'they ponder'
19	2.39208299	330/610	miettiä	**INF3**
20	2.38941082	51/64	pohtia	**PAST**:**PL3** &+MV:V:ACT:IND:**PAST**:**PL3** ⇔*p h* 'they pondered'
21	2.24091027	51/66	pohtia	**ACT:PCP1:SG**
22	2.23804303	71/111	miettiä	**PASS**:**PCP1**
23	2.13260209	635/1097	pohtia	**PRES**:**SG3** &+MV:V:ACT:IND:**PRES**:**SG3** ⇔*p h* 'he ponders'
24	2.05325910	23/28	miettiä	**IMP**

LEGEND FOR MORPHOLOGICAL TAGS IN TABLE 2(not included in Table 1)

IMP imperative mod
INF3 third infinitive forms
PASS passive voice
PCP1 first participle forms
PL3 third person plural form

One can nevertheless some make observations based on the few cases that do have such author-specific information, traversing down the full list of tag/feature combinations. Though the usage of first person singular forms is spread over 22 articles in the first research corpus, it turns out that in those two cases where it is appears more than once in the same article, it appears with the same lexeme. Furthermore, of the four identifiable individual authors who used a first person singular form, all used it in conjunction with *e a* and consequently none with*p h* – two used *e a* twice (the first in the same article and the second in two separate articles) and the other two once. Of the four authors, two belonged to the group that throughout the corpus used only *e a* in any inflected form (9 and 2 times, respectively); of the two others one used*p h* somewhat more than *e a* (10 vs. 5) and the other *e a* somewhat more than*p h* (6 vs. 3). From all this one can definitely only conclude that no interpretations on the effect or non-effect of authorship on the preference of the first person singular together with *e a* can be given or ruled out.

The first corpus appears to bring forth quite a deal of the influence of extra-linguistic factors in the selection of the two verbs, in addition to the clearly observable inflectional preference tendencies. Based on the second corpus I attempted to validate whether these tendencies of inflectional forms and features and their combinations will be continue to exist, the results of which are presented in Table 2 above. The results based on the second corpus appear to be parallel to those of the first corpus, the main distinction being that the number of cases on which to assess the strength of an tendency is clearly much higher, and thus also the corresponding t-score values. All in all, there were 4,490 tag combinations to evaluate, the number being considerably lower compared to the first corpus as a result of the lack of extra-linguistic tags. Some of the most general and most frequent features, which have already appeared in Table 1 and discussed above, e.g. the indicative mood and the present and the past tense both separately and in combination with each other have been omitted from Table 2. Resultwise, the first person singular in the indicative mood continues to be very strongly associated with *e a* (row 4), whether in the past or the present tense (rows 8 and 17, respectively). The larger corpus contributes to new features entering the picture, where the third person plural in the indicative mood (row 9) and furthermore in the present tense (row 18) and the past tense (row 20) appears to be associated with*p h* , and even regardless of mood (row 13), as is also the case with the third person singular (row 10) in the indicative mood (row 12) in the past tense (row 16) and the present tense (row 23). Regarding the passive voice in its finite usage, it leans towards*p h* in the indicative mood (rows 3, 5 and 6), somewhat lesser but still in the past tense (row 11) and even in the present tense (rows 14 and 15). As far as non-finite forms go, *e a* is the preferred one of the two (row 2), for instance in the case of the first and the third infinitive forms (rows 1 and 19), whereas*p h* has a closer tie with the finite forms (row 7). In the case of participle forms, the active singular forms of the first participle are associated with*p h* (row 21), whereas the corresponding passive forms of the first participle are more predisposed for *e a* (row 22). Finally, it is interesting to note that *e a* is predominantly the lexeme chosen in the imperative mood (row 24).

5. Discussion

On the basis of the results, the two verbs, *e a* and*p h* , do appear to differ substantially in their inflectional profiles and individual morphological features, and the results are similar in both research corpora. If I assume that the surprisingly free interchangability between the two verbs as it was assessed occurrence by occurrence in the first corpus in order to rule out the effect of the surrounding lexical context, often attributed as the cause of selection one way or another, also holds for the second corpus, I could claim on the basis of the observed numbers from this larger corpus that this study has indeed shown preferential differences between the two concerning purely individual inflectional features and their combinations. A measurement

that would be statistically more reliable than the type of t-score used in this study would nevertheless be very desirable to validate the significance of the results. One can also question whether the evaluation of interchangability is as reliable as it should be when undertaken by a single individual, representing his own idiolect and being aware of the aims of the study.

It does not suffice to merely present observations and possibly indicate where present linguistic descriptions, especially concerning morphologically rich languages such as Finnish, are lacking, but one also needs to attempt to provide an explanation. During the course of this work, especially through seeing the actual usage contexts of the two verbs both through evaluating their interchangability and validating and disambiguating their morphosyntactic analyses, it seems that the inflected forms and features by themselves cannot provide a comprehensive answer to this question. Despite all my attempts to rule out the effect of word-external context, many of the inflectional features are interconnected with the lexical and syntactic context, for instance all third person forms of a verb are the result of a third person subject in the sentence. It could very well be that the strong association of third person plural forms with*p h* is also associated with (third person) plural subjects of a particular semantic type. Which is the key determining factor remains an object of further study.

The fact that 40% of the authors in the first corpus used only one of the two lexemes could be a reason to suspect that the choice between the two words would be determined on a general level by a person's idiolect (and perhaps dialect) rather than some general preferences between the two lexemes. On the other hand, the relatively low total number of occurrences, amounting to only 65 cases (18 percent of the identifiable unique authors), in the usage of either word by these single lexeme authors might rather indicate that this is a result of these authors simply having used either verb so seldom (between 1 and 5 times per author) that they have not had a chance to vary their usage, rather than some categorical preference one way or the other. As a follow-up study, one should most probably not only try to focus on the effect of author-specific tendencies on the usage of particular verbs and features, but also other extra-linguistic features such as genre and text type, a practice that Biber (1998) strongly encourages.

6. Conclusion and suggestions for further work

This study clearly shows that synonymous words can have purely inflectional differences, but the reasons for this remain still open. On the basis of the earlier work (Kangasniemi's results) and this study, it would seem that more attention should be paid to the possible selectional restrictions or preferences of inflected forms in descriptions of lexical relationships in Finnish, and probably also in other morphologically rich languages. This would and should have an impact on lexicographical work and the contents of lexicographical products such as dictionaries for such languages. This line of research would benefit from validation with other synonym pairs or larger groups than the ones observed in this study, and from considerably larger numbers of occurrences. The effect of the writer's idiolect and dialect as well as genre and text type should very clearly be looked into. Finally, it may even be worthwhile to study this issue also in the morphologically simpler languages.

Acknowledgements

I greatly appreciate the many comments on various versions of this paper by Kari K. Pitkänen and Fred Karlsson, the theoretical and methodological discussions with Hanna Westerlund and Anu Airola, and the help received from Jussi Piitulainen and Kari T. Vasko to get statistical issues resolved. I would also like to thank , the National Technology Agency of Finland, for the financial support which has made this research possible under the auspices of the USIX/GILTA project (40943/99). Finally, I would like to thank Anu Airola and Krista Lagus for sharing the analysed version of the second research corpus.

References

Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti, M. A., and Peters W. (1998). The Linguistic Design of the EuroWordNet Database. In Vossen P., editor. *r r . A D b h r*, pages 19-43. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Biber D. (1998). *C p : r r*