Multivariate methods in the corpus-based lexicography
A study of synonymy in Finnish

Antti Arppe
Department of General Linguistics
University of Helsinki
antti.arppe@helsinki.fi

1. Introduction and background

The purpose of this paper is to present a case study of how multivariate statistical methods such as polytomous logistic regression can be adapted to discover and analyze the wide and complex range of linguistic factors which both influence and interact in the selection and usage of sets of more than two near-synonyms. The results reported in this paper are a follow-up of Arppe (2006), and a preliminary version of those to be presented in full in Arppe (forthcoming).

In the modeling of lexical choice among semantically similar words, specifically near-synonyms, it has been suggested in computational theory that (at least) three levels of representation would be necessary to account for fine-grained meaning differences and the associated usage preferences, namely a 1) conceptual-semantic level, a 2) subconceptual/stylistic-semantic level, and a 3) syntactic-semantic level (Edmonds and Hirst, 2002). With regards to the syntactic-semantic level, it has in the been shown in (mainly) lexicographically motivated corpus-based studies of actual lexical usage that semantically similar words differ significantly as to the 1) lexical context (e.g. English *powerful* vs. *strong* in Church *et al.*, 1991), the 2) syntactic structures which they form part of (e.g. English *begin* vs. *start* in Biber *et al.*, 1998), and the 3) semantic classification of some particular argument (e.g. English *shake* verbs in Atkins and Levin, 1996), as well as rather style-associated 4) text type, in which they are used (e.g. Biber *et al.*, 1998).

In addition to these studies that have focused on English, with its minimal morphology, it has also been shown for languages with an extensive morphological system, such as Finnish, that similar differentiation is evident as to the 5) inflectional forms and the associated morphosyntactic features in which synonyms are used (e.g., the Finnish adjectives *tärkeä* vs. *keskeinen* 'important, central' in Jantunen, 2001, and Finnish verbs *miettiä* and *pohtia* 'think, ponder, reflect, consider' in Arppe, 2002, Arppe and Järvikivi, forthcoming). Recently, in their studies of Russian near-synonymous verbs denoting 'try' and 'intend', Divjak (2006) and Divjak and Gries (2006) have shown that there is often more than one type of these factors in play at the same time, and that it is therefore worthwhile to observe all categories together and in unison rather than separately one by one.

All of these studies of synonymy have focused on which contextual factors differentiate words denoting a similar semantic content. In other words, which directly observable factors determine which word in a group of synonyms is selected in a particular context. This general development represents a shift away from more traditional arm-chair introspections about the connotations and range-of-use of synonyms, and it has been made possible by the accelerating development in the

last decade or so of corpus linguistic resources, i.e. corpora, and tools, e.g. parsers and statistical programs.

Entry	**Single-word definitions**
Miettiä	**Punnita**, **harkita**, **ajatella**, **järkeillä**, **tuumia**, mietiskellä, pohtia, suunnitella, **aprikoida**
Pohtia	**Punnita**, **harkita**, **ajatella**, **järkeillä**, **tuumia**, miettiä, **aprikoida**

Table 1. Single-word definitions in the electronic version of *Suomen kielen perussanakirja* 'Standard Dictionary of Finnish' (Haarala, *et al.*, 1997) provided for *miettiä* and *pohtia*; common lexemes in **boldface**.

Entry	**Synynym group**
Miettiä/Pohtia	(*miettiä*) **miettiä**, mietiskellä, **pohtia**, pohdiskella, harkita, tuumia, aprikoida, järkeillä, puntaroida, punnita, punnita, tuumata, tuumailla, hautoa, filosofoida, meditoida, spekuloida, funtsata, funtsia

Table 2. Synonym set provided for both *miettiä* and *pohtia* in *Nykysuomen sanakirja VII: Synonyymisanakirja* 'Dictionary of Modern Finnish VII: Synonym dictionary' (Jäppinen, 1989), which is constructed around *miettiä* for both lexemes, and is thus the same.

In the explanation of such observed results in both a linguistically meaningful and scientifically valid way, Gries (2003a: 32-36) has made a compelling argument in favor of holistic and multivariate approaches, in contrast to a traditional tendency for monocausal hypotheses and explanations. Nevertheless, these multivariate methods do build upon univariate and bivariate analysis, as Gries (2003a) also demonstrates. Furthermore, as has been pointed out by Divjak and Gries (2006), the majority of the above and other synonym studies appear to focus on word pairs, perhaps due to the methodological simplicity of such setups. However, it is clearly evident in lexicographical descriptions such as dictionaries that there are often more than just two members to a synonym group.

Take for instance the single-word definitions given in a current general dictionary of Finnish (Table 1) for the two earlier studied THINK lexemes *miettiä* and *pohtia*, and the set of their synonyms given in a slightly older synonym dictionary (Table 2). In both dictionaries, we can clearly see that the two lexemes form part of larger grouping of semantically similar lexemes, which in the synonym dictionary are in fact exactly the same set for both verbs, although this original pair was assessed to be closest to each other and therefore selected as the focus of deeper scrutiny in the earlier, afore-mentioned studies. Nevertheless, it may be that the observed differences between this particular pair would receive a different interpretation in the overall perspective when studied within the entire synonym group, or at least among its most frequent members. Within the larger group, the studied pair might contrast more with some other member or members than with each other.

Lexeme	Absolute frequency	Natural logarithm of relative frequency	Rank (among verbs)
Pohtia	30572	-6.7	127
Ajatella	29877	-6.7	130
Miettiä	27757	-6.8	141
Harkita	14704	-7.5	257
Tuumia	4157	-8.7	595
Punnita	2253	-9.3	828
Aprikoida	1293	-9.9	1153
Mietiskellä	995	-10.1	1345
Hautoa	536	-10.8	1939
Filosofoida	399	-11.1	2281
Järkeillä	308	-11.3	2589
Funtsata	29	-13.7	5996

Table 3. Absolute frequencies, the natural logarithms of the relative frequencies, and the corresponding rank among verbs of the entire group of THINK lexemes, calculated using the Finnish Text Collection (2001) which contains approximately 25.4 million instances of verbs (accounting for roughly 14% of all the running word tokens), sorted according to descending frequency.

Therefore, other commonly used members of the Finnish THINK synonym group, with frequencies within the same (relatively high) magnitude as the original pair, were included in this follow-up study, resulting in *ajatella*, *miettiä*, *pohtia* and *harkita* out of the entire set evident in Tables 1 and 2. The frequency counts (shown in Table 3) were based on the Finnish Text Collection (2001), which is presently the largest available corpus of uniformly processed Finnish text, containing approximately 180 million words. As can seen, the other lexemes in the THINK synonym set are clearly of a lower magnitude of frequency, and were thus excluded in this study. In addition to broadening the set of studied synonyms, the entire syntactic argument structure associated with the studied lexemes was also to be included in the examined context, as well as the semantic classifications of the most common syntactic argument types.

2. Research corpora and linguistic and statistical methods

The research corpora consisted of two months worth (January–February 1995) of written text from Helsingin Sanomat (1995), Finland's major daily newspaper, and six months worth (October 2002 – April 2003) of written discussion in the SFNET (2002-2003) Internet discussion forum, namely regarding (personal) relationships (`sfnet.keskustelu.ihmissuhteet`) and politics (`sfnet.keskustelu.politiikka`). The newspaper corpus consisted altogether of 3,304,512 words of body text, excluding headers and captions (as well as punctuation tokens), and included 1,750 representatives of the studied THINK verbs, whereas the Internet corpus comprised altogether 1,174,693 words of body text, excluding quotes of previous postings as well as punctuation tokens, adding up to 1,654 representatives of the studied THINK verbs. As can be immediately observed, the proportion of the THINK lexemes in the Internet newsgroup discussion text is more than twice as high as the corresponding value in the newspaper corpus. The individual overall frequencies among the studied THINK lexemes in the research corpora were 1492 for *ajatella*, 812 for *miettiä*, 713 for *pohtia*, and 387 for *harkita*.

The corpora were first automatically syntactically and morphologically analyzed using a computational implementation of Functional Dependency Grammar (Tapanainen and Järvinen,

1997, Järvinen and Tapanainen 1998) for Finnish, namely the FI-FDG parser (Connexor 2007). After this, all the instances of the studied THINK lexemes together with their syntactic arguments were manually validated and corrected, if necessary, and subsequently supplemented with semantic classifications. Each nominal argument (in practice nouns or pronouns) was semantically classified into one of the 25 top-level unique beginners for (originally English) nouns in WordNet (Miller, 1990). Furthermore, subordinate clauses or other phrasal structures assigned to the PATIENT argument slot were classified following Pajunen (2001) into the traditional types of participles, infinitives, indirect questions, clause propositions indicated with the subordinate conjunction *että* 'that', and direct quotes with attributions of the speaker using one of the studied THINK lexemes (e.g. *"..." miettii/pohtii joku* '"..." thinks/ponders somebody').

This covered satisfactorily AGENTS, PATIENTS, SOURCES, GOALS and LOCATIONS among the frequent argument types as well as INSTRUMENTS and VOCATIVES among the less frequent ones. However, other argument types which were also frequent in the context of the studied THINK lexemes, indicating MANNER, TIME (as a moment or period), DURATION, FREQUENCY and QUANTITY, had a high proportion of adverbs, prepositional/postpositional phrases and subordinate clauses (or their equivalents based on non-finite verb forms). These argument types were semantically classified following the *ad hoc* evidence-driven procedure proposed by Hanks (1996), in which one scrutinizes and groups the individual observed argument lexemes or phrases in a piece-meal fashion, as the contextual examples accumulate, and thus generalizes semantic classes out of them, without attempting to apply some prior theoretical model.

For instance, among the arguments of MANNER, *tarkkaan, tarkoin* < *tarkka* 'carefully/meticulously', *vakavasti* 'seriously', *oikeasti* 'really/earnestly', *perusteellisesti* 'thoroughly', *tarkasti* 'thoroughly', *huolellisesti* 'carefully', *syvään* < *syvä* 'in depth' became eventually classified to denote THOROUGHness, in contrast to *helposti* 'helposti', *pinnallisesti* 'superficially', *yksioikoisesti* 'simply', *yksipuolisesti* 'one-sidedly', *kapea-alaisesti* 'narrowly', *pinnallisemmin* 'more superficially' and *suppeasti* 'narrowly' which were understood to denote together various types of (abstract) SHALLOWness. Only in the case of MANNER arguments there emerged several levels of granurality, with the THOROUGH class going under other POSITIVE associations and the SHALLOW under generally NEGATIVE assessments, both of these two forming together the class of EVALUATIVE expressions of MANNER. Even though clause-adverbials (i.e. META-comments such as *myös* 'also', kuitenkin 'nevertheless/however' and *ehkä* 'maybe' as well as subordinate clauses with *mutta* 'but' and *vaikka* 'although') were also relatively quite frequent as an argument type, they were excluded at this stage due to their generally parenthetical nature. However, as an extension to Arppe (2006) the modality of the verb chains in which the studied THINK lexemes form part of were classified following Kangasniemi (1992) and Flint (1980), as well as were those other verbs which are syntactically in a co-ordinated (and similar) position in relation to the studied THINK lexemes, following Pajunen (2001).

After these automated and manual analyses and annotations, the frequencies of the resultant features were retrieved and counted from the research corpora. With respect to morphological variables, I chose to use analytic features characterizing the entire verb chain of which the studied THINK lexemes were components of, concerning polarity (i.e. NEGATION vs. AFFIRMATION), voice, mood, tense and person/number, and in general not to include any of the structural morphological features manifested only in the studied THINK lexemes themselves. Moreover, in a further

abstraction in comparison to Arppe (2006), the six distinct person/number features (e.g. FIRST PERSON SINGULAR, FIRST PERSON PLURAL, SECOND PERSON SINGULAR, and so on) were decomposed as a matrix of three person features (FIRST vs. SECOND vs. THIRD) and two number features (SINGULAR vs. PLURAL). In all, 108 contextual variables in the corpora turned out to exceed a minimum threshold of 24 occurrences (according to the so-called Cochran criteria) necessary for the analysis of homogeneity or heterogeneity of the distribution of such a feature among the studied lexemes by using the χ^2 statistic; however, the results of these univariate analyses will not be presented here, because the possible lack of heterogeneity in the distribution of some individual feature does not rule it out from inclusion in multivariate analysis, if its inclusion is otherwise theoretically motivated (Harrell, 2001: 56, 61, see also Bresnan *et al.*, 2007).

Among these individual features, a number were excluded due to their high (positive or negative) correlation with some other feature and/or because of obvious logical symmetric complementarity (e.g., all instances of verbs are either ACTIVE or PASSIVE but not both), partial (directed) compositionality (e.g., all instances of verbs with any person/number feature are by definition also ACTIVE forms), or simple overlap in the linguistic description (e.g., a negated form is always accompanied with a negative auxiliary verb), so as to reduce collinearity in the subsequent multivariate analysis. Furthermore, in the case that there was *only one* semantic classification for some syntactic argument type which exceeded the threshold frequency, the syntactic argument type alone was selected instead of the combined feature of syntactic role plus semantic class. Moreover, within each syntactic argument type at a time, I opted to use the same level of granularity in order to guarantee maximal coverage in the data and to avoid overlapping (in the form of double or multiple-level classifications). After this pruning, 59 contextual feature variables remained, of which 11 were morphological, 5 simple syntactic arguments, and 43 combinations of syntactic arguments with semantic classifications.

There are many possible foci of interest for multivariate analysis, but mine concern firstly the relative weights and differences in the impact of the individual feature variables which have been identified as pertinent in the preceding univariate and bivariate analyses, and secondly how well overall the selected variables are able to explain and account for the studied phenomenon. Furthermore, from prior research (e.g., Arppe and Järvikivi, forthcoming, Featherston, 2006) we know that that in practice individual features or sets of features are *not* observed in corpora to be categorically matched with the occurrence (in a corpus) of only one lexeme in some particular synonymous set and no others. Rather, while one lexeme in a synonymous set may be by far the most frequent for some particular context, others do also occur, albeit with often a considerably lower relative frequency. With this in mind, the representation of linguistic reality in multivariate analysis is probably more accurate when we study the *expected probabilities of occurrence* of all the individual lexemes belonging to a synonymous set, given some contextual features, instead of a discrete choice of only one of the four alternative lexemes at a time.

For this purpose, *polytomous* (alternatively also referred to as *multinomial*, *multiple-category*, *multiple-class*, *polychotomous*, or even, *discrete-choice*) *logistic regression* analysis (see e.g. Hosmer and Lemeshow, 2000: 260-287) is an attractive multivariate approach. As a *direct probability model* (Harrell, 2001: 217) polytomous as well as binary logistic regression yields probability estimates, corresponding to the expected proportions of occurrences, conditional on the values of the explanatory variables that have been selected for inclusion in the model. With respect

to the weighting of individual variables in polytomous logistic regression, the parameters associated with each variable reflect the increased (or decreased) *odds* of a particular outcome (i.e. lexeme) occurring (in contrast to all the rest, or some baseline category, or otherwise, depending on which practical heuristic has been selected), when the particular feature is present in the context (instead of being absent), with all the other explanatory variables being equal.

There are a number of heuristics which are all based on the splitting of the polytomous setting into a set of dichotomous cases, for which each the binary logistic regression model can then be applied and fitted either simultaneously or separately; the differences of the heuristics are in the strategies according to which the decomposition and its overall fitting is undertaken. The relevant heuristics are 1) the baseline-category multinomial model (e.g., Hosmer and Lemeshow, 2000: 260-287), 2) one-vs-rest classification (e.g., Rifkin and Klautau, 2004), 3) pairwise classification (e.g., Fürnkranz, 2002), 4) nested dichotomies (e.g., Fox, 1997: 472-475), and 5) ensembles of nested dichotomies (e.g. Frank and Kramer, 2004). A concise presentation of all these and a few more heuristics can be found in Frank and Kramer (2004). In general, it has been observed that the process of separately fitting the binarized models mostly does not have a substantial (detrimental or differentiating) effect on the overall results, in comparison to simultaneously fitting a base-line category multinomial model, the last one which is sometimes considered preferable as the most "elegant" solution (Hosmer and Lemeshow 2000: 277-278). In order to get both lexeme-specific parameters for the selected contextual features, without having to select one lexeme as a baseline category, and probability estimates for the occurrences of each lexeme, the one-vs-rest heuristic is the most appealing of the lot.

In the *one-vs-rest* heuristic, the regression coefficients (i.e. parameter values) of the individual binary models can be understood to highlight those feature variables which distinguish the individual outcome classes (i.e. lexemes) from all the rest, and the individual lexeme-specific models can be meaningfully studied together as a group. An individual odds (parameter value) which is greater than 1.0 for some feature variable and the singled-out lexeme can be interpreted as the increased chances of the occurrence of this lexeme when the feature in question is present in the context, whereas the odds less than 1.0 would denote the decreased chances of the occurrence of this lexeme, translating into corresponding increased odds in favor of any one of the three other lexemes occurring in such a context. However, one must note that the odds *do not* apply in cases when the particular feature is *not present* in the context. To its benefit, the one-vs-rest heuristic is also methodologically simple as both the parameters and the probability estimates are directly derived from binary logistic regression models of which it consists. Furthermore, Rifkin and Klautau (2004: 102) argue forcefully that, contrary to the common presumption one-vs-rest is not less accurate than other, typically more sophisticated heuristics. The statistical calculations were undertaken in public-domain *R* statistical programming environment (R Core Development Team, 2007), using both ready-made functions (`glm` for binary logistic regression) and scripts written by myself (for implementing the one-vs-rest as well as other heuristics and assessments of their performance). The lexeme-specific models can be formally stated according to the frame exemplified for *ajatella* in (1) below, where no interactions are assumed among the feature variables.

(1) ajatella <- AGENT.INDIVIDUAL + AGENT.GROUP + PATIENT.INDIVIDUAL + PATIENT.GROUP + PATIENT.NOTION + PATIENT.ATTRIBUTE + PATIENT.STATE + PATIENT.TIME + PATIENT.COMMUNICATION + PATIENT.ACTIVITY + PATIENT.EVENT + PATIENT.INFINITIVE + PATIENT.PARTICIPLE + PATIENT.INDIRECT_QUESTION + PATIENT.DIRECT_QUOTE + PATIENT.että + MANNER.GENERIC + MANNER.FRAME + MANNER.POSITIVE + MANNER.NEGATIVE + MANNER.AGREEMENT + MANNER.JOINT + QUANTITY.LITTLE + QUANTITY.MUCH + LOCATION.LOCATION + LOCATION.GROUP + LOCATION.EVENT + TIME.DEFINITE + TIME.INDEFINITE + DURATION.SEM_LONG + DURATION.OPEN + DURATION.SHORT + FREQUENCY.AGAIN + FREQUENCY.OFTEN + GOAL + CONDITION + REASON + CLAUSE-ADVERBIAL + VERB-CHAIN.PROPOSSIBILITY + VERB-CHAIN.IMPOSSIBILITY + VERB-CHAIN.PRONECESSITY + VERB-CHAIN.CONTRANECESSITY + VERB_CHAIN.TEMPORAL + VERB_CHAIN.VOLITION + VERB_CHAIN.RESULTATIVE + CO-ORDINATION. MENTAL + CO-ORDINATION.ACTION + NEGATION + INDICATIVE + CONDITIONAL + IMPERATIVE + PASSIVE + FIRST + SECOND + THIRD + PLURAL + COVERT + CLAUSE-EQUIVALENT

On the general level, this setup of multivariate statistical analysis including a wide range of different features is quite similar to that of Divjak and Gries (2006), though my foremost focus is rather on discovering features which characterize and distinguish the members of a synonym group from each other, and the relative weights of these features, than on the internal grouping of the synonym group that these features also reveal. Furthermore, Divjak and Gries used *Hierarchical Agglomerative Clustering* (HAC), which is especially adept in determining and visualizing the extent of semantic similarity between the individual lexemes in a synonym set. Such primarily visual methods do build upon and thus also contain precise numerical analysis, the results of which can be used to describe the associations of the lexemes and the features, as Divjak and Gries (2006) demonstrate. Nevertheless, such numeric data (e.g., *t-scores* and *z-scores* in the case of cluster analysis) lack the direct natural interpretation that logistic regression provides, in the form of odds for the explanatory variables and expected probabilities for the individual outcomes.

3. Results and discussion

The application of the one-vs-rest heuristic for polytomous logistic regression to the research corpora, with the selected linguistic variables as the model, reached a relative decrease in deviance, reflecting the fit of the model with the entire data, of $R_L^2 = 0.325$ (Hosmer and Lemeshow, 2000: 165-166) , which is not a bad fit at all. With respect to prediction efficiency, the overall recall rate was 65.2 percent, while the measures assessing the reduction of error (see Menard 1995) were $\lambda_{prediction}$=0.381 and $\tau_{classification}$=0.499, which are also relatively good results. When further validating the model by repeatedly (20 times) training the model with a random hold-out sample of two-thirds of the entire data and then evaluating the model against the remaining one-third of the data, this particular set of feature variables could correctly account for on the average 63.3 percent (s.d. 1.2%) of the lexical choices in the research corpora, while corresponding values for the model fit were $R_L^2 = 0.334$ (s.d. 0.006). With respect to measures assessing the reduction of error, the validation figures were $\lambda_{prediction}$=0.348 (s.d. 0.018) and $\tau_{classification}$=0.472 (s.d. 0.015).

The recall rate reached here was 6-7 percentage units higher than the 58-59 percent reported by Arppe (2006), which was a level that had been achieved on the basis of the WordNet classifications of nominal arguments alone. This shows that the *ad hoc* semantic classifications of the non-nominal arguments as a whole do hold substantial added explanatory power. Somewhat surprising, however, is that the additional classifications of modality and co-ordinated verbs, included for the

first time in this study, when compared to other earlier *ad hoc* classifications concerning only MANNER, DURATION, FREQUENCY and QUANTITY, do not seem to have caused a visible impact on the model fit and prediction efficiency. This may result from substantial number of cases representing true interchangability, as is befitting of synonyms. Or, this may be attributed to limitations in how much of the studied phenomenon can be explained using the selected types of variables representing three conventional levels of linguistic analysis, namely morphology, syntax and semantics. Furthermore, preliminary single-run trials with other techniques for turning polytomous classification problems into sets of dichotomous classifiers, e.g. *baseline-category multinomial* (recall=65.4% , R_L^2=0.328), *pairwise* classification (recall=65.3% , R_L^2=NA), or *ensembles of nested dichotomies* (recall=65.3% , R_L^2=0.328), do not appear to yield essentially better performance figures, either.

The feature-specific odds for each of the studied THINK lexemes estimated according to the one-vs-rest heuristic are presented in their entirety in Appendix 1. These results can now be scrutinized from two perspectives, either lexeme-wise or feature-wise. On the one hand, we can assess which individual features significantly increase or decrease the odds, in favor of or against the occurrence of any lexeme in the THINK synonym set (Table 4), and what are the strengths of these odds. At the same time, we can also see which features are neutral in this respect, i.e. have odds that do not significantly diverge from 1.0. On the other hand, we may be interested in which lexemes have the strongest odds in favor of or against occurring in conjunction with each studied feature, or whether any features have neutral odds for all the studied THINK lexemes (see Table 5 with respect to the semantic types of AGENTs as well as the analytic morphological features pertaining to the verb-chain).

So, we may in Table 4 see for instance that a GENERIC argument of MANNER or one indicating AGREEMENT (or disagreement), a human GROUP as a PATIENT as well as an INFINITIVE verb form, and a RESULTATIVE verb chain (e.g., *tulla ajatelleeksi* 'come to think of') exhibit the greatest odds in favor of the occurrence of *ajatella*. In contrast, an act or form of COMMUNICATION, or an indirect question or a direct quote as a PATIENT, a SHORT DURATION, or an EVENT as a LOCATION show the strongest odds against the occurrence of this same verb. On the other hand, with respect to this latter set of features, a SHORT DURATION and an indirect question as a PATIENT increase the odds for *miettiä* to occur, while an EVENT as a LOCATION or a direct quote as a PATIENT raise the odds for *pohtia*. Similar assessments can be done for each of the THINK lexemes included in the analysis.

Lexeme/Features	Strongest odds in favor of the lexeme	Strongest odds against the lexeme
ajatella	MANNER+GENERIC (22.9) MANNER+AGREEMENT (14.4) PATIENT+GROUP (8.2) PATIENT+INFINITIVE (6.4) VERB-CHAIN+RESULTATIVE (6.4)	PATIENT+COMMUNICATION (0.1) DURATION+SHORT (0.1) PATIENT+INDIRECT_QUESTION (0.1) LOCATION+EVENT (0.0) PATIENT+DIRECT_QUOTE (0.0)
miettiä	VERB-CHAIN+CONTRANECESSITY (8.1) DURATION+SHORT (8.0) PATIENT+INDIRECT_QUESTION (4.7) FREQUENCY+OFTEN (4.5) DURATION+LONG (4.4)	MANNER+FRAME (0.3) MANNER+GENERIC (0.2) MANNER+AGREEMENT (0.1) PATIENT+PARTICIPLE (0.0) PATIENT+INFINITIVE (0.0)
pohtia	LOCATION+EVENT (13.2) PATIENT+DIRECT_QUOTE (7.7) PATIENT+ATTRIBUTE (5.9) PATIENT+NOTION (4.6) AGENT+GROUP (3.7)	PATIENT+INFINITIVE (0.2) MANNER+NEGATIVE (0.2) PATIENT+GROUP (0.2) IMPERATIVE MOOD (0.1) MANNER+GENERIC (0.0)
harkita	PATIENT+ACTIVITY (8.8) CONDITION (3.1) PATIENT+STATE (2.3) FREQUENCY+AGAIN (2.2) FIRST PERSON (2.1)	PATIENT+DIRECT_QUOTE (0.0) PATIENT+GROUP (0.0) MANNER+GENERIC (0.0) VERB-CHAIN+RESULTATIVE (0.0) QUANTITY+LITTLE (0.0)

Table 4. Features per each studied lexeme with the strongest odds both in favor and against the occurrence of the lexeme in question; five features of both types per lexeme; actual odds in parentheses.

Moving on to the feature-wise observations in Table 5, we can for instance see that NEGATION increases the odds in favor of both *ajatella* and *harkita* occurring, whereas this feature decreases the odds for *miettiä* and *pohtia*; consequently, none of the four THINK lexemes are neutral with respect to NEGATION. In contrast, while PASSIVE voice in the context increases the odds of occurrence for *pohtia* and diminishes that of *ajatella*, it has no significant bearing on the occurrence of both *harkita* and *miettiä*. In the extreme, a human INDIVIDUAL as an AGENT, or the lack of an overt AGENT (denoted by the feature COVERT) is neutral with respect to all four of the studied THINK lexemes.

We can now also compare these results with an earlier multimethodological study (Arppe and Järvikivi, forthcoming) which combined both corpus and experimental data concerning the AGENT types and the associated person/number features, and which focused only on the pair *miettiä* and *pohtia*. Within the more complex syntactic-semantic network and the larger group of THINK lexemes considered in this study, it is interesting to note that the contrasts observed between *miettiä* and *pohtia* shift somewhat, but are nonetheless essentially upheld. As concluded in the combined results in the earlier study, a human GROUP as an AGENT has strong and significant odds in favor of *pohtia* in this study, but now *miettiä* is neutral with respect to this feature, instead of exhibiting a negative preference which was especially evident in the acceptability rating experiments of the former study. With respect to human INDIVIDUALS as AGENTS, the results in this study conform to the overall conclusion in the prior study that there is no significant difference between the two lexemes. Furthermore, whereas the corpus-based results in the prior study indicated a strong positive association between FIRST PERSON SINGULAR and *miettiä*, and a negative one with *pohtia*, in this study the result stays the same for *pohtia*, while the effect with respect to *miettiä* has turned neutral. It would be interesting to find out whether this dispreference of *pohtia* with respect to FIRST PERSON would diminish also in an acceptability rating experiment covering all the four THINK lexemes, similar to what was observed in such an experiment in the prior study.

Contextual feature	Lexemes with strong odds in favor	Lexemes with neutral odds	Lexemes with strong odds against
AGENT+INDIVIDUAL	-	pohtia (1.4), ajatella (1.0), miettiä (1.0), harkita (0.7)	-
AGENT+GROUP	pohtia (3.7)	miettiä (0.6), harkita (1.1)	ajatella (0.2)
NEGATION	ajatella (2.2), harkita (1.9)	-	miettiä (0.5), pohtia (0.4)
INDICATIVE MOOD	ajatella (5.8)	miettiä (1.2), harkita (0.6)	pohtia (0.3)
CONDITIONAL MOOD	ajatella (3.6),	harkita (1.7), miettiä (1.0)	pohtia (0.3)
IMPERATIVE MOOD	ajatella (5.8), miettiä (2.5)	harkita (0.5)	pohtia (0.1)
PASSIVE VOICE	pohtia (3.2)	harkita (1.2), miettiä (0.7)	ajatella (0.5)
FIRST PERSON	-	harkita (2.1), miettiä (1.6), ajatella (0.6),	pohtia (0.4)
SECOND PERSON	-	miettiä (1.7), pohtia (1.1), harkita (1.0)	ajatella (0.4)
THIRD PERSON	-	harkita (1.8), pohtia (1.6), miettiä (1.1)	ajatella (0.4)
PLURAL NUMBER	pohtia (1.6)	ajatella (1.2), harkita (1.2)	miettiä (0.6)
COVERT (AGENT)	-	ajatella (1.1), harkita (1.1), miettiä (0.9), pohtia (0.8)	-
CLAUSE-EQUIVALENT FORM	ajatella (2.7), harkita (1.9)	miettiä (0.9), pohtia (0.4)	-

Table 5. The sorting of the studied THINK lexemes per each semantic type of AGENT as well as each analytic morphological feature into ones with strong odds in favor of, neutral, and strong odds against the occurrence of each lexeme.

Finally, we can use the polytomous logistic regression model to assign expected probabilities for the same sentences in the corpus data that they were trained with. These results can in turn be used to automatically rank candidates for example sentences, which will comprehensively incorporate the contextual preferences that have been observed. Appendix 2 contains the five highest ranked sentences in the research corpora for each of the four THINK lexemes, as well as the estimated probabilities, of which Table 6 presents an excerpt with the two very top-most ranked sentences for *harkita*. We can firstly see that the combination of a REASON argument, a FIRST person human INDIVIDUAL as an AGENT, an ACTIVITY as a PATIENT and a subordinate clause denoting a CONDITION, plus a CONDITIONAL feature as well as positive NECESSITY present in the verb chain give a probability of 0.92 for *harkita* to occur (i.e. over nine times out of ten), as is indeed the case for the particular sentence in the research corpora. When the person feature in the aforementioned context type is switched to THIRD person (which is furthermore COVERT), and in addition a POSITIVE (THOROUGH) argument of MANNER is also present, the probability for the occurrence of *harkita* is only slightly less at 0.88. Interestingly, in the research corpora it is rather *pohtia* which has been used in this particular instance, though its expected probability in such a context type is only 0.10 (i.e. once out of ten times). This demonstrates on the one hand that occurrences of the studied lexemes in particular contexts are not categorically determined but rather probabilistic, and on the other hand that the selected structural feature variables can account for the occurrences of the studied THINK lexemes only to a certain extent.

P(*harkita*)	Sentence
0.9193	*"... Saatananpalvonta tai jonkinlainen pelottava kulttiuskonto*$_{REASON}$ *voisi*$_{CONDITIONAL}$ *saada*$_{VERB\_CHAIN+PRONECESSITY}$ *minut*$_{AGENT+INDIVIDUAL,\ FIRST}$ ***harkitsemaan*** *eroa*$_{PATIENT+ACTIVITY}$*, jos*$_{CONDITION}$ *kumppanini tuntuisi seonneen totaalisesti, tai jos pitäisi huolehtia lasten turvallisuudesta."* [ihmissuhteet_9584] '"Satan worship or some other type of frightening cult religion could get me to consider separating, if my partner would seem to have freaked out completely, or if one should worry about the safety of the children."'
0.8823 (0.0963)	"Hanketta$_{PATIENT+ACTIVITY}$ tulisi$_{VERB_CHAIN+PRONECESSITY,CONDITIONAL+THIRD+COVERT}$ kannanoton mukaan$_{META}$ **pohtia** rauhallisesti$_{MANNER+POSITIVE}$, koska$_{REASON}$ Töölönlahdella ei ole mahdollisuuksia nopeaan toteutukseen." [hs95_9215] '"The project proposal should according to the comment be considered calmly, as in Töölö bay there are no possibilities for quick implementation."'

Table 6. Two example sentences in the research corpora (and their approximate English translations) with the highest estimated probabilities of occurrence for *harkita*; probabilities estimated based on fitting a polytomous logistic regression model with the selected contextual features variables, using the one-vs-rest heuristic on the research corpora. Pertinent feature variables as subscripts next to the appropriate word (or head in the case of a phrase/clause)

4. Conclusions and further work

In conclusion, these results demonstrate the variety and complexity of different contextual feature categories necessary to explain the use of the studied synonym group. Furthermore, they present an adaptation of a multivariate statistical method, namely the one-vs-rest technique for polytomous logistic regression, which is applicable to multiple-category problems that synonym groups often are. This allows one to tease out complex associations from corpora which can then be used to enrich our lexicographical knowledge. Provided that one firstly had at one's disposal a sufficiently broad general semantic ontology, of the WordNet type covering the common, core semantic groupings of a language, and secondly a relatively richly annotated corpus, one could envision generating in an assembly-line fashion formalized feature descriptions as well as representative example sentences concerning the usage of one synonym group after the other, which a lexicographer could then refine further.

Further study should be undertaken to assess how robust the observed effects are, for instance to what extent they are subject to individual speaker/writer preferences, using techniques such as bootstrap resampling by treating writers/speakers as clusters, along the lines that Bresnan *et al.* (2007) have demonstrated. One might also be interested in what role the medium or domain of a text possibly play in the use of the studied synonyms. Here, a possible methodological solution is to treat such subgroupings of text as an additional explanatory variable incorporated in the model, also in accordance to Bresnan *et al.* (2007). Finally, such comprehensive corpus-based analysis makes it possible to improve the precision of our hypotheses concerning linguistic usage, which can then be evaluated against other types of linguistic evidence and methods, such as can be gathered with experimentation (exemplified in e.g. Gries, 2002, 2003b, Bresnan, 2006, Arppe and Järvikivi, forthcoming). Such multimethodological comparative work will surely increase our overall understanding of language as the multifaceted phenomenon it is.

Acknowledgements

This research has been undertaken within LANGNET, the Finnish Graduate School in Language Studies, for which financial support in am grateful. Furthermore, I want to thank Professor Martti Vainio for introducing me to logistic regression as a research method, as well as guidance and support in its practical application to this polytomous linguistic setting.

Corpora

Finnish Text Collection [FTC] (2001) ~180 million words of Finnish, consisting of 97 subcollections of Finnish newspaper, magazine and literature texts from the 1990s. Compiled by the Department of General Linguistics at the University of Helsinki, the Foreign Languages Department (General Linguistics) at the University of Joensuu, the Research Institute for the Languages in Finland, and CSC – Center for Scientific Computing, Finland, in 1996-1998 and 1999-2001. Available on-line at URL: http://www.csc.fi/kielipankki/

Helsingin Sanomat (1995) ~4 million words of Finnish newspaper articles published in Helsingin Sanomat during January–February 1995. Compiled by the Research Institute for the Languages of Finland [KOTUS] and CSC – Center for Scientific Computing, Finland. Available on-line at URL: http://www.csc.fi/kielipankki/

SFNET (2002–2003) ~100 million words of Finnish internet newsgroup discussion posted during October 2002–April 2003. Compiled by Tuuli Tuominen and Pasi Kalliokoski, Computing Centre, University of Helsinki, and Antti Arppe, Department of General Linguistics, University of Helsinki, and CSC – Center for Scientific Computing, Finland. Available on-line at URL: http://www.csc.fi/kielipankki/

References

Arppe, A. (2002) The usage patterns and selectional preferences of synonyms in a morphologically rich language, in Morin, A. and P. Sébillot (Eds.). *JADT-2002. 6th International Conference on Textual Data Statistical Analysis*, 13-15.3.2002, Vol. 1, pp. 21-32. Rennes: INRIA.

Arppe, A. (2006) Complex phenomena deserve complex explanations. *Quantitative Investigations in Theoretical Linguistics* (QITL2) Conference, Osnabrück, Germany, 1-2.6.2006, 8-11. Available on-line at URL: http://www.cogsci.uni-osnabrueck.de/~qitl/

Arppe, A. (forthcoming) Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy. [PhD dissertation]

Arppe, A. and J. Järvikivi (forthcoming). 'Every method counts - Combining corpus-based and experimental evidence in the study of synonymy'. *Corpus Linguistics and Linguistic Theory*.

Atkins, B. T. S. and B. Levin (1995) 'Building on a Corpus: A linguistic and lexicographical look at some near-synonyms'. *International Journal of Lexicography*, 8:2, 85–114.

Biber, D., S. Conrad, and R. Reppen (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press.

Bresnan, J. (2006) Is knowledge of syntax probabilistic? Experiments with the English dative alternation. Pre-proceedings of the International Conference on Linguistic Evidence. Empirical, Theoretical and Computational Perspectives, 2–4.2.2006, SFB441 "Linguistic Data Structures", University of Tübingen, Germany, 3–10.

Bresnan, J., A. Cueni, T. Nikitina, and R. H. Baayen (2007) Predicting the Dative Alternation, in Cognitive Foundations of Interpretation. Boume, G., I. Kraemer, and J. Zwarts (Eds), pp. 69-94. Amsterdam: Royal Netherlands Academy of Science.

Church, K., W. Gale, P. Hanks, and D. Hindle (1991). Using Statistics in Lexical Analysis, in Zernik, U. (Ed.) Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, pp. 115-164. Hillsdale: Lawrence Erlbaum Associates.

Connexor (2007). Language Model Tag Descriptions [for Finnish]. URL: http://www.connexor.com/demo/doc/fifdg3-tags.html (accessed: 29.5.2007)

Divjak, D. (2006) Ways on Intending. Delineating and Structuring Near-Synonyms, in Gries, S. Th. and A. Stefanowitsch (Eds). Corpora in cognitive linguistics. Vol. 2: The syntax-lexis interface, pp. 19-56. Berlin: Mouton De Gruyter.

Divjak, D. and S. Th. Gries (2006). 'Ways of trying in Russian: Clustering and comparing behavioral profiles'. *Corpus Linguistics and Linguistic Theory*, 23-60.

Edmonds, P. and G. Hirst (2002) 'Near-Synonymy and Lexical Choice'. *Computational Linguistics* 28(2), 105-145.

Featherston, S. (2005) The Decathlon Model, in Kepser, S. and M. Reis. Linguistic Evidence. Empirical, Theoretical and Computational Perspectives. Studies in Generative Grammar 85, pp. 187–208. Berlin/New York: Mouton de Gruyter.

Flint, A. (1980). Semantic Structure in the Finnish Lexicon: Verbs of Possibility and Sufficiency. Helsinki: Suomalaisen Kirjallisuuden Seura (SKST 360).

Fox, J. (1997) Applied regression analysis, linear models, and related methods. Thousand Oaks: Sage.

Frank, E. and S. Cramer (2004). Ensembles of Nested Dichotomies for Multi-class Problems, in Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.

Fürnkranz, J. (2002) 'Round Robin Classification'. *Journal of Machine Learning Research*, 2/2002, 721-747.

Gries, S. Th. (2002) Evidence in linguistics: Three approaches to genitives in English, in Brend, R. M., W. J. Sullivan, and A. R. Lommel (Eds), LACUS Forum XXVIII: What Constitutes Evidence in Linguistics?, pp. 17–31. Fullerton: LACUS.

Gries, S. Th. (2003a). Multifactorial analysis in corpus linguistics: a study of particle placement. London: Continuum.

Gries, S. Th. (2003b). 'Towards a corpus-based identification of prototypical instances of constructions'. *Annual Review of Cognitive Linguistics*, 1:1-27.

Haarala, R. and M. Lehtinen (Eds) (1997) *CD-Perussanakirja*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 94. Helsinki: Edita.

Hanks, P. (1996) 'Contextual Dependency and Lexical Sets'. *International Journal of Corpus Linguistics*, Vol. 1, No. 1, 75-98.

Harrell, F. E. (2001) Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer-Verlag.

Hosmer, D. W., Jr., and S. Lemeshow (2000). *Applied Regression Analysis* (2nd edition). New York: Wiley.

Jantunen, Jarmo H. (2001). 'Tärkeä seikka ja keskeinen kysymys'. Mitä lähisynonymia paljastaa lähisynonyymeista? *Virittäjä* 105, 170-192.

Jäppinen, Harri (Ed.) (1989). *Synonyymisanakirja*. [Nykysuomen sanakirja VII]. Porvoo: Werner Söderström.

Järvinen, T. and P. Tapanainen (1998) Towards and implementable dependency grammar, in Kahane, S. and A. Polguère (Eds). Proceedings of the Workshop on Processing of Dependency-based Grammars, COLING-ACL'98, Montreal, Canada.

Menard, S. (1995). Applied Logistic Regression Analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences 07-106. Thousand Oaks: Sage Publications.

Miller, G. A. (1990) 'Nouns in WordNet: a lexical inheritance system'. (revised August 1993). *International Journal of Lexicography*, 3 (4), 245–264. Available on-line at: URL: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Pajunen, A. (2001). Argumenttirakenne: Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä. Suomi 187. Helsinki: Suomalaisen Kirjallisuuden Seura.

R Development Core Team (2007) R: A language and environment for statistical computing. URL: http://www.R-project.org

Rifkin, R. and A. Krakatau (2004) 'In Defense of One-Vs-All Classification'. *Journal of Machine Learning Research*, 101-141.

Tapanainen, P. and T. Järvinen (1997) A non-projective dependency parser. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97), pp. 64–71. Washington, D.C.: Association for Computational Linguistics.

Appendix 1. Odds for each selected feature with respect to the occurrence of the studied THINK lexemes; odds significantly greater than 1.0 indicate an increased odds in favor of a lexeme, while odds significantly less than 1.0 indicate a decreased odds against a lexeme; significant odds in **boldface** (for which P(|Z-score|)<0.05).

Contextual feature/Lexeme	**ajatella**	**miettiä**	**pohtia**	**harkita**
(Intercept)	0.9	**0.1**	**0.3**	**0.1**
SYNTAX+SEMANTICS				
AGENT+INDIVIDUAL	1.0	1.0	1.4	0.7
AGENT+GROUP	**0.2**	0.6	**3.7**	1.1
PATIENT+INDIVIDUAL	**2.0**	**0.5**	**0.4**	1.3
PATIENT+GROUP	**8.2**	0.4	0.2	0.0
PATIENT+NOTION	**0.2**	**1.7**	**4.6**	1.0
PATIENT+ATTRIBUTE	**0.2**	1.3	**5.9**	1.0
PATIENT+STATE	0.5	1.0	1.5	2.3
PATIENT+TIME	1.1	1.1	1.2	0.6
PATIENT+COMMUNICATION	**0.1**	**2.8**	**3.3**	1.9
PATIENT+ACTIVITY	**0.1**	0.8	**1.6**	8.8
PATIENT+EVENT	1.3	1.0	1.1	0.3
PATIENT+INFINITIVE	**6.4**	0.0	0.2	1.5
PATIENT+PARTICIPLE	**6.0**	0.0	**0.2**	1.1
PATIENT+INDIRECT_QUESTION	**0.1**	**4.7**	**2.9**	0.8
PATIENT+DIRECT_QUOTE	**0.0**	**3.0**	**7.7**	0.0
PATIENT+että 'that'	**2.7**	**0.5**	**0.5**	**0.2**
MANNER+GENERIC	**22.9**	**0.2**	0.0	0.0
MANNER+FRAME	**2.5**	**0.3**	1.3	**0.3**
MANNER+POSITIVE	0.8	1.0	0.8	**1.8**
MANNER+NEGATIVE	**4.1**	0.5	0.2	0.6
MANNER+AGREEMENT	**14.4**	**0.1**	0.3	0.0
MANNER+JOINT	**0.4**	**2.0**	0.8	1.6
QUANTITY+LITTLE	0.6	**4.4**	0.5	0.0
QUANTITY+MUCH	0.9	1.6	1.0	0.7
LOCATION+LOCATION	**0.4**	0.6	**3.1**	0.5
LOCATION+GROUP	**0.4**	**2.6**	1.0	0.7
LOCATION+EVENT	**0.0**	0.4	**13.2**	0.3
TIME+DEFINITE	**0.4**	1.0	**2.2**	0.7
TIME+INDEFINITE	**0.6**	**1.5**	0.9	1.3
DURATION+SEM_LONG	**0.1**	**4.4**	0.8	1.5
DURATION+OPEN	**0.2**	1.5	1.7	1.8
DURATION+SHORT	**0.1**	**8.0**	0.9	0.3
FREQUENCY+AGAIN	0.5	1.1	0.7	**2.2**
FREQUENCY+OFTEN	**0.3**	**4.5**	0.5	0.4
SYNTAX				
GOAL	**4.1**	0.6	0.5	**0.2**
CONDITION	**0.5**	1.3	0.5	**3.1**
REASON	0.6	1.1	0.9	1.6
CLAUSE-ADVERBIAL (META-COMMENT)	0.8	1.0	0.8	**1.6**
VERB-CHAIN+MODALITY and CO-ORDINATION				
VERB-CHAIN+PROPOSSIBILITY	1.2	1.2	**0.6**	1.6
VERB-CHAIN+IMPOSSIBILITY	1.6	1.0	1.5	**0.2**
VERB-CHAIN+PRONECESSITY	**0.4**	**1.8**	0.8	**1.6**
VERB-CHAIN+CONTRANECESSITY	**0.3**	**8.1**	0.4	0.2
VERB_CHAIN+TEMPORAL	**0.3**	**1.8**	**2.3**	**0.1**
VERB_CHAIN+VOLITION	0.6	1.6	1.2	0.6

VERB_CHAIN+RESULTATIVE	**6.4**	0.5	0.4	0.0
CO-ORDINATION+ MENTAL	**0.4**	**2.4**	0.9	0.8
CO-ORDINATION+ACTION	1.1	1.8	0.5	1.0
VERB-CHAIN MORPHOLOGY				
NEGATION	**2.2**	**0.5**	**0.4**	**1.9**
INDICATIVE MOOD	**5.8**	1.2	**0.3**	0.6
CONDITIONAL MOOD	**3.6**	1.0	**0.3**	1.7
IMPERATIVE MOOD	**5.1**	**2.5**	**0.1**	0.5
PASSIVE VOICE	**0.5**	0.7	**3.2**	1.2
FIRST PERSON	0.6	1.6	**0.4**	2.1
SECOND PERSON	**0.4**	1.7	1.1	1.0
THIRD PERSON	**0.4**	1.1	1.6	1.8
PLURAL NUMBER	1.2	**0.6**	**1.6**	1.2
COVERT (AGENT)	1.1	1.1	0.9	0.8
CLAUSE-EQUIVALENT FORM	**2.7**	0.9	0.4	**1.9**

Appendix 2. Example sentences in the research corpora with the highest estimated probabilities of occurrence for each of the four studied THINK lexemes (five sentences per each lexeme); probabilities estimated based on fitting a polytomous logistic regression model with the selected contextual features variables, using the one-vs-rest heuristic on the research corpora. Pertinent feature variables as subscripts next to the appropriate word (or head in the case of a phrase/clause)

P(*ajatella*)	**Sentence**
0.9980	"Huom. itse$_{\text{AGENT+INDIVIDUAL}}$ en$_{\text{NEGATION+FIRST}}$ **ajattele**$_{\text{INDICATIVE+PRESENT}}$ kirjastoista$_{\text{SOURCE}}$ näin$_{\text{MANNER+GENERIC}}$." [politiikka_12561]
0.9972	"Voihan$_{\text{VERB_CHAIN+PROPOSSIBILITY, INDICATIVE+PRESENT+THIRD+COVERT, AGENT+INDIVIDUAL}}$ ihmisten ruumiillisista ominaisuuksista$_{\text{SOURCE}}$ **ajatella** ihan samalla tavalla$_{\text{MANNER+AGREEMENT}}$, että$_{\text{PATIENT+että}}$ ei ole 'oikeaa' ja 'väärää', 'tervettä' ja 'sairasta'". [ihmissuhteet_7237]
0.9968	"... Miten$_{\text{MANNER+GENERIC}}$ **ajattelit**$_{\text{INDICATIVE+PRESENT+SECOND+COVERT, AGENT+INDIVIDUAL}}$ erota$_{\text{PATIENT+INFINITIVE}}$ mitenkään jostain SAKn umpimielisistä luokka-ajattelun kannattajasta?" [politiikka_9967]
0.9967	"Tähän pitää vastata sen mukaan, miten$_{\text{MANNER+GENERIC}}$ **ajattelee**$_{\text{INDICATIVE+PRESENT+THIRD+COVERT, AGENT+INDIVIDUAL}}$ niiden vastaavan$_{\text{PATIENT+PARTICIPLE}}$ jotka ruuan arvonlisäveron laskua haluaisivat ei eivät haluaisi, vaikka valinta tarkastiottaen poissulkee muut laskuvaihtoehdot, maltillisemmat ja radikaalimmat. [politiikka_17004]
0.9964	"Kyynikko$_{\text{AGENT+INDIVIDUAL}}$ voisi$_{\text{CONDITIONAL+THIRD}}$ tästä$_{\text{SOURCE}}$ tosin$_{\text{META}}$ **ajatella** niin$_{\text{MANNER+GENERIC}}$, että$_{\text{PAT+että}}$ joku kenties jollekulle tulisi mieleen siirtää kirjoittelunsa painopistettä nyysseistä bloggiin välttääkseen omiin ah niin rakkaisiin, mutta valitettavan haavoittuviin teorioihinsa kohdistetun ilkeämielisen kritiikin ..." [ihmissuhteet_6882]
P(*miettiä*)	**Sentence**
0.9641	"Ai että vastuu jäi nyt sitten minulle", sanoo$_{\text{CO-ORDINATION+MENTAL}}$ Sievinen$_{\text{AGENT+INDIVIDUAL}}$ ja **miettii**$_{\text{INDICATIVE+PRESENT+THIRD}}$ pitkään$_{\text{DURATION+LONG}}$ mitä$_{\text{PATIENT+INDIRECT_QUESTION}}$ vastaisi. [hs95_11987]
0.9437	"**Mietipä**"$_{\text{IMPERATIVE+SECOND}}$ nyt$_{\text{TIME+DEFINITE}}$ hiukan$_{\text{QUANTITY+LITTLE}}$ itsekin$_{\text{MANNER+JOINT}}$ juttujasi$_{\text{PATIENT+COMMUNICATION}}$. [ihmissuhteet_2952]
0.9422	"Simonsuuri toki muistuttaa tuntevansa mytologioiden antropologisia analyysejä,

	hän$_{\text{AGENT+INDIVIDUAL}}$ varoittaa$_{\text{CO-ORDINATION+MENTAL}}$ antiikin draaman naiskuvien anakronistisista tulkinnoista tai **pohtii** hetken$_{\text{DURATION+SHORT}}$, miltä$_{\text{PATIENT+INDIRECT_QUESTION}}$ tuntuisi ajatella alitajuisia prosesseja ja myyttikertomuksia "historiallisen todellisuuden ja kulttuuristen prosessien" tuloksina. [hs95_1681]
0.9395	"Yksityistetyt laitokset ovat edelleenkin turvallisia alansa monopoleja , joissa ei johtajan$_{\text{AGENT+INDIVIDUAL}}$ juuri$_{\text{QUANTITY+LITTLE}}$ tarvitse$_{\text{VERB_CHAIN+CONTRANECESSITY,INDICATIVE+PRESENT+THIRD}}$ **miettiä** kilpailuriskejä$_{\text{PATIENT+ATTRIBUTE}}$." [hs95_5913]
0.9329	"**Mieti**$_{\text{IMPERATIVE+SECOND+COVERT, AGENT+INDIVIDUAL}}$ nyt$_{\text{TIME+DEFINITE}}$ vähän$_{\text{QUANTITY+LITTLE}}$ miten$_{\text{PATIENT+INDIRECT_QUESTION}}$ ennakkoluuloinen olet!" [ihmissuhteet_5598]
P(*pohtia*)	**Sentence**
0.9716	"Asiaa$_{\text{PATIENT+NOTION}}$ **pohditaan**$_{\text{PASSIVE+INDICATIVE+PRESENT}}$ klo$_{\text{TIME+DEFINITE}}$ 13 Helsingin työttömien viikkotapaamisessa$_{\text{LOCATION+EVENT}}$ Tennispalatsissa, Freda 65 B, 2. krs." [hs95_1085]
0.9588	"Viimeksi suomalaiset teatterintekijät$_{\text{AGENT+INDIVIDUAL}}$ **pohtivat**$_{\text{INDICATIVE+PAST+THIRD+PLURAL}}$ noin runsas puoli vuotta sitten$_{\text{TIME+DEFINITE}}$ Tampereen teatterikesässä$_{\text{LOCATION+EVENT}}$, miksi$_{\text{PATIENT+INDIRECT_QUESTION}}$ varsinkin monet naisohjaajat haluavat tarkastella elämän ikuisia peruskysymyksiä juuri myyttien näkökulmasta." [hs95_10041]
0.9587	"Aiemmin punavihreää yhdistelmää ajanut puheenjohtaja Claes Andersson$_{\text{AGENT+INDIVIDUAL}}$ **pohti**$_{\text{INDICATIVE+PAST+THIRD}}$ hallituspohjaa$_{\text{PATIENT+NOTION}}$ puolueensa vaalikampanjan avajaisissa$_{\text{LOCATION+EVENT}}$ Tampereella keskiviikkona$_{\text{TIME+DEFINITE}}$." [hs95_8591]
0.9576	"Kaupunkisuunnittelulautakunta$_{\text{AGENT+GROUP}}$ **pohti**$_{\text{INDIVATIVE+PAST+THIRD}}$ kokouksessaan$_{\text{LOCATION+EVENT}}$ myös$_{\text{META}}$ vaihtoehtoja$_{\text{PATIENT+NOTION}}$ siirtää raitiolinja 1:n päätepysäkki Havis Amandan patsaalta lähemmäksi Olympiaterminaalia . [hs95_1358]
0.9558	"Järjestyykö$_{\text{PATIENT+INDIRECT_QUESTION}}$ päivähoito Vantaalla", **pohditaan**$_{\text{PASSIVE+INDICATIVE+PRESENT}}$ paneelikeskustelussa$_{\text{LOCATION+EVENT}}$ klo $_{\text{TIME+DEFINITE}}$ 18.30 Peltolan koululla Tikkurilassa , Lummetie 27. ... [hs95_9277]
P(*harkita*)	**Sentence**
0.9193	"... Saatananpalvonta tai jonkinlainen pelottava kulttiuskonto$_{\text{REASON}}$ voisi$_{\text{CONDITIONAL}}$ saada$_{\text{VERB_CHAIN+PRONECESSITY}}$ minut$_{\text{AGENT+INDIVIDUAL, FIRST}}$ **harkitsemaan** eroa$_{\text{PATIENT+ACTIVITY}}$, jos$_{\text{CONDITION}}$ kumppanini tuntuisi seonneen totaalisesti, tai jos pitäisi huolehtia lasten turvallisuudesta." [ihmissuhteet_9584]
0.8823 (0.0963)	"Hanketta$_{\text{PATIENT+ACTIVITY}}$ tulisi$_{\text{VERB_CHAIN+PRONECESSITY,CONDITIONAL+THIRD+COVERT}}$ kannanoton mukaan$_{\text{META}}$ **pohtia** rauhallisesti$_{\text{MANNER+POSITIVE}}$, koska$_{\text{REASON}}$ Töölönlahdella ei ole mahdollisuuksia nopeaan toteutukseen." [hs95_9215]
0.8769	"huomauttaisin vielä, että itse$_{\text{MANNER+JOINT}}$ en$_{\text{NEGATION+FIRST+COVERT}}$ ainakaan koskaan$_{\text{TIME+INDEFINITE}}$ edes$_{\text{META}}$ **HARKITSISI**$_{\text{CONDITIONAL}}$ avioitumista$_{\text{PATIENT+ACTIVITY}}$ ihmisen kanssa, jolle seksi mekaanisena suorituksena on tärkeämpi kuin minä ihmisenä." [ihmissuhteet_3117]
0.8750	"Olisin$_{\text{CONDITION+FIRST}}$ itse$_{\text{AGENT+INDIVIDUAL}}$ - kuten olen jo toisessa yhteydessä todennut - valmis$_{\text{VERB_CHAIN+PROPOSSIBILITY}}$ **harkitsemaan** tuenilmaisua$_{\text{PATIENT+ACTIVITY}}$

	hänelle, jos$_{\text{CONDITION}}$ hänellä olisi vedenpitävä Marshallin suunnitelma Irakin jälleenrakentamiseksi ja demokratisoimiseksi mahdollisen sodan jälkeen.” [politiikka_12203]
0.8700	“Toivottavaa on, että Helsingin kaupunki$_{\text{AGENT+GROUP}}$ **harkitsisi**$_{\text{CONDITIONAL+THIRD}}$ vielä$_{\text{DURATION+OPEN}}$ vakavasti$_{\text{MANNER+POSITIVE}}$ Pasilanväylä-hanketta$_{\text{PATIENT+ACTIVITY}}$ ja ottaisi$_{\text{CO-ORDINATION+ACTION}}$ päätöksenteossaan tavoitteeksi.” [hs95_2218]