Population Testing: Extracting Semantic Information On Near-Synonymy From Native Speakers

Ulla Vanhatalo
Department of Finno-Ugrian Studies,
University of Helsinki
Department of Scandinavian Studies,
University of Washington
ulla.vanhatalo@helsinki.fi

Hilary Chan
Department of Near Eastern
Languages & Civilizations,
University of Washington
hilaryc@u.washington.edu
someguy@hilarychan.com

Abstract

Measuring differences between near-synonyms constitutes a major challenge in the development of electronic dictionaries and natural language processing systems. This paper presents a pilot study on how *Population Test Method* (PTM) may be used as an effective, empirical tool to define near-synonyms in a quantifiable manner. Use of PTM presumes that all knowledge about lexical meaning in a language resides collectively in the mind(s) of its native speakers, and that this *intersubjective* understanding may be extracted via targeted surveys that encourage creative, thinking responses. In this paper we show (1) examples of such tests performed on a group of high school students in Finland, (2) resulting data from the tests that is surprisingly quantifiable, and (3) a web-based visualization program we are developing to analyze and present the collected data.

1 Introduction

The problem of near-synonym discrimination presents a formidable challenge to computer-based natural language processing systems (Edmonds 1999; Edmonds and Hirst, 2002), as well as to humans who are attempting to acquire near-native competency in a foreign language. In both cases, a comprehensive lexical database specifically designed for near-synonymy in the target language is a pre-requisite for the further development of practical applications in their respective domains.

Some promising approaches have appeared in recent literature. These include corpus based procedures (Inkpen and Hirst 2001, 2002), and applied componential analysis, in particular continuing work on cross-lingual semantic primitives by Wierzbicka and her colleagues (Wierzbicka 1996, 1999).

Corpus-based approaches are, however, constrained by the kind and scope of pre-existing corpora and tools that are currently available; while componential analysis necessarily depends heavily on the subjective judgment of its investigators. Under such conditions, it may prove difficult to achieve complete and evenly distributed lexical coverage that truly reflects the diversity of the language community.

In this paper we propose another approach that we hope would complement these existing methods. In this approach, we go directly and repeatedly, in an iterative process, to the native speakers of the speech community to acquire and to verify the semantic information thus collected.

We also briefly describe a visualization tool (a Java applet) that we are currently developing to aid us in analyzing the collected data, and in further refining the semantic model.

2 Extracting Lexical Semantic Data with the Population Test Method (PTM)

2.1 General Background

Population Test Method (PTM) is based on the assumption that the semantics of human language is *intersubjective* in nature. The term *intersubjectivity* has long been associated with theories and practice in philosophy, cognitive science, and experimental and developmental psychology derived from, or influenced by *phenomenology*, a branch of philosophical thinking pioneered by the German philosopher Edmund Husserl in early 20th Century. It is also known, in the field of semiotics, as a central concern in the works of Walker Percy (Percy 1976). In this paper, however, we generally use this term in a more restricted sense, namely, to refer to the guiding principles for a specific empirical method, due to Raukko, for acquiring semantic information from non-expert informants in a speech community (Raukko 1999).

Another background framework of PTM is inspired by an idea from Wierzbicka's Natural Semantic Meta-

language (NSM) — that all complex meanings are decomposable into constituent parts that can be readily expressed in natural language.

Unlike NSM, however, PTM has a more practical goal and a more narrow scope, namely, that of extracting information to help differentiate a relatively small group of closely related words. Thus, instead of searching for and verifying whether a semantic feature is a proper universal primitive, we take a more ad-hoc approach, i.e. if it is evident from empirical data that a new feature would help distinguish one group of words from another, then we will adopt it at the next iteration of our investigation as one of the dimensions to test the population with, and deal with the theoretical issues later.

2.2 Practical Considerations

Since the very nature of PTM is to examine the productive use of actual everyday words in their natural settings, the tests need to be specifically tailored both for the words of interest, and for the study population. Also, it should be noted that PTM is by design intended as a re-iterative process, where each test round generates hypothesis to be tested in the following round.

2.2.1 Tailoring the Tests for Features Specific to the Words under Investigation

While some semantic dimensions are common to all vocabulary, many words or word groups also have their own unique semantic characteristics that are not apparent at first, even to a trained semanticist. These subtle nuances often do not come out automatically in conscious explanations, but can nevertheless be drawn out very prominently with the right kind of testing (see Vanhatalo 2002a, 2002b).

For instance, while most native English speakers would instinctively choose either *shout* or *yell* in his or her speech, such speakers are often at a loss *at first* when asked to explain why one choice is made over the other.

To draw out such hidden linguistic intuition, we need to think of some non-rigid way of testing that encourages creative brainstorming. In PTM this often comes in the form of a natural-sounding, open-ended task given in a non-pressured setting, such as a free-form question framed in a plausible context, e.g. "You've just met an exchange student from Japan. She would like to know what the difference is between *shout* and *yell*. How would you explain the difference to her?"

Finally, a practical concern is that the number of semantic features for words in any given word group can be quite large. However, since we are only interested in differentiation among these closely related words, we can choose only features that contribute to such differentiation. Furthermore, because of the re-iterative nature of PTM, the feature set for each group of words can grow or shrink as we go.

2.2.2 Testing in Settings that are Realistic for the Informants

In order to generate data that are as authentic as possible, our test settings are tailored so that they are natural for each informant group. For instance, since it would appear more natural for high school students to explain the difference between words to their friends, or to place themselves in situations that are plausible for an average teenager, our tests for them are designed accordingly.

Below in Fig. 1 is an example of a multiple choice task where the various near-synonyms for the Finnish version of the verb "to nag" is used in a realistically plausible setting (for the Finnish high school students who were our informants):

HOW ANGRY WAS YOUR MOTHER?

1 -- a little angry; 5 -- very angry.

	1	2	3	4	5
Yesterday I came home late and Mom *jäkätti*.	—	—	—	—	—
Yesterday I came home late and Mom *valitti*.	—	—	—	—	—
Yesterday I came home late and Mom *marisi*.	—	—	—	—	—
....	...	...	...	...	...

Fig. 1 A Multiple Choice Question

2.3 Details of one Pilot Study: Procedures and Results

We have conducted several pilot studies with over 450 subjects in Finland and Estonia to date. One such study was carried out with 154 high school students in Finland. The tests were delivered on paper. The tested vocabulary comprised of 18 speech act verbs that describe complaining (e.g. English "to nag" or "to carp") in Finnish (see Appendix A for the list with English glosses.) According to existing dictionaries, these words are considered near-synonyms. The tasks constituting the testing were either *production tasks* or *multiple choice tasks*.

In most production tasks (i.e. open-ended tests), the informants were asked to compare two or more near-synonyms, often by explaining them to their non-native peers. In the analysis phase, features in their descriptions were extracted and collected into matrices, which were then used to generate frequency charts for compilation of further test series. Semi-quantitative comparisons were also performed with the results from multiple choice tasks. The most surprising observations were the abundance of discriminating features between words,

and the high frequency of some answers (e.g. reasons for a certain speech act).

In multiple choice tasks (i.e. difference evaluation tests), the informants were requested (1) to choose the best word for the given context, (2) to choose the best context for the given word, or (3) to rate/rank the word in a given semantic dimension. All these results were analyzed statistically. Tasks requiring word ranking or rating yielded direct numerical values with measures of variance.

An example of numerical rating of a semantic dimension is given in Figure 2, where the informants were asked to rate volume of the speech act on a scale of 1 to 5. It appears that the assumed near-synonyms are clearly distinguishable in this semantic dimension, and the calculated confidence intervals (short vertical bars) demonstrate the high consensus among the informants.

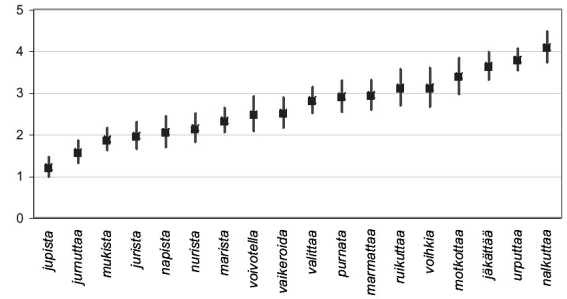

Fig. 2 Volume of the Speech Act

An example of ranking between near-synonyms is given in Figure 3, which shows the result of a task to select the gender (of the agent) in the speech act. The result reveals that some verbs are clearly associated with female or male gender, while others are not as clearly gender-associated.

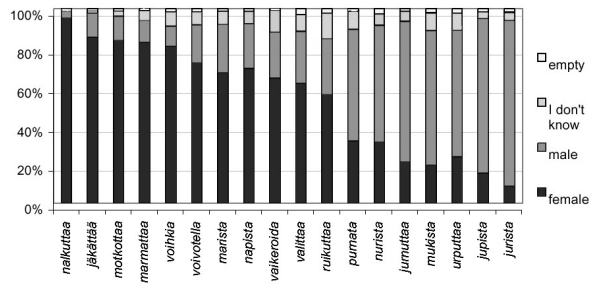

Fig. 3 Gender (of Agent) in the Speech Act

3 Visualizing Near-Synonymy in a Quasi 3-D Space

3.1 The Need for an Intuitive Way to View and Review Semantic Information

While we are in general satisfied with the results from the first round of our pilot study, we have come to realize that, in order to pass the results back to the informants for the next iteration of our test process, we need to present our findings in a more intuitive format.

Furthermore, as researchers engaging in the design of modern electronic dictionaries and thesauri for human users, we are interested in creating a user friendly interface for a thesaurus like application. Indeed, we have in mind that our informants would also be the users of such a thesaurus/dictionary, and thus have an incentive to make contribution to its continuing update and improvement. The general configuration of such a setup is illustrated in Fig. 4 below and described in more detail in a forthcoming paper (Vanhatalo 2003).

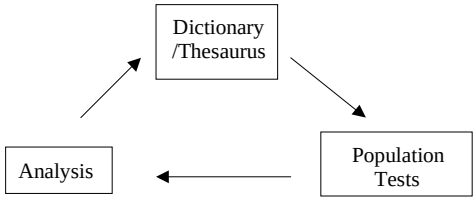

Fig. 4 A New Kind of Dictionary/Thesaurus

The convergence of these interests and requirements resulted in the prototype visualization tool, currently implemented as a Java applet, described in the following sections.

Fig. 5 Single-Axis Layout Based on *gender*

The screen-shot of the visualization tool in Fig. 5 above contains essentially the same information as in the bar chart of Fig. 3, except here the words themselves are the main objects being displayed. The words are distributed along a diagonal axis based on gender (of the agent), with lower left being more "male"-like, and upper right being more "female"-like.

The view shown in Fig. 6 is similar, but in this case we use x-axis for gender and the y-axis for volume. In other words, Fig. 6 contains the same information as those in Fig. 2 and Fig. 3 combined.

Actually, there is more. In both Fig. 5 and Fig 6 there is a third (z-) dimension shown via type size and color. This dimension is currently used to represent the *semantic distance* of each word from the *focus*, i.e. the currently selected word of interest highlighted in a box. The basic idea is that the word of interest would be closest to the viewer, and thus largest in type and darkest in color; while the other words (its near-synonyms) will be nearer or further from the viewer depending on how close they are semantically to this focus word. In other words, we hypothesize that the viewer would have an intuitive feel for this notion of 'semantic distance', and that he or she would instinctively translate this mental distance into a perceived visual distance, and vice versa.

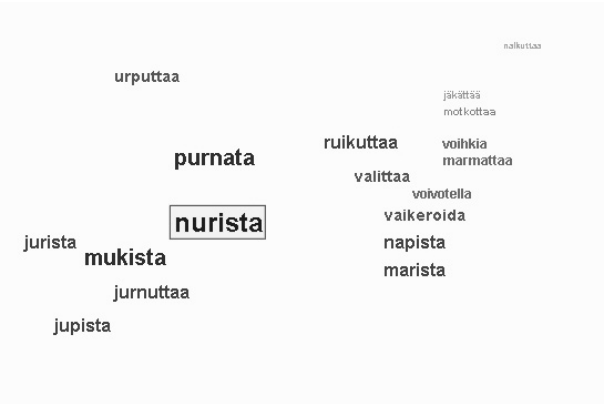

Fig. 6 Dual-Axis Layout (*gender* x *volume*)

For example, in both Fig. 5 and Fig. 6 above, where the semantic distance is calculated as a weighted average of six semantic dimensions, one could see at a glance which word(s) are the closest near-synonyms to the selected focus word. Thus, in Fig. 5 for *vaikeroida* it is *voivotella*; while in Fig. 6 for *nurista*, they are *purnata* and *mukista*.

3.2 Towards a Web-Based, Visually Enriched Extension to PTM

We envision this Web-based, visual extension to PTM to work mostly the same as the paper-and-pencil version, except that (1) it would be conducted over the Internet; and (2) some, though not all, of the tests would be more visually oriented.

More specifically, the visual tests would still consist of both multiple-choice tasks and open-ended tasks as before.

For an open-ended task, one option is to present one of these displays, say Fig. 5 or Fig. 6 above, and ask the informants if the picture makes sense, and if not, to explain in their own words what in the picture appears *odd* to them, and why.

As to multiple-choice tasks, consider the four views in Fig. 7a - 7d below:

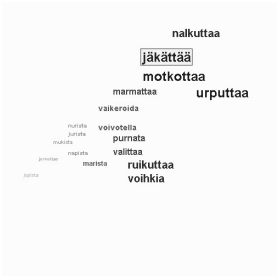

Fig. 7a

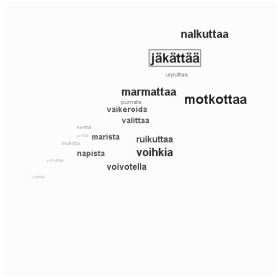

Fig. 7b

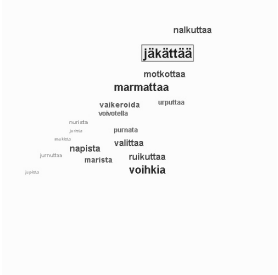

Fig. 7c

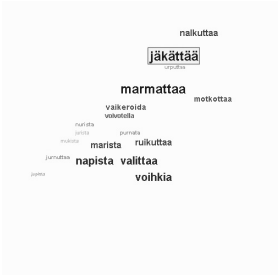

Fig. 7d

These four views are essentially variations on the same theme, i.e. they have the same x- and y-axis layout and the same focus word. The only difference among them is in the weight assignment for calculating semantic distance from the focus word *jäkättää*. The viewer will be asked to rank the four displays in terms of their "naturalness".

Another way to do this will allow the weights to be assigned by the viewer directly, e.g. via a set of sliders similar to those for photo manipulation programs. While this would require more work for the informant, it could actually be more *fun* and thus perhaps would have a greater potential as a successful method for an Internet based approach.

4 Discussion

We are primarily researcher and practitioners in the field of foreign language studies, and our interests are focused on the design and implementation of electronic dictionaries and thesauri for human users who are studying a second language. Nevertheless, we have benefited

greatly from exposure to research done in computational linguistics, and look forward to the exchange of ideas that we hope would benefit both of our fields.

In this paper we presented an approach towards solving the problem of building a large scale lexical database with specific emphasis on near-synonymy. As we are still at a very early stage of our investigation, many unanswered questions remain.

First of all, while we have confidence that our 'informant-friendly' *intersubjective* approach can extract good semantic information, we are less sure that *all* the information thus collected can always be easily converted into some numeric format, or be intuitively representable in some visualized form.

Another concern has to do with the assumption that the informant would intuitively perceive visual similarity as semantic similarity. As this has not yet been tested, we simply do not know if it will work as hoped. We also wonder if purely visual design factor (e.g. color clashes, compositional imbalance) could inadvertently skew an informant's judgment on a particular display's semantic "naturalness".

Lastly, doing survey of any sort on the Internet involves a whole set of issues that we are aware of, but have not yet seriously investigated.

Despite these uncertainties, we are in general very optimistic about the direction we are heading. We envision the next phase of our research to involve scaling up the testing, to include both more word groups and many more (in the thousands, ideally) informants, possibly via the Internet but more probably a large university's internal network, in our first venture into the Web-based survey world.

We would also like to do some more experiments with the visualization tool, e.g. to try out different schemes for calculating semantic distances, to use data from other databases (and in other languages, e.g. English), or to create a more appealing, 3-D game like user interface. Perhaps even a "space war" type game for near-synonymy. *Maybe*.

Appendix A. "Nag" Verbs in Finnish

jupista	mutter, mumble; grumble
jurnutta	(colloq.) annoy, vex
jäkättää	(colloq.) [yakety-] yak; nag
marista	whine, whimper, fret, grumble
marmattaa	grumble
motkottaa	carp, nag
mukista	grumble, grouse
nalkuttaa	nag, carp
napista	grumble, gripe, murmur
nurista	grumble
purnatta	grouse, grumble
ruikuttaa	whine, whimper, complain, (colloq.) moan, wail, (colloq.) pester
urputtaa	-- N/A --*
vaikeroida	moan, groan, wail, lament, bemoan
valittaa	groan, moan, wail, lament, complain
voihkia	groan, moan
voivotella	moan, whine, bewail

(edited from Finnish-English General Dictionary 1984)

* The word *urputtaa* is not yet found in current Finnish-English dictionaries, though it has been collected into the more recent monolingual Finnish dictionaries.

References

Edmonds, Philip: *Semantic Representation of Near-Synonyms for Automatic Lexical Choice*. PhD Thesis, University of Toronto (1999)

Edmonds, Philip and Hirst, Graeme. "Near-synonymy and lexical choice." *Computational Linguistics*, 28(2), June 2002, 105--144.

Finnish-English General Dictionary. Uusi suomi-englanti suursanakirja (1984). Raija Hurme, Riitta-Leena Malin, Olli Syväoja. WSOY. Helsinki.

Inkpen, Diana Zaiu and Hirst, Graeme. "Acquiring collocations for lexical choice between near-synonyms." In *SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Linguistics*, Philadelphia, June 2002.

Inkpen, Diana Zaiu and Hirst, Graeme. "Building a lexical knowledge-base of near-synonym differences." In *Proceedings of the Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, June 2001.

Percy, Walker: *The Message in the Bottle*, Farrar, Strauss and Giroux, New York, 1976

Raukko, Jarno 1999: “An ‘intersubjective’ Method for cognitive-semantic Research on Polysemy: The case of GET”. – In Masako K. Hiraga, Chris Sinha, and Sherman Wilcox (ed.), *Cultural, Psychological and Typological Issues in Cognitive Linguistics. Selected papers of the bi-annual ICLA meeting in Albuquerque, July 1995* pp. 87–105. John Benjamins. Amsterdam/Philadelphia. [Current Issues in Linguistic Theory, 152.]

Vanhatalo, Ulla 2002a: “Population Tests in Lexicography”. In Geoffrey Stewart Morrison & Les Zsoldos (ed.), *Proceedings of the Northwest Linguistics Conference 2002*. pp. 83-94. Available online at *http://edocs.lib.sfu.ca/projects/NWLC2002/NWLC2002_Proceedings_Vanhatalo.pdf*

Vanhatalo, Ulla 2002b: “Naiset motkottaa aiheesta ja nalkuttaa syyttä”: Kyselytestit verbien semanttisten sisältöjen arvioinnissa. [Using Questionnaires to Assess Semantic Content of Verbs]. In *Virittäjä* 106: 3. pp. 330-353.

Vanhatalo, Ulla 2003: “Finnish Electronic Dictionaries: Present and Future Challenges in Incorporating Semantic and Pragmatic Information”. In *Journal of Finnish Studies*. Forthcoming.

Wierzbicka, Anna 1996: *Semantics: Primes and Universals*. Oxford University Press.

Wierzbicka, Anna 1999: *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press.