Tracking the Lexical *Zeitgeist* with WordNet and Wikipedia

Tony Veale[1]

Abstract. Most new words, or neologisms, bubble beneath the surface of widespread usage for some time, perhaps even years, before gaining acceptance in conventional print dictionaries [1]. A shorter, yet still significant, delay is also evident in the life-cycle of NLP-oriented lexical resources like WordNet [2]. A more topical lexical resource is Wikipedia [3], an open-source community-maintained encyclopedia whose headwords reflect the many new words that gain recognition in a particular linguistic sub-culture. In this paper we describe the principles behind *Zeitgeist*, a system for dynamic lexicon growth that harvests and semantically analyses new lexical forms from Wikipedia, to automatically enrich WordNet as these new word forms are minted. *Zeitgeist* demonstrates good results for composite words that exhibit a complex morphemic structure, such as portmanteau words and formal blends [4, 5].

1 INTRODUCTION

Language is a dynamic landscape in which words are not fixed landmarks, but unstable signposts that switch directions as archaic senses are lost and new, more topical senses, are gained. Frequently, entirely new lexical signposts are added as newly minted word-forms enter the language. Some of these new forms are cut from whole cloth and have their origins in creative writing, movies or games. But many are patchwork creations whose origins can be traced to a blend of existing word forms [1]. This latter form of neologism is of particular interest to the computational lexicographer, since such words possess an obviously compositional structure from which one can begin to infer meaning. In this paper, we demonstrate that, if given enough semantic context, an automated system can assign a sufficiently rich semantic structure to these words to allow them to be automatically added to an electronic database like WordNet [2]. When tied to a system for harvesting new word forms from the internet, this capability allows for a dynamic WordNet that grows itself in response to a changing language and cultural context.

Most neologisms bubble beneath the surface of widespread usage before they gain entry to a conventional dictionary. This is to be expected, since the internet is awash with idiosyncratic neologisms that lack both charm and staying power. Nonetheless, to experience the variety and inventiveness of the most creative new words in English, one need look no further than Wikipedia [3], an open-source electronic encyclopedia that is continuously updated by a on-line community of volunteers. If such words are likely to be encountered in any text to which NLP technologies are applied, from deep text understanding to shallow spell-checking, we should expect our lexical databases to possess a basic interpretation capability.

In this paper, we describe an automated system, called *Zeitgeist*, that harvests neologisms from Wikipedia and uses the semantic context provided by Wikipedia's topology of cross-references to add corresponding semantic entries to WordNet. In section two we briefly describe WordNet and Wikipedia, and outline the properties of each that are central to *Zeitgeist*'s operation. Our goal is to exploit only the topology of cross-references, rather than the raw text of the corresponding Wikipedia articles (which would necessitate heavy-duty parsing and analysis methods). Since some topological contexts are more opaque than others, *Zeitgeist* employs a multi-pass approach to acquiring new word forms. In the first pass, only clear-cut cases are harvested; these exemplars are then generalized to underpin schemata that, in a second pass, allow less obvious neologisms to be recognized and semantically analyzed. Both passes are described in sections three and four. In section five, an empirical evaluation and discussion of *Zeitgeist*'s results is presented, while concluding thoughts are offered in section six.

2 LINKING WORDNET AND WIKIPEDIA

WordNet and Wikipedia each blur the traditional semiotic distinction between dictionaries and encyclopedias – which views the former as a source of *word* knowledge and the latter as a source of *world* knowledge – in different ways. WordNet is primarily an electronic dictionary/thesaurus whose structure is informed by psycholinguistic research (e.g., it uses different representations for nouns, verbs, adjectives and adverbs), but in eschewing alphabetic indexing for a semantic organization, it imposes an encyclopedia-like topic organization on its contents. Its coverage is broad, containing entries on topics such as historical events, places and personages more typically found in an encyclopedia. Unsurprisingly, it tends to be used in NLP applications not just as a lexicon, but as a lightweight knowledge-base for reasoning about entities and events.

For its part, Wikipedia's topic articles are surprisingly word-oriented. One finds many more headwords than in a conventional encyclopedia, and a richer level of interconnectedness. In many cases, composite headwords (such as "feminazi") are explicitly linked to the entries for their component parts, while detailed articles on lexical phenomena such as blended (or portmanteau) word-forms [4, 5] and political epithets provide links to numerous topical

[1] School of Computer Science and Informatics, University College Dublin, Belfield, Dublin, Ireland.

examples. Additionally, a sister project, Wiktionary [6], aims to exploit the Wikipedia model for an open-source dictionary.

The advantages accruing from an integration of such complementary resources are obvious. To Wikipedia, WordNet can give its explicit semantic backbone, as found in the *isa*-taxonomy used to structure its noun senses. To WordNet, Wikipedia can give its rich, open-textured topology of cross-references [7], as well as its larger and constantly growing set of topical headwords. To achieve this integration, the headwords of Wikipedia must be sense-disambiguated, though [8] report positive results for this task. In this paper, we explore the extent to which the semantic head of a neologism (that part which contributes the suffix, partially or completely, such as "pub" in "Gastropub" and "economics" in "Enronomics") can be disambiguated by the priming effects of other links emanating from the same Wikipedia article. General purpose WSD techniques (e.g., [9,10]), applied to the text rather than the links of an article, can then be used to resolve those ambiguous heads that are not primed in this way.

Toward thus end, we introduce two connectives for relating Wikipedia headwords to WordNet lexical entries. The first is written x *isa* y, and states that a new synset {x} is to be added to WordNet as a hyponym of the appropriate sense of y. Thus, *superhero isa hero* assumes that WSD is used to identify the intended sense of "hero" in the "superhero" context. The second is x *hedges* y, as in *spintronics hedges electronics*. As described in Lakoff [11], a hedge is a category-building relationship that allows one to reason as if a concept belonged to a given category, in spite of strict knowledge to the contrary (e.g., most people know that whales are not fish, but reason about them as if they were). In WordNet terms, hedge relationships will ultimately be instantiated via taxonomic coordination: {spintronics} will not be added as a hyponym of {electronics}, rather both will share the common hypernym {physics}. Hedges allow us to sidestep the awkward issues of hyperbolae and metaphor that frequently mark new coinages. Though "affluenza" ("affluence + influenza") is not, strictly speaking, a kind of "influenza", the hedge allows an NLP system to reason as if it were a real virus; this is apt, since the blend is used to depict affluence as a contagious affliction.

3 PASS I: LEARNING FROM EASY CASES

We employ a string-matching approach to recognizing and analyzing Wikipedia neologisms, in which specific schemata relate the form of a headword to the form of the words that are cross-referenced in the corresponding article. Let $\alpha\beta$ represent the general form of a Wikipedia term, where α and β denote arbitrary prefix and suffix strings that may, or may not, turn out to be actual morphemes. In addition, we use $\alpha\rightarrow\beta$ to denote a reference to headword β from the Wikipedia article of α, and use $\alpha\rightarrow\beta\,;\gamma$ to denote a contiguous pair of references to β and γ from article α.

As noted earlier, Zeitgeist seeks out neologisms that are a formal blend of two different lexical inputs [4, 5]. The first input contributes a prefix element, while the second contributes a suffix element that is taken to indicate the semantic head of the neologism as a whole.

The first schema below illustrates the most common arrangement of lexical inputs (as we shall see in section 5):

Schema I: Explicit Extension

$$\frac{\alpha\beta\rightarrow\beta \;\wedge\; \alpha\beta\rightarrow\alpha\gamma}{\alpha\beta \;\textit{isa}\; \beta}$$

This schema recognizes blended word forms like "gastropub" and "feminazi" in which the suffix β is a complete word in itself (e.g., "pub" and "Nazi"), and in which the prefix α is a fragment of a contextually linked term (like "gastronomy" or "feminist"). The suffix β provides the semantic head of the expansion, allowing the new term to be indexed in WordNet under the appropriate synset (e.g., {Nazi} or {pub, public_house}). The textual gloss given to this new entry will be a simple unpacking of the blended word: "$\alpha\gamma$ β" (e.g., "gastronomy pub" and "feminist Nazi"). To avoid degenerate cases, α and β must meet a minimum size requirement (at least 3 characters apiece), though in some exceptional contexts (to be described later), this threshold may be lowered.

Many neologisms are simple variations on existing terminology. Thus, "fangirl" is a male variation on "fanboy", while "supervillain" is a criminal variation on "superhero". When an explicit Wikipedia reference exists between these alternating suffixes, the new composite word can be identified as follows:

Schema II: Suffix Alternation

$$\frac{\alpha\beta\rightarrow\alpha\gamma \;\wedge\; \beta\rightarrow\gamma}{\alpha\beta \;\textit{hedges}\; \alpha\gamma}$$

This schema identifies a range of alternating suffix pairs in Wikipedia, from *man↔boy* to *woman↔girl* to *genus↔genera*, *bit↔byte* and *bacteria↔toxin*.

We can now begin to consider portmanteau words in which the suffix term is only partially present. Words like "Rubbergate" are understood as variations on other terms (e.g., "Watergate") if the prefix term (here, "rubber") is explicitly linked. In effect, a partial suffix like "gate" becomes evocative of the whole, as follows:

Schema III: Partial Suffix

$$\frac{\alpha\beta\rightarrow\gamma\beta \;\wedge\; (\alpha\beta\rightarrow\alpha \vee \alpha\beta\rightarrow\delta\rightarrow\alpha)}{\alpha\beta \;\textit{hedges}\; \gamma\beta}$$

This schema additionally covers situations where the prefix is only indirectly accessible from the neologism, as in the case of "metrosexual" (where "metro" is accessible via a link to "metropolitan"), and "pomosexual" (where "pomo" is only accessible via a mediating link to "postmodernism"). We note that this schema ignores the obvious role of rhyme in the coinage of these neologisms.

This indirection means that, in words like "metrosexual", both the prefix and the suffix may be partially projected to

form a true portmanteau word. In Wikipedia, the lexical inputs to a portmanteau word are often stated as contiguous references in the corresponding article. For instance, Wikipedia describes "sharpedo" as a "shark torpedo", while "Spanglish" is explicitly unpacked in the corresponding article as "Spanish English". We can exploit this finding in the following schema:

Schema IV: Consecutive Blends

$$\frac{\alpha\beta \rightarrow \alpha\gamma\ ;\ \delta\beta}{\alpha\beta\ \textit{hedges}\ \delta\beta} \qquad \textit{e.g., sharpedo} \rightarrow \textit{shark torpedo}$$

Indeed, portmanteau terms are so striking that the corresponding Wikipedia articles often explicitly reference the headword "portmanteau", or vice versa. In such cases, where $\alpha\beta \rightarrow$ *portmanteau*, we can safely reduce the minimum size requirements on α and β to two characters apiece. This allows *Zeitgeist* to analyze words like "spork" (spoon + fork) and "sporgery" (spam + forgery).

4 PASS II: RESOLVING OPAQUE CASES

The foregoing schemata anchor themselves to the local topological context of a headword to curb the wild over-generation that would arise from string decomposition alone. But even when this topological context is uninformative, or absent entirely (since some Wikipedia articles make no reference to other articles), a system may be able to reason by example from other, more clear-cut cases. For instance, there will be many exemplars arising from schemas III and IV to suggest that a word ending in "ware" is a kind of software and that a word ending in "lish" or "glish" is a kind of English. If E is the set of headwords analyzed using schema III and IV, and S is the corresponding set of partial suffixes, we can exploit these exemplars thus:

Schema V: Suffix Completion

$$\frac{\alpha\beta \rightarrow \gamma\beta \wedge\ \gamma\beta \in E\ \wedge\ \beta \in S}{\alpha\beta\ \textit{hedges}\ \gamma\beta}$$

Since the Wikipedia entries for "crippleware", "donationware" and "malware" – but not "stemware" or "drinkware" – make reference to "software", the above schema allows us to infer that the former are kinds of software and the latter dishware. Suffix completion reflects the way neologisms are often coined as reactions to other neologisms; for example, once "metrosexual" is recognized using schema III (partial suffix), it provides a basis for later recognizing "retrosexual" using schema V, since "sexual" will now suggest "metrosexual" as a completion. Similarly, "Reaganomics" serves as an exemplar for later analysing "Enronomics".

If P denotes the set of prefix morphemes that are identified via the application of schemas I, II and III, we can also formulate the following generalization:

Schema VI: Separable Suffix

$$\frac{\alpha\beta \rightarrow \beta \wedge \alpha \in P}{\alpha\beta\ \textit{isa}\ \beta} \qquad \textit{e.g., antiprism} \rightarrow \textit{prism}$$

This is simply a weakened version of schema I, where α is recognized as a valid prefix but is not anchored to any term in the topological context of the headword.

Though the entry "logicnazi" makes no reference to other headwords in Wikipedia, one can immediately recognize it as similar to "feminazi" (a "feminist Nazi" as resolved by schema I). Conceptually, "Nazi" appears an allowable epithet for an extreme believer of any ideology, and in part, this intuition can be captured by noting that the "Nazi" suffix overwrites the "ism" / "ist" suffix of its modifier. If T is a set of tuples, such as <ism, Nazi>, derived from the use of schema I, we have:

Schema VII: Prefix Completion

$$\frac{\alpha\gamma \rightarrow \alpha\ \wedge <\gamma, \delta\beta> \in T}{\alpha\beta\ \textit{isa}\ \beta}$$

Zeitgeist recognizes "logicnazi" as a kind of "Nazi", in the vein of "feminazi", since, from "logic" it can reach an "ism" or belief system "logicism" for this Nazi to extol. Likewise, it recognizes "Zionazi" as an extreme Zionist (allowing for a shared "n"), and "Islamonazi" as an extreme Islamist (allowing for an added "o" connective).

Finally, the collected prefixes and suffixes of pass one can now be used to recognize portmanteau words that are not explicitly tagged (as in schema V) or whose lexical inputs are not contiguously referenced (as in schema IV):

Schema VIII: Recombination

$$\frac{\alpha\beta \rightarrow \alpha\gamma\ \wedge \alpha\beta \rightarrow \delta\beta \wedge \alpha \in P \wedge \beta \in S}{\alpha\beta\ \textit{hedges}\ \delta\beta}$$

Thus, a "geonym" can be analyzed as a combination of "geography" and "toponym".

5 EVALUATION AND DISCUSSION

To evaluate these schemata, each was applied to the set of 152,060 single-term headwords and their inter-article connections in Wikipedia (as downloaded as a SQL loader file in June, 2005). Version 1.6 of WordNet was used to separate known headwords from possible neologisms. In all, 4677 headwords are decomposed by one or more of the given schemata; of these: 1385 (30%) are ignored because the headword already exists in WordNet, 884 (19%) are ignored because the hypernym or hedge determined by the analysis does not itself denote a WordNet term. Thus, though "bioprospecting" is correctly analyzed as "biology

prospecting", "prospecting" is not a lexical entry in WN1.6 and so this term must be ignored. The remaining 2408 (51%) of cases[2] are analyzed according to the breakdown of Table I:

Table 1. Breakdown of performance by individual schema.

Schema	# Headwords		# Errors	Precision
I	710	29%	11	.985
II	144	5%	0	1.0
III	330	13%	5	.985
IV	82	3%	2	.975
V	161	6%	0	1.0
VI	321	13%	16	.95
VII	340	14%	32	.90
VIII	320	13%	11	.965

Each *Zeitgeist* analysis was manually checked to find errors of decomposition and provide the precision scores of Table I. Two schemas (II in pass one, which e.g., derives Rubbergate from Watergate, and V in pass two, which e.g., derives retrosexual from metrosexual) produce no errors, while the most productive schema (explicit extension, schema I) has an error rate of just 1.5%. In contrast, schema VII (prefix completion in pass two, which derives logicnazi via the exemplar feminist/feminazi) is cause for concern with an error rate of 10%. High-risk schemata like this should thus be used in a controlled manner: they should not update the lexicon without user approval, but may be used to hypothesize interpretations in contexts that are more ephemeral and where more information may be available (e.g., a spellchecking or thesaurus application invoked within a particular document).

Some obvious factors contribute to an overall error rate of 4%. Company names (like Lucasfilm) comprise 12% of the erroneous cases, organization names (like Greenpeace and Aerosmith) 6%, place names (like Darfur) 11% and product names (like Winamp) 2%. Another 5% are names from fantasy literature (like Saruman and Octopussy). In all then, 35% of errors might be filtered in advance via the use of a reliable named-entity recognizer.

5.1 Word Sense Disambiguation

For 51% of the Wikipedia neologisms recognized by *Zeitgeist*, the semantic head (i.e., the word that contributes the suffix to the neologism) denotes an unambiguous WordNet term. The remaining 49% of cases thus require some form of WSD to determine the appropriate sense, or senses, of the semantic head before the neologism can be added to WordNet. While one can employ general purpose WSD techniques on the textual content of a Wikipedia article [9, 10], the topological context of the headword in Wikipedia may, to a certain degree, be self-disambiguating via a system of mutual priming.

For example, the intended WordNet sense of "hero" in the headword "superhero" (not present in WN 1.6) is suggested by the link *superhero→ Hercules*, since both "hero" and "Hercules" have senses that share the immediate WordNet hypernym {Mythological-Character}. In general, a given sense of the semantic head will be primed by any Wikipedia term linked to the neologism that has a WordNet sense to which the head relates via synonymy, hyponymy or hypernymy.

Priming can also be effected via an intersection of the textual glosses of WordNet senses and the topological context of the Wikipedia article (in a simple Wikipedia variation of the Lesk algorithm [9]). For example, the Wikipedia headword "kickboxing" suggests the ambiguous "boxing" as a semantic head (via schema I). However, because the Wikipedia link *kickboxing→fist* is echoed in the gloss of the WordNet sense {boxing, pugilism, fisticuffs} but not in the gloss of {boxing, packing}, only the former is taken as the intended sense.

More generally, the elements of the Wikipedia topological context can be viewed as a simple system of semantic features, in which e.g., *fist* is a feature of *kickboxing*, *fascism* is a feature of *Nazi*, and so on. Furthermore, because blending theory [4,5] claims that blended structures will contain a selective projection of elements from multiple inputs, this projection can be seen in the sharing of semantic features (that is, topological links) between the neological headword and its semantic head. For instance, the Wikipedia terms "Feminazi" and it semantic head, "Nazi", share three Wikipedia links – to Totalitarianism, Fascism and Nazism – which may be taken as the contribution of the lexical component "Nazi" to the meaning of the word as a whole. In the terminology of blending theory [4,5], these features are *projected* from the input space of Nazi into the blended space of Feminazi. Projection of this kind occurs in 64% of the neologisms recognized by *Zeitgeist*.

By understanding the projective basis of a word blend, *Zeitgeist* has yet another means of performing disambiguation of the semantic head, since the intended sense of the head will be that sense that visibly contributes semantic features to the blend. In the case of "kickboxing", the feature *fist* is directly contributed by the pugilistic sense of "boxing". However, for the blended word "emoticon", the feature *pictogram* is indirectly contributed by the user-interface sense of "icon" via its hypernym {symbol}.

Overall, topological priming resolves 25% of neologisms to a single WN1.6 sense, while another 1% are resolved to multiple WN senses, which is to be expected when the head element is a polysemous word. For instance, "photophone" ("photograph" + "telephone") is deemed to hedge both the equipment and medium senses of "telephone", while "subvertising" ("subversion" + "advertising") is deemed to

[2] Interestingly, the distribution for WN2.1 is much the same: 1570 analysed headwords (33%) are ignored because the headword is already in WN2.1, while 789 headwords (17%) must be ignored because their semantic heads are *not* in WN2.1. This leaves 2319 valid neologisms (49%) to be added to WN2.1, as opposed to 2408 for WN1.6. The number of neologisms remains relatively stable across WN versions because greater lexical coverage presents a greater opportunity to recognize neologisms that cannot be integrated into lesser versions. For instance, the "cyberpunk" entry in WN2.1 means that while this word is not treated as a neologism for this version (as it is for WN1.6), its presence allows "steampunk" and "clockpunk" to be recognized as neologisms.

hedge the message and industry senses of "advertising". In all, total WSD coverage in Zeitgeist is 77%. Recourse to more general WSD techniques is thus needed for just 23% of cases.

5.2 Literal Versus Figurative Interpretations

Our evaluation reveals that over half (57%) of the neologisms recognized by *Zeitgeist* (via schemas I, VI and VII) are realized in WordNet via a simple hypernymy relationship, while the remainder (43%) are realized (via schemas II, III, IV, V and VII) using the more nuanced *hedge* relationship. It seems clear, for instance, that "Gastropub" really is a kind of "pub" and "cocawine" really is a kind of "wine" (with added cocaine). However, it is not so clear whether Feminazis are truly Nazis (in the strict, National Socialist sense), so hedging may be more prevalent than these figures suggest. Though WordNet defines {Nazi} as a hyponym of {fascist}, the word is often used as a highly charged pseudo-synonym of the latter. "Nazi" seems to be used here in a sense-extensive, metaphoric fashion to suggest totalitarian zeal rather than political affiliation.

Two factors alert us that this use of "Nazi" is hyperbolae rather than literal extension. The first is the orthographic form of the word itself, for while "Nazi" is a proper-named class, "Feminazi" employs the word in an uncapitalized form which suggests a process of semantic bleaching or generalization. The second factor is the relative contribution, in terms of projected features, of the semantic head to the blend as a whole. Recall that the word "Nazi" shares the Wikipedia linkages {Totalitarianism, Fascism, Nazism} with "Feminazi", so these features may be said to originate from this input. However, "fascist" also references the terms {Totalitarianism, Fascism, Nazism} in Wikipedia, suggesting that there is no obvious loss of semantic import if Feminazi is considered an extension of {fascist} rather than of {Nazi}.

In 36% percent of neologisms, one or more semantic features are projected into the blend by a hypernym of the semantic head. In just 2% of neologisms this projection occurs in the context of an *isa* relation (i.e., via schemas I and VI) and is such that all features that are projected from the head are also redundantly projected from the hypernym of the head. (As it happens, only in the case of "Feminazi" does the semantic head denote a proper-named concept). While not conclusive, such redundancy is sufficient cause either to hedge the relationship or to prompt for human guidance in these cases.

6 CONCLUSIONS

We have presented a linguistics-lite approach to harvesting neologisms from Wikipedia and adding them to WordNet. *Zeitgeist* does not employ an explicit morphological analyser, but relies instead on a marriage of partial string-matching and topological constraints. Nonetheless, many of the words that are successfully recognized exhibit a creative and playful use of English morphology. Furthermore, by grounding is analyses in the local link topology of Wikipedia articles, *Zeitgeist* gains a semantic insight that one cannot obtain from morphology rules alone. For instance, not only is "microsurgery" recognized as a micro-variant of surgery, the specific meaning of "micro" in this context is localized to the headword "microscopy" via schema I. The concept "microsurgery" is not just "micro-surgery", but surgery conducted via a microscope.

Even a lightweight approach can, however, bring some degree of semantic insight to bear on the analysis of new words. In this respect, Wikipedia's link topology deserves further consideration as a source of semantic features. Certainly, Wikipedia has great promise as a semi-structured semantic representation. For instance, one can distinguish two kinds of semantic feature in Wikipedia. Strong or highly-salient features are those that are reciprocated; thus, *charity*→*altruism* and *altruism*→*charity* implies that altruism is a highly salient feature of charity, and vice versa. Weak features are those that are not reciprocated in this way. It remains to be seen how far one can go with such a representation without imposing a more rigid logical framework, but we believe that the initial foray described here suggests the scheme has yet more mileage to offer.

We conclude by noting that the linguistics-lite nature of *Zeitgeist's* approach means that is not intrinsically biased toward English. In principle, its mix of string matching and topological constraints should validly apply to other languages also. Whether phenomena like lexical blending spring forth with equal regularity in the non-English languages supported by Wikipedia is a subject of future research.

REFERENCES

[1] Dent, S., Fanboys and Overdogs: The Language Report III. Oxford University Press (2003).

[2] Miller, G. A., WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38 No. 11 (1995).

[3] *www.wikipedia.org*

[4] Fauconnier, G., Turner, M., Conceptual Integration Networks. *Cognitive Science*, 22(2), pp 133–187 (1998).

[5] Veale, T., O'Donoghue, D., Computation and Blending. *Cognitive Linguistics*, 11(3-4).

[6] *www.wiktionary.org*

[7] Ruiz-Casado, M., Alfonseca, E., Castells, P., Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. *LNAI 3513, pp 67 (2005).*

[8] Ruiz-Casado, M., Alfonseca, E., Castells, P., Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. Springer *LNAI 3528, pp 280 (2005).*

[9] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of ACM SigDoc Conference, Toronto: ACM*, pp 24-6 (1986).

[10] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11, pp 95-130, (1999).

[11] G. Lakoff, *Women, Fire and Dangerous Things: How the Mind forms Categories.* University of Chicago Press (1987)