Harvesting and understanding on-line neologisms

Tony Veale, Cristina Butnariu

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 (Dent 2003). This latter type 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 chapter, 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 dictionary like WordNet (Miller 1995). When tied to a system for harvesting new word forms from the internet, this capability allows for a dynamic dictionary that expands 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 (*www.wikipedia.org*), an open-source electronic encyclopedia that is continuously updated by an on-line community of volunteers. If such words are likely to be encountered in any text to which NLP (Natural Language Processing) technologies are applied, from deep text understanding to shallow spell-checking, we should expect our lexical databases to possess a basic interpretation capability for these neological forms. Indeed, the entire tone of a text-message (whether email, SMS, or an on-line discussion forum) can pivot around a single neologism with a highly polarizing affect, such as

"logicnazi" or "Feminazi" (to consider just two words discussed in this chapter). Such neologisms serve more than a decorative role in a text, to mark their user as a wordsmith; they often serve as dense descriptors that convey (or allude to) a great deal of information in a single lexical design.

1.1. Lexical Creativity as Variation and Combination

Though it is possible to cut a new word form from entirely whole cloth, most neologisms employ a combination of existing elements, and many of these new forms are clear variations of well-known words and phrases. Suppose you ask for a drink and a friend offers you a "virgin Mary". Knowing that a "bloody Mary" is a cocktail made from tomato juice, vodka and Worcester sauce, you can assume that a "virgin Mary" is a variant of this particular drink. Interestingly, though it is the word "bloody" in the normative version of the cocktail that is replaced with the word "virgin", you are unlikely to assume that it is the blood-red tomato juice that is replaced, but the alcoholic vodka. In this case, "virgin" suggests "chastity", which suggests "abstinence", which suggests "temperance", which implies a lack of alcohol. As such, one can argue that this linguistic concoction works better at a deep conceptual level than at a superficial linguistic level.

In contrast, there is no such cross-talk between the lexical and the conceptual levels in the Australian name for this cocktail, a "bloody shame". This second variation is arguably the more ingenious and humorous of the pair, for a number of reasons: first, because "bloody shame" is already a familiar phrase in English, and so this variation establishes a punning relationship between the new cocktail and its source norm; second, because this pre-existing phrase has a negative connotation, of "regret" or "tragedy", and this allows the variant name to express a negative view of the underlying concoction; and third, this negative perspective also expresses a strong cultural preference for alcohol that serves to reinforce the stereotypical hard-drinking Australian self-image. Drinkers who order a "bloody shame" thus communicate a disdain for their own choice while implying a desire to order something else, something a good deal more alcoholic, and order it in a way that humorously seems to crave our sympathy. Indeed, because the phrase

"bloody shame" has connotations of tragedy, it also works as a form of epic irony when used as the name of an alcohol-free cocktail.

The names "virgin Mary" and "bloody shame" are both variants on a shared norm that use a single-word replacement strategy to achieve a new but somewhat familiar meaning in a new but somewhat familiar form. The key to their success is their reuse of recognizable elements (either "Bloody" or "Mary") from the original phrasing. But all variants on a norm are *not* creatively equal: while both variations work quite well (as evidenced by their widespread use on cocktail menus), the latter achieves the greatest degree of creative duality, compressing multiple levels of meaning and perspective into a simple two-word name. In this chapter we shall focus on how multiple meanings can likewise be compressed and fused into a new two-part unit, to create not phrases but single neological word-forms. As with phrase level creativity, these new forms will reuse and combine recognizable (sub-word) elements from existing forms, to create something new that is at once surprising and familiar.

1.2. Principles of Ergonomic Word Design

Words are everyday things, as central to our daily lives as the clothes we wear, the tools we use and the vehicles we drive. As man-made objects, words and phrases are subject to many of the same design principles as the consumer artefacts that compete for our attention in the marketplace. In his book *The Psychology of Everyday Things* (later reissued as *The Design of Everyday Things*), Donald A. Norman (1988) identifies two key principles of artifact design: visibility and mapping. A good design makes it easy for a user to mentally visualize, or conceptualize, the inner workings of a product, while a bad design causes a user to construct an inaccurate conceptual model that leads to misuse of the product and inevitable human error. If well-designed, the external elements of a product will yield a natural mapping to its internal functions, but if badly designed, the mapping between appearance and function will be confusing and counter-intuitive. These principles are just as applicable to phrases like "Virgin Mary" as they are to refrigerators and car stereos. In the case of a "Bloody Mary" we have a partially visible conceptual model, in which the redness of tomato juice matches the redness of blood, but in which "Mary" does not appear to match anything at all. In

the case of a "Virgin Mary", even this partial visibility is completely undone. If Norman were to analyze phrases such as these in the same way he analyzes the workings of a Mercedes Benz, he would conclude that "Bloody Shame" is the phrase design with the highest visibility (of meaning) and the most natural mapping (of linguistic form to underlying conceptual structure).

Manufacturers place new kinds of ovens, televisions and automobiles on the market all the time, but users do not need to relearn basic behaviours like baking, watching TV or driving to work. These new products are usually variants of existing models, adding new functionality and subtlety to familiar forms that retain their underlying structures. Likewise in language, new coinages frequently borrow the form of existing phrases, allowing a user to reuse the same underlying conceptual model. Thus, when presented with the novel coinage "Ghost airport", we do not attempt to construct a new conceptual model from first principles; rather, we reuse the conceptual model of "Ghost town", by accepting that an airport is sufficiently similar to a town for the meaning of "Ghost airport" and "Ghost town" to be analogous (towns and airports tend to be filled with people, thoroughfares and businesses, while "ghost" variants are empty and desolate). The same principle also applies to the creation of new words from existing elements; you may never have encountered the term "Twitchhiking" before, but the word shares enough structure with "hitchhiking" to strongly suggest that the conceptual model for the latter can safely be reused. Depending on your technological savvy, you may well guess that "Twitch" is a blend of "hitch" and "Twitter", and integrate your knowledge of this new form of electronic communication into the conventional model of hitchhiking.

Variation of an established word or phrase is a common strategy in linguistic creativity, and indeed, the lexicographer Patrick Hanks (2004) argues that it is our dominant means of doing meaningfully novel things with language. But of course, not every variation will be creative. For instance, variations in how a word is spelled or pronounced can yield a more or less creative pun, but random typing errors are highly unlikely to yield anything we might consider creative. We should allow for serendipitous creativity that is unintentional, or the product of purely random combination or mutation, but almost all random variations will be uncreative, or else the very idea of creativity becomes devalued. What gives a linguistic variation its creative value is the transformation it yields in our understanding of the underlying idea. The variations "virgin

Mary" and "bloody shame" yield non-alcoholic variations of a popular vodka cocktail, and the latter name even serves as a negative judgement on the resulting concoction. The phrases "Virgin Mary" and "Bloody Shame" have long been in common usage outside the domain of mixed drinks, and have obvious similarities to the phrase they are candidates to replace, so variation and integration of word-elements at the lexical level is relatively straightforward to achieve. What is most interesting is how well, and how clearly, the combined elements map to the conceptual level. In this chapter we shall explore a particularly ergonomic means of combining word elements to generate *designer* word-forms that exhibit the visibility and mapping that is called for by principles of good design.

1.3. Designer Words

With visibility and mapping our chief concerns, we confine our computational exploration to one particular kind of designer word – the *portmanteau* word (Deleuze, 1990:42). As first coined by Lewis Carroll in *Through the Looking Glass*, a portmanteau has "two meanings packed up into one word" (1887:114). Carroll delighted in creating apparently nonsensical words using the portmanteau principle, such as "slithy" ("lithe" and "slimy") and "snark" ("snake" + "shark"), which are lexically suggestive if not exactly semantically transparent. Modern uses of the portmanteau principle generally aim for greater transparency, allowing a reader to infer the constituent words (and thus, ideas) from which the neologism is blended. For instance, the historian Niall Ferguson recently coined a new portmanteau, "Chimerica", to describe the heavily inter-dependent relationship between the U.S.A. and the People's Republic of China[1]. As Ferguson (2007) puts it, America and China are no longer two distinct countries from an economic perspective, but one blended economic whole that Ferguson chooses to call *Chimerica*. Ferguson's coinage results in a rather ugly word, but it is a word with some interesting properties nonetheless. For one, the word *Chimerica* resembles the Greek word "Chimera", a mythical monster that combines parts of other fabulous beasts, such as the body of lioness and a tail with a snake's head. The word "Chimera" is also used in modern genetics to describe a single organism with genetically distinct cells from two different zygotes. This is essentially what a portmanteau word

is: a neologism that results from the cross-breeding of words. Another interesting property, then, is the suggestion of conceptual unity that arises from the structure of a portmanteau word: the tight lexical integration of two distinct word-forms into a unified lexical whole suggests an equally tight integration of ideas at the conceptual level.

Humorous effects can arise when integration at the lexical level forces together ideas that one might inconsider incompatible at the conceptual and pragmatic levels. For instance, the portmanteau term "Feminazi" attempts to equate the strident expression of feminism with the Socialist Nationalism of Nazi Germany in World War II. The humour arises here from the clash of semantic frames associated with the terms "Feminist" and "Nazi", though the frames can be reconciled somewhat (at least at a superficial level) by recognizing both types of agent as zealous advocates of a particular social philosophy. Nonetheless, the interpretation process asks us to shift or project key elements of the "Nazi" frame into that of the "Feminist" frame to achieve an integration of both (see Coulson, 2000; Raskin, 1985 offers an earlier "semantic" interpretation along the same lines). Because this conceptual integration is somewhat less successful than the lexical integration (at least to those with moderate political views), the resulting neologism highlights rather than downplays the incompatibilities between both world-views; so while the lexical integration blurs the differences between the surface words, the conceptual integration highlights the differences between the deeper ideas, resulting in a humorous artifact (Pollio, 1996).

1.4. A computationally-driven approach to Neologisms

In this chapter we adopt an applied, computationally-driven approach to the analysis of neological portmanteau words. No single cognitive or linguistic theory is employed as a theoretical motivation for the work, though the approach is undoubtedly compatible with a number of such theories (Plag 1999, Kemmer 2003, Fauconnier and Turner, 1998). For instance, portmanteau blend words have been studied within the context of conceptual integration networks as championed by Fauconnier and Turner (1998), where such words belong to the category of formal blends. The theory of Conceptual integration networks, more commonly known as blending theory, posits a number of optimality princi-

ples for understanding how content from multiple conceptual spaces can be selectively projected and integrated into a new conceptual space, called the blend space. For instance, the word "chunnel" is a formal blend of two words, "channel" and "tunnel", which frequently collocate via the compound phrase "channel tunnel" (an undersea tunnel that links Britain to the European continent). Blending theory equips us with a rich descriptive framework for simultaneously discussing both the conceptual and the linguistic insights that go into such a neological formation: e.g., not only are "channel" and "tunnel" phonologically and orthographically similar (both comprise two syllables and both end with the same one), they are conceptually similar too, since both are members of the category Pathway. Fauconnier and Turner argue, for instance, that those elements projected into the blend space should be sufficiently linked (in well-formed blends) to their original input spaces that the resulting blend can be decomposed to reveal these inputs (thus, "chunnel" has identifiable vestiges of "channel" and "tunnel"). Likewise, they argue that only those elements that have a good reason to reside in the blend space are actually projected into the blend space (thus, "ch" is present in the blend because it represents "channel", while "unnel" is there only because it represents "tunnel").

Though primarily a cognitive-linguistic theory of compositional meaning rather than an algorithmic model of composition, there are aspects of blending theory that are amenable to computational implementation. For instance, Veale and O'Donogue (2000) describe how a variety of the optimality principles that guide the blending process can be interpreted in terms of well-understood computational ideas, such as semantic networks, graph-theoretic representation, structure-mapping and spreading activation. The computational approach that we present here is certainly compatible with this computational view of blending theory, though ultimately, the principles we take from blending theory are not that different from the design principles advocated by Norman (1988). Whether designing words or artifacts, mapping and visibility are important considerations in any new design. Norman's terminology intuitively captures the optimality principles of *Web, Topology* and *Good Reason*, while allowing us to firmly fix our focus on neologisms as human-designed products that are intended to appeal to a particular marketplace and user-base.

In keeping with our computation-oriented and relatively theory-lite approach to the treatment of portmanteau words, we simply assume

that a suffix is any arbitrary subsequence of a word that is anchored to the end of that word, while a prefix is an arbitrary subsequence that is anchored at the beginning of a word. As such, prefixes and suffixes do not have to represent morphemes, free or bound, nor are they constrained to necessarily begin or end at morpheme boundaries. This assumption reflects the observation that many portmanteau words do not respect morpheme boundaries, while, more importantly, allowing our computational model to operate without a fixed inventory of known morphemes.

1.5. Structure and Rationale of this paper

In this chapter, we describe a fully-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 introduce 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.

Zeitgeist is clearly not intended as a cognitive model of how humans comprehend and create neological portmanteau words. Nonetheless, because *Zeitgeist* implements the simplest possible assumptions that actually work in a real, unsupervised environment, its mechanisms demonstrate the empirical validity of these assumptions. In particular, as the results of section 5 shall bear out, a computational approach such as *Zeitgeist* allows us to appreciate, in the context of real data, the relative merits and risks of different strategies for coining portmanteau words.

2. Linking WordNet and Wikipedia

WordNet is a structured network of word senses that offer comprehensive coverage for the English language (Miller 1995). In WordNet, word-senses correspond to lexical concepts – concepts that are directly lexicalized in the language as individual words or as stable collocations (compounds) of such words. Each lexical concept is thus represented extensionally, as a set of near-synonymous words that can each denote that sense in some particular context. These sets, called *synsets*, in turn serve as the vertices of the semantic network that is WordNet, linked together via a small set of lexicosemantic relationships such as hypernymy (X is a kind of Y), hyponymy (X is a generalization of Y), meronymy (X has a part Y) and holonymy (X is part of Y). For instance, the synset *{surgeon, operating_surgeon, sawbones}* is connected to the synset *{doctor, physician, dr., md, doc, medico}* by a hypernymy relation and to *{neurosurgeon, brain_surgeon}* by a hyponymy relation. WordNet is widely used by computationally-minded researchers of language not only for its wide selection of English word-senses (numbering more than 150,000), but also for its free availability and its unencumbered terms of usage.

Wikipedia does for print encyclopedias what WordNet has done for print dictionaries. Besides, Wikipedia is a knowledge repository that is entirely defined by its users. Whereas synsets provide the building blocks of WordNet, Wikipedia comprises a large, user-defined and richly interconnected space of head-terms (words and phrases) and their associated text articles. Each Wikipedia topic/article may reference any other, so that the backbone of Wikipedia can be perceived, much like WordNet's, to be a semantic network of connected lexical concepts. However, the scale of Wikipedia is such that most of nominal terms in WordNet correspond to head-terms with their own articles in Wikipedia, while most Wikipedia head terms do not occur in WordNet. As one might expect from such an open resource, Wikipedia users are free to add their own head-terms and articles on pet-topics of their own choosing, giving Wikipedia a distinctive diachronic edge over more traditional and centralized sources of knowledge.

WordNet and Wikipedia each blur, in different ways, the traditional semiotic distinction between dictionaries and encyclopedias, which distinguishes the former as a source of *word* knowledge from the latter as a source of *world* knowledge. 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 organiza-

tion, 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 (Fauconnier and Turner 1998, Veale and O'Donoghue 2000) and political epithets provide links to numerous topical examples. Additionally, a sister project, Wiktionary (*www.wiktionary.org*), 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 (Ruiz-Casado *et al*. 2005a), 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 Ruiz-Casado *et al.* (2005b) report positive results for this task. In this chapter, 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 Word Sense Disambiguation (or WSD) techniques (Lesk 1986, Resnik 1999), 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.

For this purpose, 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 1987), a hedge is a category-building relation-

ship that allows the speaker 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 (Fauconnier and Turner 1998, Veale and O'Donoghue 2000). 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

$$\frac{\alpha\beta\rightarrow\beta\ \wedge\ \alpha\beta\rightarrow\alpha\gamma}{\alpha\beta\ \mathit{isa}\ \beta}$$

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"). In other words, this schema is designed to handle impure cases of portmanteau word-formation, in which the right-most source-word is carried over whole while the leftmost is only partially projected. We consider this schema first since it is the simplest, both in terms of ease and precision of applicability (our analysis of section 5 will empirically bear out this fact) and in terms of semantic interpretability. For the fact that the suffix β is a free morpheme means that portmanteau words that adhere to this schematic form have an easily identifiable semantic head, under which we can consider the word αβ as a whole as denoting a specialization of this head meaning. The suffix β thus 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: "αγ 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. Note how the theory-lite computational philosophy described in section 1 manifests itself in this schema, by allowing us to opportunistically treat words like "fangirl" as portmanteau words whenever it is computationally expedient to do so. While it may seem more appropriate to treat "fangirl" as a solid compound, more like "bulldog"

or "fireman" than "Feminazi" or "gastropub", the Wikipedia context here is strongly suggestive of an analogical derivation from "fanboy". As a clear counterpart to "fanboy", "fangirl" is thus best seen as a integration of "fanboy" and "girl" in which the former is only partially projected, rather than an integration of "fan" and "girl" in which both elements are fully projected into the resulting blend space. To use the specific terminology of Fauconnier and Turner, the Wikipedia link topology here mirrors the most likely workings of the optimality principle called *Web*, inasmuch as it explains why an unpacking of the concept Fangirl should yield not Fan and Girl but Fanboy and Girl.

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 indirect accessibility 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

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

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 analyzing "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

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

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 as 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. The resulting portmanteau is orthographically economical because "Feminist" and "Nazi" are allowed to share the same "n", but it is also semantically economical insofar as this over-writing produces a tighter integration at the conceptual level. The "-ist" morpheme in English denotes an agent with a strong connection to a given idea (denoted by a preceding morpheme), and since "Nazi" too can be seen as a kind of sociopolitical agent, the resulting blend can be seen as a form of conceptual specialization: thus, if a "feminist" is an agent of "feminism", a "Feminazi" is a Nazi-like-agent of "feminism". Indeed, one could argue that the "ist" suffix triggers a generic frame structure, like Activity-Agent, that ultimately guides this blend, allowing "Feminism" to instantiate the Activity slot and "Nazi" to instantiate the Agent slot. However, we note that *Zeitgeist* makes no such determination, simply because it has neither the means nor the resources to avail of such linguistic and conceptual knowledge. As such, *Zeitgeist* represents a

minimal or baseline attempt at the understanding of portmanteau words, whose empirical performance will indicate whether such knowledge is actually needed in practice. 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 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	No. Headwords		No. 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 (cf. Elsen, this volume for an analysis of such names from a Gestalt-perspective).

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 (Lesk 1986, Resnik 1999), 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 (Lesk 1986)). 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 (Fauconnier and Turner 1998, Veale and O'Donoghue 2000) 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 (Fauconnier and Turner 1998, 2002, Veale and O'Donoghue 2000), "Feminazi" is a single-scope blend at the formal level but a double-scope blend at the conceptual level. Formally, which is to say from the perspective of word form, "Nazi" is projected unaltered into the resulting blend word while "Feminist" is necessarily truncated for reasons of euphony. However, from the perspective of conceptual structure, the integration process is highly selective about which elements from our consensus understanding of Feminists and Nazis are projected into the blend space. Only the most profiled (cf. Langacker 1991) and easily-caricatured aspects of each concept are considered for the blend, which is fitting since the result is memorable but glib: the blend views feminists as shrill and strident, and views Nazis as totemic embodiments of this stridency while freeing them (and thus, the resulting portmanteau word) of their devastating historical associations. However, though the neologism defines a hybrid entity that is simultaneously a feminist and a (metaphorical) Nazi, "Nazi" serves as the linguistic head of the new term and one can thus appreciate that it is the Nazi-like traits of the hybrid that are most emphasized by the neologism. 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 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 that suggests it has undergone a process of sense-extension and semantic re-profiling (cf. Langacker 1991). That is, "nazi" and "Nazi" are different words with related meanings, where the meaning of the former profiles only a minor (yet quite potent) aspect of the meaning of the latter. 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 analyzer, 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 its analyses in the local link topology of Wikipedia articles, *Zeitgeist* gains a semantic insight that cannot be retained from morphological 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 it 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.

Notes

1. A similar coinage is "Chindia", which alludes to the growing dominance and inter-dependence of China and India (e.g., see Sheth, 2007). Though "Chimerica" clearly does not denote a real country, it has been used to denote a fictional country in a politically-inspired video game called *Hidden Agenda*. Interestingly, this game was released in 1988, nine years before Niall Ferguson re-invented the term to describe a real-world power grouping.

2. 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.

References

Carroll, Lewis
1887 *Through the Looking Glass*. London: Plain Label Books.

Coulson, Seana
2000 *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction*. New York/Cambridge: Cambridge University Press.

Deleuze, Gilles
1990 *The Logic of Sense*. New York: Columbia University Press.

Dent, Susie
2003 *Fanboys and Overdogs: The Language Report* III. Oxford: Oxford University Press.

Fauconnier, Gilles, Mark Turner
1998 Conceptual Integration Networks. *Cognitive Science* 22(2): 133–187.
2002 *The Way We Think*. New York: Basic Books.

Ferguson, Niall
2007 Not two countries, but one: Chimerica. *The Daily Telegraph* newspaper (UK), March 4th edition.

Hanks, Patrick
2004 The syntagmatics of metaphor. *International Journal of Lexicography*, 17(3):245-274.
Kemmer, Suzanne
2003 Schemas and lexical blends. In *Motivation in Language: From Case Grammar to Cognitive Linguistics. A Festschrift for Gunter Radden*, Hubert Cuyckens, Thomas Berg, René Dirven, and Klaus-Uwe Panther (eds.), 69–97. Amsterdam/Philadelphia: Benjamins.
Lakoff, George
1987 *Women, Fire and Dangerous Things: How the Mind forms Categories*. Chicago: University of Chicago Press.
Langacker, Ronald W.
1991 *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Cognitive Linguistics Research. Berlin/New York: De Gruyter.
Lesk, Michael
1986 Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SigDoc, ACM*, 24–26.
Miller, George A.
1995 WordNet: A Lexical Database for English. *Communications of the ACM*. Vol. 38, No. 11.
Norman, Donald A.
1988 *The Design of Everyday Things*. New York: Basic Books.
Plag, Ingo
1999 *Morphological Productivity. Structural Constraints in English Derivation.* Berlin/New York: De Gruyter.
Pollio, Howard R.
1996. Boundaries in Humor and Metaphor. Jeffrey Scott Mio and Albert N. Katz (Eds.), *Metaphor: Implications and Applications*. Mahwah, New Jersey: Laurence Erlbaum Associates, 231–253.
Raskin, Victor
1985 *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel.
Resnik, Philip
1999 Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11:95–130.
Ruiz-Casado, Maria, Enrique Alfonseca, Pablo Castells
2005a Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. *Springer LNAI* 3513: 67.
2005b Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. *Springer LNAI* 3528: 280.

Sheth, Jagdish N.
2007 *Chindia Rising*. New Delhi: McGraw Hill India.
Veale, Tony, Diarmuid O'Donoghue
2000 Computation and Blending. *Cognitive Linguistics* 11(3-4): 253–282.