Exploring Linguistic Creativity via Predictive Lexicology

Tony Veale [1] and **Cristina Butnariu** [1]

Abstract. Creativity is not just a matter of generation, but of interpretation, since for creativity to be recognized, it must be interpreted (and not rejected) by other agents. In the domain of lexical creativity, which concerns the generation of innovative word forms, we can describe two kinds of creative process: *explanatory lexicology*, in which the lexical creativity of others is appreciated and understood; and *predictive lexicology*, in which an agent creates lexical innovations *ab initio*. We focus on the latter process in this paper.

1 INTRODUCTION

Morphemes, as the smallest meaning-bearing parts of language, serve as the semantic atoms from which complex linguistic meanings are built. From these atoms, speakers may synthesize an open-ended array of semantic molecules, from complex word forms to sentences and whole texts. This analogy between linguistics and chemistry is a productive one, allowing linguists to catalogue and understand the subtle interactions between language elements that produce meaning. Alas, the same cannot be said for linguistic creativity, and creativity in general, whose study is still more akin to alchemy than chemistry. Despite some influential metaphors, involving the exploration and transformation of conceptual spaces, creativity remains an elusive phenomenon to pin down in formal, computational terms.

In this paper, we examine one important area of linguistic creativity - neologism generation - from a perspective that is more akin to chemistry than alchemy. New word forms are constantly entering the lexicon to reflect the changing cultural environment [4], so a creative system capable to anticipating and understanding these lexical creations will enable a lexicon to grow itself automatically. There are two broad approaches to capturing these lexical innovations as they arise. The first, which can be called *explanatory* lexicology, attempts to structurally analyse newly minted word forms whenever they are found in an input text. For instance, once the word “metrosexual” is encountered for the first time, an explanatory system can attempt to assign a meaning by first decomposing the word into its constituent atoms of meaning, in this case the morphemes “metro” and “sexual”. Reasoning by example with similar words, such a system may decide that “metrosexual” is a variant of “heterosexual”, especially if the latter word is found to occur nearby. The local context may also provide enough information to decide that “metro” here means “metropolitan”, allowing the system to conclude that a “metrosexual” is a “metropolitan heterosexual”. This approach is entirely reactive; only when a new word has been encountered does the system attempt to assign it a meaning. Explanatory lexicology is thus better suited to analysing the creativity of others than to producing innovations of its own.

In contrast, the second approach - which we dub *predictive* lexicology - anticipates lexical creativity before the fact, thus allowing a system to exhibit genuine creativity. Predictive lexicology employs lexical and world knowledge to recognize the potential utility of a new word, by first identifying the concept that it would denote. A novel word form is then constructed from the appropriate morphemes to describe this concept. In the terminology of Boden [1,2], such a neologism is P-Creative - it is creative with respect to the knowledge of the system itself. A search for the word form in a large corpus, such as the world-wide-web, will then reveal whether the word-form is indeed creative in the wider historical sense, which Boden labels H-Creativity. Suppose that a system contains the following three language facts: the word “cartographer” denotes a person who makes maps; the bound morpheme “astro-” denotes the concept Star; and the collocation “star map” denotes a kind of map. From these three facts, predictive lexicology can be used to suggest the new word form “astrocartographer” as a label for people who make star maps.

A dynamic NLU system needs to employ both explanatory lexicology and predictive lexicology to address the issues of lexical creativity; the former allows a system to react appropriately when creative word forms are discovered in text, while the latter allows a system to generate creatively concise lexical descriptions of its own. We describe and evaluate an implementation of both approaches in this paper. We begin in section 2 by describing how the necessary combinatorial elements, morphemes, are identified and tagged using WordNet [5], a large-scale database of English words. We then outline, in section 3, how predictions about novel word forms can be made from these elements, and how these neologisms can be verified even when they are H-Creative. In section 4 we complement this approach with a simple model of explanatory lexicology, which looks to the on-line encyclopaedia Wikipedia [6] as a source of new word forms to analyse. We provide an empirical analysis in section 5, and conclude with some closing observations in section 6.

2 CONSTRUCTING A MORPHEME INVENTORY

Clearly, a sizeable inventory of semantic atoms will be needed to fuel a combinatorial approach to word creation. In this section we outline a semi-automatic means of constructing this inventory using WordNet. The approach is an example-based one: given a set of morphemes that typify the kind of semantic atom we require, this set then seeds a search process to find similar elements within a given search horizon. Consider the morpheme “-ology”, a meaning element of Greek origin that denotes a field of specialized study. Used as a seed, this morpheme will allow us to retrieve other morphemes, also (for the most part) of Greek origin, with which it can meaningfully combine.

From a given seed, other morphemes are retrieved using a process

[1] School of Computer Science and Informatics, University College Dublin, Ireland, email: {Tony.Veale, Ioana.Butnariu}@UCD.ie

of substitution. For instance, from “-ology” we retrieve from WordNet the word “astrology”, which provides the morpheme “astro-”. From “astro-” we retrieve “astronomy”, which provides the morpheme “-onomy”, and so on. To reduce the effects of noise, we require that the prefix morpheme and the suffix morpheme overlap by one character; in the case of morphemes of Greek origin, this yoking character will be “o”. More generally, we construct from a given morphemic stem α all paths of the form $*\alpha\rightarrow\beta{:}\alpha\rightarrow\beta*\rightarrow\beta{:}\gamma\rightarrow*\gamma\rightarrow...\rightarrow\delta{:}\epsilon$ up to a specified search horizon. This approach falls far short of identifying every productive morpheme in WordNet, but the branching factor of the search, combined with a liberal search horizon and a handful of productive seed morphemes, allows it to identify a highly productive inventory from which initial experiments with lexical creativity can proceed.

For example, the seed “-ology” retrieves (with a horizon of 10) 907 morphemes, like “psycho-“, “caco-” and ”-olyte”. As we might expect, different seeds from the same language family exhibit a strong overlap in the morphemes they retrieve. The morpheme “-ophobia” retrieves 378 morphemes (at horizon 5), 56% of which are also retrieved by “-ology”; the seed “-oscope” retrieves 333 morphemes, 95% of which are retrieved by “-ology” or “-ophobia”; and the morpheme “-ometer” retrieves 212 morphemes (at horizon 5), all of which are retrieved by either “-ology”, “-ophobia” or “-oscope”. This strong coherence suggests that these retrieved morphemes should interact well to frequently produce meaningful combinations.

2.1 Morpheme Annotation

Once the core inventory of morphemes is identified in this way, they are hand-annotated with their semantic interpretations. This annotation has two parts: the first is a word gloss, such as *astro=star*, *ology=study*, *onaut=explorer/traveller*, *oplasty=repair*, and so on; the second is a WordNet sense identifier to indicate where, in the WordNet noun taxonomy, a new word-form with a given morphemic suffix should be placed. For instance, the morpheme “-onaut” is associated with the WordNet synset {*traveller*}, while “-oplasty” is associated with the WordNet synset {*surgery, operation*}. New words with the “-oplasty” suffix can thus be entered into WordNet as specialized kinds of surgical procedure. While one can conceive of a machine-learning approach to this kind of annotation, the inventory size easily facilitates human annotation. The hand annotation process also allows errors of decomposition to be identified and the corresponding false morphemes to be filtered out.

3 PREDICTIVE LEXICOLOGY

The core inventory thus contains two types of elements, morphemes (partial words) and their annotations (whole words). These two types suggest two forms of combination and, therefore, two approaches to predictive lexicology.

3.1 Scattershot Generation

The first concerns itself directly with the combination of morphemic elements to create new word forms that are then validated in a corpus or on the world-wide-web. The form of predictive lexicology, which we dub *scattershot generation*, is not sufficiently sophisticated to recognize when a new word-form has lexical value if it cannot be validated in this way. Is the word “chrononaut” a creative neologism or a random mishmash of morphemes? Such a concoction may be genuinely creative, but without understanding the conceptual motivation for the combination, the scattershot approach cannot distinguish it from combinations that are simply inept. Scattershot is capable of just P-Creative innovation then - that is, it can only replicate neologisms that are new to itself but which are already known to the wider language community.

3.2 Conceptually-Constrained Generation

It is also meaningful to generate combinations of morpheme annotations, rather than of the morphemes themselves, to create not words but phrases. For instance, the phrase “food explorer” can be generated by combining the annotation “food” of the morpheme “gastro-” and the annotation “explorer” of the morpheme “-onaut”. The sensibility of this combination can then be validated using a corpus or web-search. Once validated, we then have a conceptual motivation for combining the associated morphemes, in this case to generate the new word “gastronaut”.

If the resulting neologism is validated in this way, we now possess strong evidence that the word means what our combination mechanism believes it to mean. In this instance, the neologism is still a P-Creative artifact, but one that can safely be added to the lexicon, with (in the case of WordNet) a textual gloss composed from the validated combination of annotations. The synset {*gastronaut*} can thus be added to WordNet as a hyponym of {*traveller*} with the textual gloss “food explorer”.

More importantly, if the resulting neologism is not validated via corpus or web-search, the validity of the corresponding gloss is still evidence that the neologism is both meaningful and H-Creative, rather than linguistically worthless. For example, the web-validated collocation “time traveller” provides evidence that “chrononaut” is not just new, but meaningfully creative. By demonstrating that an utterly new word denotes a legitimate concept in this way, a predictive system is capable of verifying its own H-Creative outputs. The parallel views offered by morpheme combination and annotation combination thus give a predictive lexicology system the ability to evaluate itself, rather than to wait for human approval.

4 EXPLANATORY LEXICOLOGY

Predictive lexicology is sharply limited in scope by the size of its morpheme inventory. While this makes the exhaustive combinatorial search of scattershot generation computationally feasible, it means that many creative neologisms cannot be replicated or even understood. This problem is exacerbated by the fact that many new word-forms exploit combinations of bound and free morphemes. The combination “gastropub”, for instance (a bistro-like pub with restaurant-quality food) combines the annotated morpheme “gastro-” with the freely occurring word-form “pub”. The set of such neologisms is considerably larger and less predictable than predicative lexicology can handle. If we cannot predict these word-forms, our system should at least be capable of explaining or understanding them whenever they are encountered.

As a reactive strategy, explanatory lexicology can only harvest new word forms, rather than create them itself. The web contains many idiosyncratic word forms, but most lack the staying power or charm to merit a place in the NLU lexicon. However, one reasonably authoritative and topical web-repository, the Wikipedia open-source encyclopaedia [6], contains a wealth of interlinked neologisms that can be harvested. Wikipedia not only provides a growing stock of words, but the means to analyze them via its rich network of inter-headword references [9,10]. For instance, the Wikipedia article for "gastropub" links to both the article for "pub" and for "gastronomy" (which we denote *gastropub→pub* and *gastropub→gastronomy*), providing a solid basis for a decompositional analysis. Wikipedia supports the following three types of analysis:

1. Bound morphemic prefix and suffix (e.g., *metrosexual, retrosexual*)
   Schematic form: $\alpha{:}\beta \rightarrow \alpha{:}\gamma \wedge \alpha{:}\beta \rightarrow \delta{:}\beta$
2. Bound morphemic prefix, free morphemic suffix (e.g., *gastropub*)
   Schematic form: $\alpha{:}\beta \rightarrow \alpha{:}\gamma \wedge \alpha{:}\beta \rightarrow \beta$
3. Free morphemic prefix, bound morphemic suffix (e.g., *Reaganomics, pomosexual*)
   Schematic form: $\alpha{:}\beta \rightarrow \alpha \wedge \alpha{:}\beta \rightarrow \gamma{:}\beta$

These schemata allow a system to unpack many of the novel headwords it harvests from Wikipedia. For instance, schema (3) allows the words "Reagonomics" and "Enronomics" to be unpacked as "Reagan economics" and "Enron economics" respectively, and indexed in WordNet as new hyponyms of economics.

5 EMPIRICAL ANALYSIS

For the experiments described here, we employ an inventory of 1097 unique morphemes: 625 suffix morphemes and 472 prefix morphemes - extracted from WordNet using the seed morphemes "-ology", "- ophobia", "-oscope", "-ometer" and "-oglyph" with a search horizon of 10. This inventory is hand checked and annotated as described in section 2.

5.1 Predictive Lexicology

This inventory allows for 295,000 different combinations of prefix and suffix (and over 138 million combinations of two prefix morphemes and one suffix morpheme). Only 1792 (1%) of these 295,000 prefix/suffix combinations already exist in WordNet 1.6. When each of these combinations are sought on the web (using AltaVista), 42810 (15%) are verified as P-Creative.

Given the considerable combinatorial potential of even this small inventory, we limit our web investigations into H-Creativity to the following sample of suffix morphemes: "-ographer", "-ography", "-ology", "-omania", "-onaut", "-ophobia" and "oscope". Combining these with all 472 prefixes generates 3304 new word forms. Of these, only 252 (7%) are already in WN1.6, while 1684 (51%) are web-validated using AltaVista as P-Creative. Those that cannot be validated via the web are subjected to a secondary analysis, whereby the corresponding annotations are combined to form a noun description (such as "time traveller" for "chrononaut") and this word combination is validated instead. This secondary web validation reveals that 654 (20%) of the 3304 word-forms are actually H-Creative (that is, novel and useful). Table 1 provides a breakdown of the analysis on a morpheme by morpheme basis.

5.2 Explanatory Lexicology

Looking to Wikipedia, we find that 1914 Wikipedia headwords (or 1number of atomic headwords, as downloaded in August 2005) can be explained in terms of schema (1) from section 4, that is, as a combination of a two morphemes - like "psychonaut" - from our acquired inventory. In addition, 502 of these headwords can be understood via schema (2), that is, as a combination of a catalogued prefix morpheme with another headword that references, or is referenced by, the original headword. For instance, "homomasculinity" can be understood as a variant of "masculinity" via the addition of the prefix "homo-". Indeed, 282 of these 502 cases, the prefix can be tied to another headword that references, or is referenced by, the headword under analysis. Thus, the "homo-" of "homomasculinity" can be understood in this context not to mean "man" but "homosexual". A mere 70 headwords can be analyzed by schema (3), that is, as a combination of a catalogued suffix morpheme and another headword that references, or is referenced by, the headword under analysis. Nonetheless, some creative combinations are uncovered this way, including "fontographer", "inkometer", "anvilology", "Islamophobia" and "Christianophobia". Indeed, the latter two suggest a productive pattern, *<religion>-ophobia*, that can feed back into the process of predictive lexicology.

Table 1. Results of web-validation for a selection of 7 suffix morphemes: *ographer=writer, ography=writing, ology=study, omania=obsession, onaut=traveller, ophobia=fear_of_, oscope=viewer*

Suffix	in WN1.6	on Web	P-Creative	H-Creative
-ographer	18(4%)	311(66%)	171(36%)	140(30%)
-ography	45(9%)	343(73%)	292(62%)	51(11%)
-ology	135(29%)	288(61%)	269(57%)	19(4%)
-omania	12(3%)	364(77%)	271(57%)	93(20%)
-onaut	3(1%)	317(67%)	147(31%)	170(36%)
-ophobia	2(1%)	373(80%)	281(60%)	92(20%)
-oscope	24(5%)	355(75%)	266(56%)	89(19%)

These schemata can also enlarge the morpheme inventory by acquiring new bound morphemes and their annotations. For instance, the word "parasitology" combines a suffix from the annotated inventory - "-ology" - with an unknown prefix morpheme, "parasito-". Nonetheless, because of the Wikipedia reference *parasitology→parasite*, this morpheme can be acquired with the annotation "parasite". In this way, an 130 additional prefix morphemes (such as "planeto"-, "cometo-", "matho-" and "blogo-") and their annotations, are automatically gleaned from Wikipedia. The topology of Wikipedia often suggests non-obvious annotations for these new prefixes. For instance, the trio of references *Danophone→Denmark*, *Denmark→Dane* and *Dane→Denmark* suggests that "Denmark" is a valid annotation for the newly acquired prefix "Dano-". Likewise, the nexus of references *Selenology→moon*, *moon→Selene* and *Selene→moon* suggests that "Moon" is a valid annotation for the prefix "Seleno-".

6 CONCLUSIONS

Predictive and explanatory lexicology have complementary strengths and weaknesses. Predictive lexicology has a limited perspective on the space of possible neologisms, but can anticipate H-creative innovations before they are created by the language community at large. Explanatory lexicology has a much broader perspective, one that encompasses the use of arbitrary free morphemes as word elements, but can only react to new words in a P-Creative fashion.

Nonetheless, explanatory lexicology can serve as a valuable input to the process of predictive lexicology. Firstly, its explanatory analyses can increase the inventory of annotated morphemes that fuel lexical predictions. In schema (1), for instance, this analysis includes an expansion of the bound morphemic prefix (as in "gastronomy" for "gastropub"), which in turn yields a corresponding suffix morpheme that may not already by part of the annotated inventory of morphemes. In such a case, the expansion itself serves as an annotation for this newly acquired suffix. Secondly, as noted in section 5, these explanatory analyses can be generalized to provide patterns for further predictive generation. For instance, from “Reaganomics”, the system can generate the pattern *<president>-omics is a hyponym of* {*economics*}, where WordNet provides the hypernymic generalization of “Reagan” as {*president*}. This pattern can then be used to generate other forms like “Clintonomics”, “Bushonomics” and “Nixonomics”. The extent to which these P-Creative forms can be validated on the web is an indication of the predictive value of the generalized pattern. Though considerable further research awaits on this topic, predictive lexicology is clearly a technology that can prove valuable in the development of adaptive, dynamic lexicons.

ACKNOWLEDGEMENTS

We would like to thank the Enterprise Ireland for supporting this research through a grant from the Commercialization Fund.

REFERENCES

[1] M. Boden, *The Creative Mind: Myths and Mechanisms*, Weidenfeld and Nicholson, London, 1990.

[2] M. Boden, Computational models of creativity, *Handbook of Creativity*, 351-373, 1999.

[3] A. D. Cruse, *Lexical Semantics*, Cambridge University Press, London, 1986.

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

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

[6] *http://www.wikipedia.org*

[7] A. Newell, J. G. Shaw, and H. A. Simon, The process of creative thinking.
In: H. E. Gruber, G. Terrell and M. Wertheimer (Eds.), *Contemporary Approaches to Creative Thinking*, 63-119. New York: Atherton, 1963.

[8] G. Ritchie, The Transformational Creativity Hypothesis, *New Generation Computing, special issue on Creative Systems* (forthcoming), 2006.

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

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