Acceptance as an Integral Factor in the Interpretation of Novel Words

Dorota Smyk-Bhattacharjee

Two contributing factors in the interpretation of the novel word are discussed in this paper: (1) speaker acceptability, and (2) the expected unfamiliarity with the new formation, of the potential reader. The study presented here is based on a corpus that comes from samples taken from English language blogs, and currently consists of one million words of text. Writers often give clear clues as to whether the words form part of their lexicon. They also make predictions of whether the word familiarity will be shared by the readers. After filtering out the already established words, the remaining novel formations are classified according to the lexical distance manifested by the writer.

Key words: word-formation, interpretation, acceptance, blogs, corpus

1. Introduction

There is general agreement that out of all language domains "[v]ocabulary items tend to be added, replaced, or changed in meaning more rapidly than any other aspect of language" (Aitchison 2001: 16-17). Being at the forefront of language change, novel formations do indeed offer a fascinating and at the same time challenging object of study. Let me mention three problem areas that are relevant to this paper. Firstly, there is no agreement in the literature of what should be considered a novel formation and when the word loses its novelty (see for example Bauer 2001, Plag 1999). Secondly, the impact of the individual language user in many morphological studies tends to be perceived as marginal. Thirdly, for the most part the initial stages of word existence are difficult, if not impossible, to observe, and as such, the coining process itself, and the accentuation levels (that is pre-institutionalisation levels) often remain neglected due to the lack of tools. Of course, it is very difficult to observe the actual process of word coinage unless we talk about elicitation tasks where it is the purpose of the elicitation exercise. At the same time the observation of the reception and use of a new (or infrequent) word by a small group of people is also very difficult.

The work reported here is part of a larger project, which aims at observing morpho-lexical language change in progress, that is, the initial levels of word coinage and spread in the domain of computer mediated communication (CMC). In this paper I will discuss two integral and contributing factors in the interpretation of the novel word: (a) the issue of speaker acceptability, and (b) the anticipated unfamiliarity of the potential reader (or listener) with the new formation. This discussion is based on the observation of the language of individual users as the "individual speaker is the central factor with regard to all linguistic phenomena" (Koefoed and van Marle 2000: 311). In morphological studies, preference should be given to "naturally occurring data" in order to avoid the observers' paradox. This is where CMC comes in. As stated by Herring (1996: 5), "[...] interactions come already entered as text on a computer; [...] observers can observe without their presence being known, thus avoiding the 'Observer's Paradox' that has traditionally plagued research in the social sciences." For morpho-lexical research, data offered by CMC seems an ideal departure point. I will first explain the relevance

of the data source. Then I will briefly outline the project and the methodology behind it. Finally I will look at the acceptance levels and the ways they manifest themselves in my corpus.

2. The data

The data in this study comes from a corpus of blog samples. Blogs are the most recent type of CMC. Language purists often frown at the use of 'nternet language'as data source perceiving it as corrupt. In fact CMC uses many non-standard features, which are often misinterpreted as errors (see Runkehl et. al. 1998, Herring 2001, Shortis 2001). Rather than sloppiness and inattentions they are motivated by the economy of effort or mimicking the oral language. Blogs are a particularly special type of CMC, due to their unique structure, their popularity and a new way of information, and what follows language, dissemination.

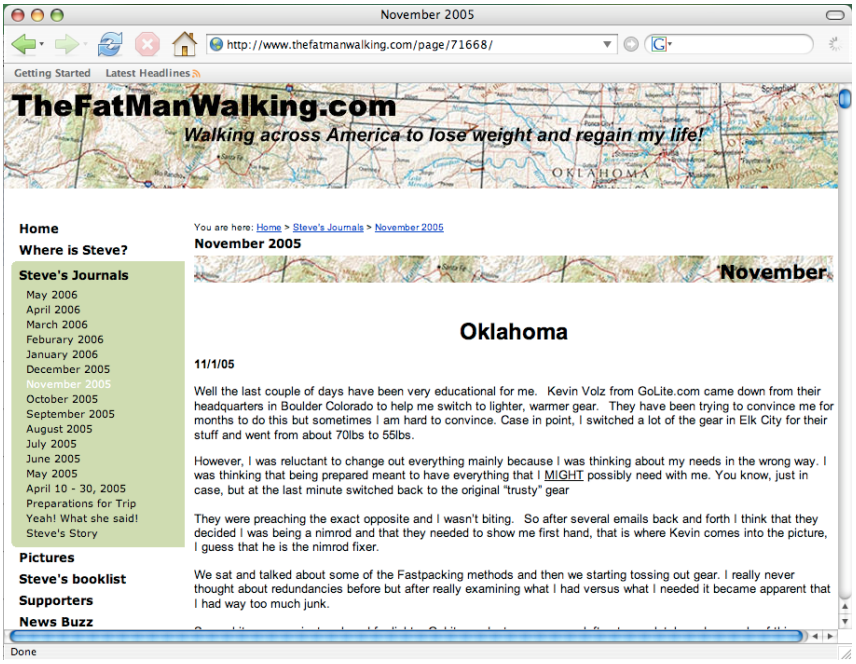

Figure 1 *Sample blog*

A blog is a website consisting of dated entries usually arranged in reverse chronological order (see Figure 1). The term *blog* (a clipped version of weblog) originates from a compound form: website log, that is a log, or diary, on the web. Blogs first appeared in the early nineties and gained popularity around the turn of the 21st century. A blog is characterised by the way information in it is organized, independent of the topic. In the words of Hourihan (2002), "[w]hat we write about does not define us as bloggers; it's how we write about it (frequently, ad nauseam, peppered with links. (…) As with free speech itself, what we say isn't as important as the system that enables us to say it." Blogs revolutionized content creation, giving the public broadcast tools to the hands of the private users and making them content creators. It is as if

anybody can run their own radio and/or TV station which is moreover available licence-free to any audience in the world who has the internet access. The most popular blogs attract over a million visits a day.[1] If we think about the language used in blogs – the impact they make should not then be underestimated.

3. Methodology

I have compiled a corpus of 100 English language blogs selected arbitrarily. Each blog sample has 10,000 words. So, the corpus currently amounts to one million words of running text, covering the period from 2000-2005.

In a one million-word corpus looking for *hapax legomena* is not a suitable approach. In my corpus over 50 % of the words occur only once. Instead, in order to facilitate analysis, a software tool, named INDIANA (INternet DIctionary ANAlyser), has been developed to extract potential neologisms from input files. It combines a text converter, a cumulative database and a series of online and offline filters. In brief, the steps from the input blog to the list of potential neologisms are as follows: The HTML content of the blog is first converted into plain text. Each new input file of text is converted into a sorted list of words. INDIANA first checks every word from the list against the existing database. If the word is not already present in the database, it is checked against two external reference sources: the data of the British National Corpus (BNC) and the Webster online dictionary (M-W).[2] If a match is found in one of these reference sources, this information is added to the database. Otherwise the word is marked as a potential neologism. Proper names and misspelled forms are manually deleted from the set of potential neologisms. The words remaining in list are verified and classified, again manually. INDIANA also includes a variety of filters which enable not only quick extraction of potentially new words but also offer information on type/token frequency, distribution across the input files, and easy view of each word in context. Additionally distribution information of every text file, grouped as information about the author (when available), text, linguistic features etc. is encoded. This enables detailed cross-analysis using any combination of filters. The result is a list of novel formations, with context information about their use, and information about the user. The writers often give clear clues, using various lexical, punctuation, syntactic or register strategies, as to whether these words form part of their lexicon. They also make predictions of whether the word familiarity will be shared by the blog readers.

As you can see, in the context of this project, a novel formation is the word-form that was not recorded in the material from BNC corpus (100 million words) or found in the Merriam-Webster dictionary online version. This of course means that the words classified as potential novel formations might as well be simply infrequent words, at least not frequent enough to make their way into the large BNC corpus or the dictionary at the time of the analysis. This might be seen as a drawback. However, it does not affect the research methodology in a serious way. Kjellmer (2000: 207) observes that "[w]hen a new word emerges and becomes accepted as part of the common word-stock, it is frequently the case that it has previously existed in some remote corner of the language (…) Still, it is a "new word" to the public at large." I will also be unable

to observe the semantic development of the form that already existed unless it underwent a graphemic change.

So far this method helps me filter out the words which, in the big picture, are new but it does not tell us much (if anything at all) about the individual users. Having access to the context in which the novel words are used however, I can observe acceptance patterns at the level of the individual speaker.

4. Acceptance manifestations

The life of a word can be seen as a cline of the levels of acceptability a new word goes through. Crystal (1995: 132) argues that "a neologism stays new until people start to use it without thinking, or alternatively until it falls out of fashion, and they stop using it altogether." It is important to point out that no distinction is made here between 'new' in the sense *newly formed* and 'new' as in *newly used/spread*, that is, every new encounter of the use of a word is treated the same way as its actual coinage.

When discussing the novelty of a word, we can have two different levels of analysis: general language observation, often based on quantitative institutionalisation principle (for example, whether the word is listed in a dictionary), and individual acceptance and dissemination patterns seen as manifestation of familiarity or with the word. As blogs focus on communication and information sharing, and, in this respect, are similar to public discourse, the second analysis level extends to the anticipated (un)familiarity with the word of the potential reader/listener. Discussing novel formations we, therefore, have three important parameters to consider: word, speaker/author, speaker/reader. In other words, we have to ask the following questions:

- Is the word at the pre-institutionalisation level?
- Is the word part of the speaker/writer's lexicon
- Is the word expected to be part of the speaker/reader's lexicon?

In my data I have looked at the acceptance patterns of words already filtered by Indiana as neologisms, that is, words at pre-institutionalisation level. For each such word, I compare the level of general linguistic acceptance/novelty with the individual one. Specifically, I compared the quantitative information based on the general language observation and the qualitative information based on the individual user's attitude to lexical items. To illustrate the different levels of acceptability, and the strategies used to signal them, I have selected the following pre-institutionalised words[3] from my corpus: *antihype, fisk, down-ness, destressedness, dragonology, dyndns, blogaholic,e-promos, meatspace, celebutante, blogorrhoea.*

4.1 *Author's lexicon*

Let us first look from the blog author's perspective, that is, is the word under consideration part of the author's lexicon? We may assume that if a word is already accepted from the point of view of the speaker, it will not be marked in any way. In other words, it has become part of author's language. Examples (1) and (2), illustrate this case:

(1) antihype:

"Stowe Boyd has posted an interesting True Voice show about spreading blog antihype. Stowe recorded interviews during Les Blogs and I'm in the show with Darren Barefoot, Doc Searls and Lee Bryant. Interesting stuff."

(2) fisk:

"Jones is also fisked by Jarvis and Ernie Miller."

In both cases no distancing strategies are used and the novel words are not marked in any way. It is important to add that, at first glance the use of *fisk* and *antihype* do not seem to follow the same acceptance pattern. This is evident if we look at the text in the original layout. The word *fisk* is additionally hyperlinked (see figure 2). Yet when we follow the hyperlink, instead of finding a definition or the explanation we go to the actual webpage where the whole action of John being fisked by Jarvis and Ernie Miller takes place.

Figure 2 *Acceptance pattern for 'fisk'*

In my data the lack of explicit marking, quite unsurprisingly, is rather frequent. It commonly applies to cases which are morphologically and semantically more transparent (like in example (1)).

The author can introduce various distancing-strategies to mark the word as either new to his lexicon, or as one he has not accepted (yet), or both. These strategies can be grouped and labelled as *implicit* and *explicit* strategies. Implicit strategies are those where the marking is relatively subtle. The implicit strategy used by one author is, for example, basing new words on pattern repetitions or using lexical prompting (that is, a word in a pre-derived/base form appearing first). Examples (3) and (4) illustrate these cases, respectively.

(3) down-ness:

> "This is a belated post due to the busy-ness of the schedule and the down-ness of the blogger."

(4) destressedness:

> "This day, so far, has been working incredibly hard to destress me. And so far, it's worked. However, the 5:30 - 7 Carmina rehearsal back to back with 7 - 8:30 play practice is forcefully working to undo earlier destressedness."

One of the explicit strategies is the use of hedges. These are metalinguistic comments that, in my data, manifest themselves in various ways. The author may simply and directly state that a concept is new, such as "new field" in example (5). Another strategy is to use impersonal distancing as in *sometimes labelled as* in example (6), and "what is referred to within certain circles" in example (7). The author can also personalize the hedge by assigning the coining to a particular person as in *he coins* in example (8). The author can also directly negatively assess the word coined by somebody else. In (9) the author not only writes: "I never heard the word" but also adds the he prefers a different term, and calls the new formation *icky*.

(5) dragonology:

> "New field of study: Dragonology "

(6) dyndns:

> "Dynamic DNS is sometimes labeled as **dyndns**. This is fun."

(7) blogaholic:

> "Maybe someone should recommend a stay at a blog addiction rehab center like the Central Vermont Blog Addiction Rehab Treatment Center, which just recently had their Website put up for them.
> opps, meant to write:
> ..., he still could be *what* is referred to within certain circles as a dry blogaholic."

(8) e-promos:

> "The writer writes about a phenomena, he coins (?) e-promos"

(9) meatspace:

> "Interesting terminology in that article. I've never heard the word "meatspace", before. I prefer the acronym "RL" (Real Life). "Meatspace" sounds ... kind of icky."

Explicit strategies often combine several ways of marking a word. In example (5) the hedge is accompanied by a hyperlink that points to a definition and illustration page. In example (6) the hedge is followed by the new word highlighted in bold print. In fact, the distance can also

be introduced by punctuation or any set of typographical devices used in CMC such as font-type, colour, size, spacing, and so on. Hedges can be combined with typographical devices as in example (10), where, apart from the bold print and inverted commas, the new word is hedged by "here's an even more annoying one" and "Fox coined it".

(10) celebutante:

> "here's an even more annoying one: "**celebutantes,**" to describe wealthy young women who are famous for being wealthy and young. (Fox coined it for their show "The Simple Life," in which hotel heiress Paris Hilton and Lionel Richie's improbably blonde daughter Nicole take the incredibly daring step of living on [gasp!] a farm for several weeks.)"

4.2 *Anticipated reader's lexicon*

Language use in CMC is predominantly interactive. In the blogging environment, where the author is producing public broadcast and where the life of a blog largely depends on readership, the author might take certain specific measures to negotiate the common ground and reduce the processing effort and miscommunication. It is even more understandable if we recall that similarly to radio/TV broadcasts or newspapers, the blogger has extremely limited information about the audience, so in order to communicate efficiently he might make extra effort to establish common ground. The author might use strategies already well established in journalism or literature.

We can imagine following situations. If the word is not new to the speaker, or if he himself coined the word, he may then leave it unmarked (examples (1) and (2)) or mark it using implicit strategies (examples (3) and (4)). Example (11) is the extreme case where the dictionary-like definition, including a part-of-speech (POS) information, is given for the word newly coined by the speaker before the word is actually used in context.

(11) blogorrhoea:

> Blogorrhoea (n): Psychic condition occasioned by global condition, occasioning bouts of public whimpering and fulmination.

If the word is new, very often the role of the speaker and reader overlap, in the sense that if a word is unknown to the speaker he might assume it will also be new to the reader, and so he can use explicit strategies not only to mark the distance but also to help the decoding process of the reader. This additional help can come in a variety of forms, such as an explanation (10), providing the origin of the expression (6), or a direct translation when a foreign word is used.

5. Conclusion

In this paper, I have demonstrated one of the contributing factors in the interpretation of the novel word, namely, the analysis of new formations at the level of an individual speaker. I

looked at the interpretation of the novel word from speaker's acceptability perspective, and the expected unfamiliarity of the potential reader with the new formation.

Quantitatively and qualitatively all words filtered by Indiana are neologisms (in the sense discussed before). Comparing the general linguistic level with the individual one, the unsurprising observation is that most of the cases of novel formations in my blog corpus show high levels of acceptability. In other words, they are not marked in any way, which directly supports Kjellmer's observations mentioned above. In the remaining words various implicit and explicit marking strategies are used.

The concept of newness is a very complex one. The discussion of novel words usually focuses on their morphological and semantic properties at the institutionalisation level. As a result the role of an individual is often marginalized. Hohenhaus (this issue) says that the "reception of real words by real speech communities in the real word – remains largely unobservable directly". In my project, with the corpus I compiled and the tools I use, I can overcome these shortcomings. The linguistic relevance of this study is that it combines the research done on the macro and micro levels using the help of the modern analysis tools.

Notes

[1] For a blog tracking tool see e.g.http://truthlaidbear.com/TrafficRanking.php

[2] The correctness of the selection made by my computerised tool is further supported by the language tools in the text editor that highlight all these words suggesting alternative spellings, or simply different words altogether. For example *antitype* for *antihype* or *distractedness* for *destressedness*. It simply shows that these words are not (yet) included in the existing wordlist in this particular word editor.

[3] The correctness of the selection made by my computerised tool is further supported by the language tools in the text editor that highlight all these words suggesting alternative spellings, or simply different words altogether. For example *antitype* for *antihype* or *distractedness* for *destressedness*. It simply shows that these words are not (yet) included in the existing wordlist in this particular word editor.

References

AITCHISON, Jean. 2001. *Language change: Progress or decay?* 3rd edition. Cambridge, New York, Melbourne: Cambridge University Press.

BAUER, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

CRYSTAL, David. 1995. *The Cambridge Encyclopedia of the English language*. Cambridge: Cambridge University Press.

HERRING, Susan C. (ED.) 1996. *Computer-Mediated Communication - Linguistic, Social and Cross-Cultural Perspectives*, Amsterdam, Philadelphia (Pragmatics & Beyond. New Series 39)

HERRING, Susan C. 2001. Computer-mediated discourse. In Schiffrin, D., Tannen, D. and Hamilton H. (EDs.) *The Handbook of Discourse Analysis* pp. 612-634.

HOHENHAUS, Peter. (this issue) Bouncebackability – a Web-as-corpus-based case study of a new formation, its interpretation, generalization, spread and subsequent decline.

HOURIHAN, Meg. 2002. What We're Doing When We Blog. [online] Available at: http://www.oreillynet.com/pub/a/javascript/2002/06/13/megnut.html

KJELLMER, Göran. 2000. Potential Words. In *Word* 51:2, pp. 205-228.

KOEFOED, Geert and VAN MARLE, Jaap. 2000. Productivity. In BOOIJ, G., LEHMANN, C., MUGDAN, J. (eds.) *An International Handbook on Inflection and Word-Formation*. Vol. I. Berlin: Walter de Gruyter, pp. 303-311.

PLAG, Ingo. 1999. *Morphological Productivity. Structural Constraints in English Derivation*. Berlin/New York: Mouton de Gruyter.

RUNKEHL, Jens, SCHLOBINSKI, Peter and SIEVER, Torston 1998. *Sprache und Kommunikation im Internet*. Wiesbaden: Wissenschaftlicher Verlag.

SHORTIS, Tim. 2001. *The Language of ICT. Information and Communication Technology*. London: Routledge.

Dorota Smyk-Bhattacharjee
Englisches Seminar
University of Zürich
Switzerland
dsmyk@es.unizh.ch