ORTHOGRAPHY, STRUCTURE AND LEXICAL CHOICE AS IDENTITY MARKERS IN SOCIAL TAGGING ENVIRONMENTS

Emma Tonkin
UKOLN
University of Bath, UK

ABSTRACT

Lexical choice, idiosyncratic composition and orthography receive more interest in the area of linguistics than in community informatics. However, language use and the underlying referents are understood to be deeply embedded in the wider social context. Stylistic attributes of text can become a defining feature in environments such as the Web, on which text-based communication still predominates. Therefore, features of even a short text are potentially valuable functional markers for participants and researchers alike. This paper briefly examines the role of lexical choice – the choice of a term to convey a given meaning – and stylistic choices in orthography in a social informatics environment, taking as an exemplar a dataset from eight tagging services. A number of specific communities are examined from the perspective of specialist terminology employed by each group. Through these brief studies, a sense is gained of various parameters of language use, and potential approaches for further research in the area.

KEYWORDS

Social tagging, orthography, lexical choice, community informatics, multi-word expressions

1. INTRODUCTION

This paper examines selected features of variation in language use across a number of groups and social tagging systems, with the aim of selecting a number of practically useful and relevant features. There is a long history of analysis of language use in on-line communities, the most widely known examples being Herring's work on gender and CMC (Herring, 1996) and Crystal's research into the characteristics of language use on-line (Crystal, 2001). Haythornthwaite and Gruzd (2007) in particular describe a method of applying noun phrase analysis toward content analysis of online communities. We examine the possibilities across a range of collaborative tagging systems. Collaborative tagging, also described in certain implementations as social tagging or classification and social indexing, describes a range of online services that allow ordinary users to assign keywords, or tags, to items (Tonkin et al, 2008). It is, however, often criticised as 'noisy', which is to say that the metadata retrieved is often considered to be of low quality. This criticism has sparked discussion and research into typologies and classes of tag use, such as personal information management and joint resource description and discovery.

Many historical personalities have asserted that present use of language as less appropriate or nuanced than that of previous years. Deutscher (2005) describes individuals and organisations who, since Roman times, have found fault with the current vernacular used by speakers of their language, and have either bemoaned it or made active efforts to halt the decline of the language. Deutscher quotes Cicero who, in 46 B.C., compared the speech of public figures of the day with that of a century before, and concluded that ’practically everyone... in those days spoke correctly. But the lapse of time has certainly had a deteriorating effect in this respect’. We might instead follow the poet T. S. Eliot, who reflected that 'last year's words belong to last year's language and next year's words await another voice'.

Sociolinguistics and historical linguistics might regard variation in language as an indication of continuous adaption rather than a sign of the decay of culture or lack of literacy. We may portray language as one of a large class of emergent systems that, through a series of local interactions between agents, develops

an intricate structure. To seriously entertain this possibility that language as a flexible, moderately coherent structure is emergent from local interactions, it is necessary to describe a system that is sufficiently flexible for changes to occur, but sufficiently rigid that the system does not lose coherence entirely. Variation is necessary to ensure that the system continues to evolve. Analysis of systems where natural language is applied would therefore expect to encounter some level of variation.

In knowledge management (KM), Ogden and Richards (1923) described the *semiotic triangle* - a structure that links reference, referent and symbol, showing that the object which is referred to by a given symbol or word is not static, but relative to each language user.A well-known model is the *word-space* model (Sahlgren, 2006), in which words are placed on a high-dimensional vector spaces – the vector space being a classical model in information retrieval. There is not sufficient information to clearly reconstruct the *concept space* underlying the many instances of term use; the semiotic triangle is important in that it formalises the idea that the symbol and the referent are not equivalent – the concept to which the symbol points may be neither entirely reproducible nor entirely unique. The 'map' of memories or concepts, and the collection of perhaps 60,000 words (Crystal, 2003) by which these are accessed are in all likelihood unique and dynamic structures in their own right. Thus, communication is sometimes described as a process of negotiation (eg Krauss and Fussell, 1996).

In the following section we examine various examples of language use in the area of tagging in order to investigate various aspects of adaption or variation, broadly focusing on specific areas of subject, structure and orthography. We will then examine datasets taken from several tagging services in order to establish a rough idea of the variation in use and types of terms applied. In this area, variation is expected to be the rule, but the majority of social tagging studies focus on gaining a deep understanding of the usage of specific sites, rather than a broader and less detailed understanding of the wider landscape of usage patterns across platforms; in this case, the latter approach is taken in order to gain an overview of the differences between platforms and hence, in a broad sense, between user communities. The term 'community' is used here predominantly to refer to Swales' (1990) discourse communities - spatially dispersed, formed around sociorhetorical functions, and mainly mediated by text.

1.1 Word Clusters and Meaning

We have discussed the viewpoint that there exists a concept space that is not itself directly visible, but may be investigated indirectly through examination of the way in which symbols are applied. The simplest metric is a raw frequency count. Examining terms applied to data objects in the popular photo tagging system Flickr for geographical locations in the area of Tokyo, we find that the term 'Tokyo' and the Japanese equivalent appear with a similar relative frequency and in similar contexts.

Indirect examination is generally done via examining cooccurrence. This is a variant on the *distributional hypothesis* – that two terms applied similarly may have similar characteristics. If an approach such as PCA or SVD (two similar mathematical methods enabling identification and evaluation of the significant dimensions in a multidimensional space) are applied, it is possible to produce a reduced-dimensional visual representation of the system and examine a 2D representation visually:

Table 1. Raw frequency of top terms from a sample of geotagged images taken from Flickr

Tag	Relative frequency
Tokyo	**547**
Japan	**433**
Night	**91**
東京 (Tokyo)	**89**
日本 (Japan)	**74**

As can be seen from this image, there is – even in two dimensions – some evidence of clustering of similar terms. Terms relating to location are gathered together, as are nomenclature ('nihon' and 'nippon'), landmarks, types of photographs and light conditions ('dark' and 'dusk'). This is an effect resulting from co-occurrence; terms that appear in similar positions are expected to share some similarity in meaning or function. Of course, we would not expect to find a perfect segmentation over a two-dimensional

representation. Vector-space models in practice could be expected to operate across hundreds, if not thousands, of dimensions. We might imagine a few of these dimensions: word type (noun, verb...), tallness, warmth, the frequency with which a term is applied by a certain age group – but in practice, SVD dimensions typically have only mathematical relevance. The major difficulty with methods such as SVD is one of computational expense of calculation.

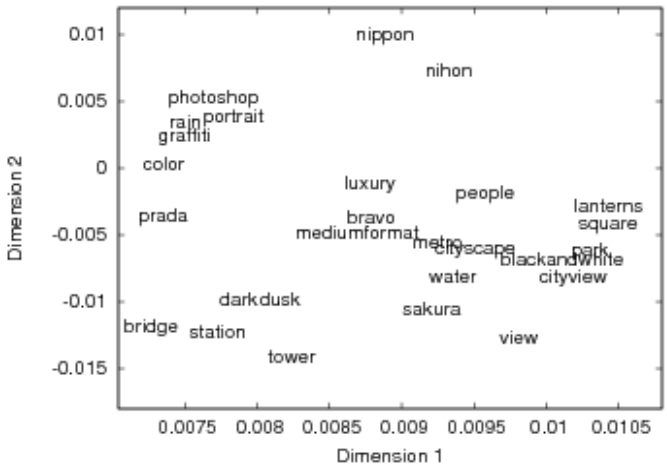

Illustration 1. Partial two-dimensional word space of tags applied to images of Tokyo

Considering a set of terms as lexical items overlaid across an underlying and relatively ambiguous set of concepts or ideas, what strategies can optimally communicate these concepts? The mathematics behind Illustration 1 is essentially deterministic across a given dataset, but in practice, each individual 'owns' a subtly different collection of concepts. The obvious motivation for lexical choice is clarity of communication, but there are others – particularly social factors. For example, it may relate to the construction of an identity (eg. Bucholtz, 1999) - the choice of a term is sometimes designed to express something about the identity of the speaker. A similar phenomenon has also arisen in online social network software. Public visibility of profiles leads users of Web sites such as Linked In and Facebook to manage the information, posts and users who are marked as 'friends'/'contacts' with caution, in the everyday course of managing a visible online identity. Frequency and cooccurrence are two easily calculable metrics, but there are many other characteristic features in written language. In the following section, we concentrate on three types of feature.

2. MEMES, SNOWCLONES AND THE MWE

As is so often the case in an interdisciplinary world, the least precise of these terms is perhaps the best known. The meme, of course, is the 'unit of cultural transmission, or a unit of imitation' first described by Richard Dawkins in *The Selfish Gene* (1976). The term has been widely adopted on the Web in order to describe a particular form of transmission of attractive ideas or activities, popular words or phrases and cultural references, and references to shared activities such as Web-based questionnaires or in-jokes; an overview of the Internet meme is presented by Knobel (2005).

We take as our example the loosely-knit collective of self-selected individuals known as 'Anonymous'. Defining identity through shared cultural reference, the participants do not in general provide real or even stable pseudonymous identities on-or-offline. This is at least in part due to the risks and political nature of their shared aim (activism). The group makes frequent use of internet memes on-and-offline, such as 'lolcats' - images of cats or other animals with an idiosyncratically misspelt caption, references to in-joke terms and resources such as the theme tune to *Portal*, a recent game by Valve Software. For a worldwide, distributed group of individuals to organise with such success, sharing a set of idiosyncratic linguistic habits and references, is a testament to the power of the meme.

Leaving behind the Internet meme, we move on to the phrasal template, dubbed 'snowclone' by Glen Whitman in 2004 (Pullam, 2004) - a 'some-assembly required adaptable cliché for lazy journalists'. His

examples include, “If Eskimos have N words for snow, X surely have Y words for Z.” and “In space, no one can hear you X.” Internet memes often appear in the form of compound phrases; examples taken from the popular blogging site Slashdot include the “In Soviet Russia” subject-object reversal meme - “In Soviet Russia, paper writes *you*!” - “I for one welcome our [subject area] overlords”, and “Pirates of the [topic]”. Lolcat slang contains its own stock phrasal templates, such as “Im in ur ..., ...ing yr ....” - for example, “Im in ur dictionaries, verbing yr nouns”. However, the definition of 'snowclone' includes the stipulation that the use of the compound should contribute a novel meaning to the phrase. In the majority of cases, stock phrasal constructions taken from internet memes do not do so, although the second example taken from Slashdot perhaps comes closest.

Finally, we come to the multi-word expression. MWEs are cohesive lexemes that cross word boundaries. Many different types of construction may be referred to as MWEs; compound nouns, verbal idioms, verb-particle constructions and so forth. According to (Baldwin et al, 2003), MWEs can be typified into three categories; in particular, these are non-decomposable, idiosyncratically decomposable and simple decomposable. Non-decomposable MWEs are 'semantically impenetrable', which is to say that almost no alterations can be made to the expression if the meaning is to be preserved. Idiosyncratically decomposable MWEs permit a certain amount of syntactic variation, but the terms within the MWE nonetheless take on a different meaning once they are decomposed. Simple decomposable MWEs display high syntactic variability whilst retaining the same meaning. An example of the first is "kicking the bucket"; of the second "spilling the beans", and of the third, "traffic light". An MWE is distinguished from ordinary functional combination of words by the fact that where an MWE exists, it is in general the preferred combination. Where no MWE is available, alternative terms may be used that carry the same semantics. For example, 'little boy' and 'small boy' are similar. However, use of the term "stop light" is not freely interchangable with the expression "halt light".

These constructions are of great relevance to the study of tagging, despite the fact that numerically, the majority of tags assigned to resources on most websites appear to be single-word expressions. Proportions found during this study (see section 2.3), range from 98% of unique expressions on YouTube to approximately half of all tags on Last.fm. Some terms are 'compound' in the sense that they contain more than one unit of meaning, such as an author's name and a date, or perhaps a landmark's name and location. Others are 'atomic' -- that is, they are full lexical chunks. However, it is observed that tags are often 'less-than-atomic' -- that is, they do not contain sufficient information to decode their meaning without reference to the terms with which they co-occur. Such terms are in many cases components in some form of MWE.

2.1 Orthographic Variation

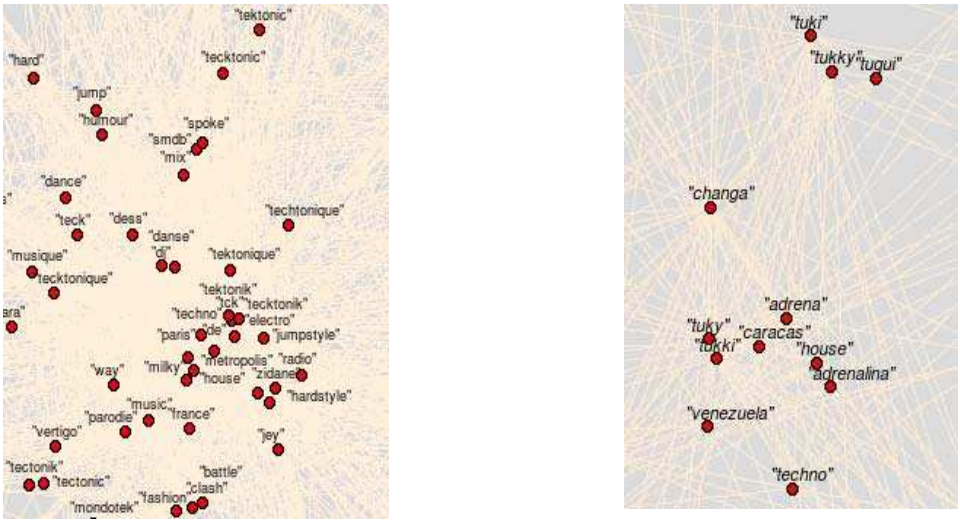

Illustration 2. A partial landscape of terms applied to 'Tecktonic Killer' and ' Changa Tuki' videos on YouTube

Moulaison (2007) recently examined the Techtonik Killer dance phenomenon, presently popular in France. Tecktonic Killer (or Tektonik, Tektonick, Techtonik...) is a movement presently visible on video-

sharing sites such as YouTube, in which young people dance to techno music and attempt to outperform their peers, with feedback from video comments and rating mechanisms. According to Moulaison, the tags applied to these videos vary widely in spelling. The area offers a fascinating example of the creation and use of many variant terms, and could be described as an example of a battleground in which selection between variant terms is taking place. The name 'Tecktonik' is summarised by one blogger, who discourages the use of the term, as 'le nom d'une organisation d'événementiel! Une marque quoi...' - 'the name of an event organiser – a trademark'. This may be one contributing factor for the variation in spelling – to arrive at a term that is owned by the community – but another explanation arises from an observation by Tavosanis (2007) that deviations from standard orthography are often actively sought in online communication.

There is undoubtedly considerable variation in the use of the term – too much to lay blame at the feet of poor spelling (competence error). Taking a dataset from the tags applied on YouTube videos and organising them in order of stress centrality, that is, identifying nodes that are in the set of 'shortest paths' between nodes in the graph, we see a large number of variations of the term and its abbreviations. Illustration 2 was produced in this manner, using the open-source social-network-analysis software tool *socnetv*. Compare to the variations in spelling of the Venezualan youth dance phenomenon Changa Tuki (Illustration 2, right). Several variants are visible. Particularly with the Tecktonic variants, left, it is difficult to dismiss these as competence faults. Variation is systematic within and between taggers. Instead, this can be seen as intentional deviation for stylistic purposes, as described by Tavosanis (2007), perhaps influenced by linguistic interference – that is, systematic misspelling as a result of the application of standards and customs from the writer's native language.

2.2 Variation Across Communities

For this study, datasets were taken from eight tagging services. The experimental method necessarily differs slightly between tagging systems as level of access, available APIs etc differ a great deal. A randomised sampling of terms, applied to randomly selected data objects, were harvested. The precise means by which tags are gathered and randomised depended on the available interfaces and structure of each site, ranging from the use of provided sample data to data extracted via a purpose-built web spider. 100 component tags from each sample are classified according to several metrics; tag length in words and characters, tag structure according to part-of-speech tagging of component elements, and the status of each tag as an example of a performance tag. The tagging services that were examined in this manner included Amazon, CiteULike, Connotea, Del.icio.us, last.fm, Panoramio, Slashdot and YouTube.

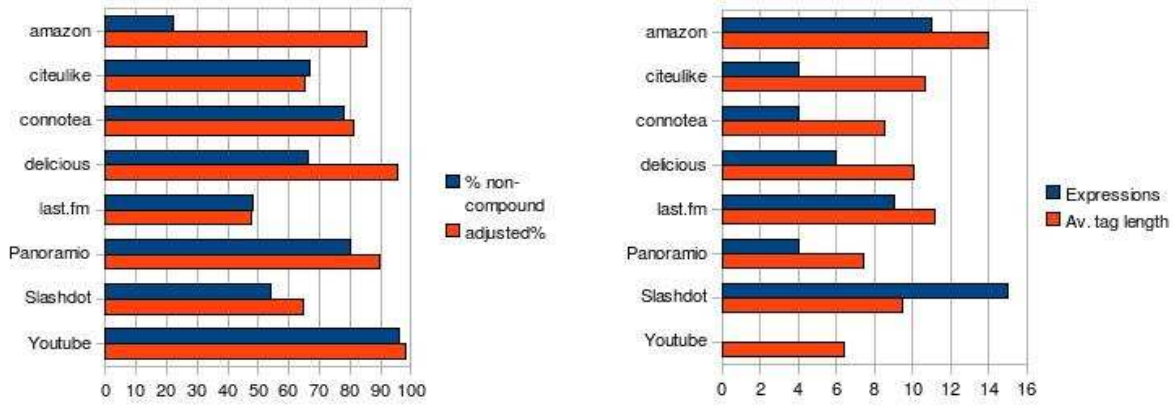

Illustration 3. Appearance of singular tags as a proportion and adjusted for estimated tag frequency. Right: freq. of recognised compound expressions, average tag length

The results show considerable variation in the use of tags in general, both within and between systems. Noun phrases appear more commonly on academic sites than on popular culture sites; the converse is true of adjectives. Occurrence of 'performance tags' (Zollers, 2007) range from 0% in the case of Panoramio to 47% in the case of Slashdot. Variation in appearance of performance tags, and indeed tag syntax in general, raises further questions about the factors influencing tag choice and structure. Sites that publicly display tags

in proximity to content such as product or artist descriptions tend to receive more performance tags, suggesting a socially-motivated pattern of use. The existence of tag sharing/browsing facilities alone does not have this effect. However, it is also clear that a great deal of variation is owed to peculiarities of the interface; some systems are presented as solely a bookmarking system. Others display the tags in such a way that they instead appear to be a space for brief phrases or comments destined for public viewing. In particular, Slashdot provides an example of this type of behaviour.

This study also demonstrated a difficulty with the methodology of the investigation; the 'less-than-atomic' nature of many tags on certain sites, like YouTube, means that a given set of tags typically contain more than one expression. In other words, the space set aside for tags is commonly used for short comments, plus tags, meaning that another approach must be taken to decompose the set. For example, the following example 'phrase' from YouTube decomposes into short phrases as follows: *(new york) (anime festival) 2007 fandomania animation cosplay costumes interview (book characters)* This is also somewhat true of CiteULike tags; many multi-word tags are not full expressions at all, but multiple separate terms, such as 'Advertisements Solar'. It is possible that this is deliberate, to 'fool' automatic indexing (as described by Tavosanis, 2007). Detecting these reduces to the known problem of MWE identification and extraction.

Perhaps the most interesting finding is the level of variation in the type of characteristic phrase. For 'social sites' such as Slashdot or Last.fm, types of phrase found are frequently 'performance tags', idiosyncratic uses of language, cultural references or in-jokes. Phrases found in academic or scholarly sites are more frequently noun phrases. This raises the possibility that the approach described by Haythornthwaite et al, which focuses specifically on noun phrase extraction, could be adapted and applied across a wider range of texts and types of term.

3. CONCLUSION

The examination of linguistic peculiarities is an extraordinarily addictive way of passing time, but also constitutes a powerful means of exploring how members of various communities choose to interact – and to limit interaction. There exist many excellent and influential studies of the characteristics of given groups, but it is relatively uncommon for details such as syntactic construction to be given a great deal of explicit attention in the wider area of community informatics, although this information is, in practice, already tacitly applied in some cases.

For system designers, it is often the case that such variation is seen as a source of inconsistency in the system, to be eradicated where possible. For users intent on leveraging secondary uses of the communication channel – such as expression of identity and self-identification as a community member – the attempt to engineer out this variation simply reduces the value of the channel. Furthermore, the level of variation seen suggests that continuing the trend in community informatics towards close investigation of the markers present in language use, and taking note of the applicable research taking place in linguistics linguistics in these areas, would be as valuable to the field of social network analysis as to interaction design.

Further research focusing on a deeper analysis of any or all of these categories of variation may be valuable to those examining the propagation of ideas, memes or references. The general approach is also of use to those with an interest in examining the dynamics of community membership. Additionally and perhaps most prosaically, teaming the concepts discussed here with techniques taken from machine learning and pattern recognition may permit a better segmentation of tag information, potentially improving the precision of information retrieval as well as the visual presentation of information taken from user-submitted tags. Finally, a better understanding of the wide range of variation seen in use of tagging systems is advantageous in developing systems that are satisfactory to both the system designer and the user. Recognising the causes and consequences of unexpected patterns of use enables the designer to respond appropriately if necessary or advisable (for example, if the variation is a symptom of a flawed design), and to work more effectively with the resulting dataset.

ACKNOWLEDGEMENTS

The author would like to thank Heather Lea Moulaison, Edward Corrado and Alla Zollers.

REFERENCES

Bucholtz, M., 1999. Why be normal?: Language and identity practices in a community of nerd girls. *Language in Society* 28.2: 203-223.

Crystal, D., 2001. *Language and the Internet*. Cambridge, UK: Cambridge University Press.

Crystal, D., 2003. *The Cambridge Encyclopedia of the English Language*. Cambridge University Press, Cambridge, second edition.

Dawkins, R.. 1976. *The Selfish Gene*. Memes:the new replicators, Oxford University, 1976.

Deutscher, G., 2005. *The Unfolding of Language*. Arrow Books.

Haythornthwaite, C. and Gruzd, A., 2007. *A noun phrase analysis tool for mining online community conversations*. In Steinfield, Pentland, Ackerman, and Contractor (eds.), Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference, Michigan State University, 2007, London: Springer, 67-86.

Herring, S.C., 1994. Gender differences in computer-mediated communication: Bringing familiar baggage to the new frontier. American Library Association convention, Miami, FL.

Knobel, M., 2005. Memes, Literacy and affinity spaces: implications for policy and digital divides in education. Policy options and models for bridging digital divides. March 14-15.

Krauss, R. M. and Fussell, S. R., 1996. *Constructing Shared Communicative Environments*, chapter 9, pages 172–202. American Psychological Association: Washington, DC, 3rd edition.

Moulaison, H. L., 2007. Social tagging in France: The evolution of a phenomenon. In panel: Corrado, E. M. et al. Tagging and social networks: The impact of communities on user centered tagging. ASIS&T 2007 Annual Conference. Milwaukee, Wisconsin.

Ogden, C. K. and I. A. Richards, I. A., 1923. *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism.* London: Routledge & Kegan Paul.

Pullam, Geoffrey K. 2004. Snowclones: Lexicographical dating to the second. *Language Log.* Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/000350.html

Sahlgren, M., 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD thesis, Stockholm University.

Swales, J. M., 1990. *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tavosanis, M., 2007. A causal classification of orthography errors in web texts. *AND 2007.*

Tonkin, E.L., Corrado, E. M., Moulaison, H. L., Kipp, M.E.I., Resmini, A., Pfeiffer, H. D. and Zhang, Q., 2008. Collaborative and Social Tagging Networks. Ariadne 54 (1).