Following Conversational Traces

Part I: Creating a corpus with the ICWSM dataset

Stephanie Hendrick
Umeå University, Department of
Modern languages and HUMlab
90732 Umeå, Sweden
stephanie.hendrick@humlab.umu.se

Abstract

This poster will present the methodology behind the creation of a linguistic corpus based on a subset of the 2007 International Conference on Weblogs and Social Media dataset. Posts from a small group of political bloggers were tagged for parts of speech and indexed into a corpus using the program Xairia. From this corpus, the political blogger subset will be investigated for register and referential information. Referential information, especially with regards to new and given information, will be compared against network placement both to identify network innovators as well as to compare network placement as a catalyst for innovation. The final section, Further Research, will outline the modifications necessary for the creation of a full-scale corpus based on the entire ICWSM 2006 dataset (currently in progress).

General Terms

Documentation, Performance, Design, Experimentation, Languages, Theory

Keywords

Corpus building, XML indexing, Xaira, CLAWS POS tagging, sociolinguistics, ICWSM 2006 weblog dataset.

1. Introduction

'We will leave here today and our language will have changed by the interaction that has taken place.' –Nev Shrimpton, Krio Linguistic Seminar, Umeå University – December 1st, 2006.

The quote above came from the closing arguments of the December 1st 2006 linguistic seminar on Krio at Umeå University. While Nev was exaggerating the extent of language change in order to emphasize his point, there is also truth to his statement. Communication is a constant state of negotiation, and language in continuous flux. Those with whom we come into contact modify our language. We speak a certain way with a certain group, and even with ourselves. And while that thought in itself is interesting, even more so is how and to what degree we vary our communications. Do bloggers vary their language according to perceived readership - the 'invisible readers' versus speculative social networks?

Blogs are an exceptional object of research to answer this question of 'how' because of their social nature. They form weak social networks with dense, fluid, clusters [4]. The clustering/small world effects that occur in blog networks (sometimes called the echochamber effect) allow us to look for potential variation for both perceived general audience and perceived social network. But exactly how do we answer the 'how'? Linguistically, from the merger of three methodological frameworks: *Social Network Analysis, corpus investigation, and sociolinguistic relations.*

Social network analysis: Where are people positioned in their network? How fluid are those positions? How often (if at all) do their interact with members of other networks?

Corpus Linguistics: Do different social networks use different registers, and if so, what differences/similarities do their communicative functions have? Are some networks more speech like than others? Are some more matter-of-fact, some more questioning? Where do they fall on the continuum of speech and writing? Does this differ between the different types of weblog networks?

Sociolinguistic relationships: How do network positions relate to language maintenance and variation (is there a relationship between the fluidity of placement and variation?)? What about other social variables? Does 'real-world position' (i.e. professor rather than a grad student in an academic network) make a difference? Gender? Geography?

It is in the area linguistic terra incognito - where sociolinguistic factors (gender, socioeconomic status, etc.) are different than in real-world situations - that corpus linguistics can become an important tool in identifying and explaining variation between communities. In other words, the ICWSM dataset can provide a valuable opportunity to create corpus which can help explain why we blog the way we do and if our network placement influences our choice of topics or formality of language.

2. A pilot corpus

As this dataset is so large, it was decided to test a small subset, day 1, before creating a corpus using the dataset in its entirety. Day 1's list of links generated 244, 988 nodes (weblogs) and 93, 284 edges (relationships). This file was still so large, however, that it was quite difficult to visualize (see Figure 1). In order to visualize stronger relationships, we mined the links for stronger ties. Rather than looking at all blogs who have an outbound link, we coded for those who have a reciprocal link – those who link back to each other at least one time. In this set, we generated a list with 1.362 nodes and 3174 edges (see Figure 2).

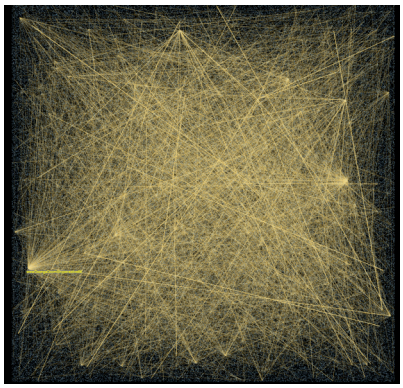

Figure 1: *Day 1 of ICWSM Dataset - whole. (*GUESS -Random*)*

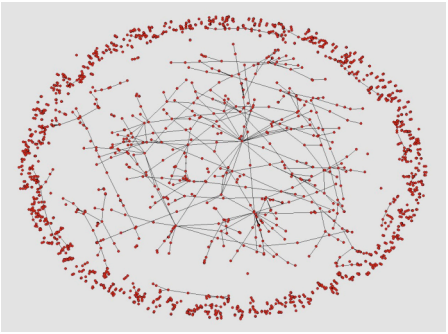

Figure 2: *Day 1 of ICWSM dataset – (*Pajek – KK Layout*)*

The purpose of this pilot study is to create a methodological framework for identifying weblog register, as well as possible relationships between register and network formation. It will not suppose weblog register in general; rather register for this specific COP.

2.1 Representativeness and Tagging

The first decision made was how much of the sample group text to use. For the pilot study, it was decided to use the entire entrie(s) for each member as there may be a relationship between the type of noun modifiers used and the length of entry.

Next, the data had to be tagged. The first step in this process was to determine if the meta-data tagged in the XML files was of linguistic interest, as well as to determine which tags I would need in order to define community of practice (COP) register and any possible relationships between the spread of information and COP placement. It was decided that, for the pilot study, metadata (author, time, permalink, etc) was not necessary as each blog's entries were contained in a separate file and indexed together in the corpus program, Xaira.

Part of Speech tagging (POS tagging) was applied to each file using the CLAWS online version - the C5 tagset (see figure 3). XML tagging was used rather than traditional POS tagging as Xairia is XML based. The decision to use Xairia was based on potential use of existing metadata for the complete corpus.

```
<w id="17.17" pos="CJT">that</w> <w id="17.18" pos="PNP">he</w>
<w id="17.19" pos="VHZ">has</w> <w id="17.20" pos="VVN">run</w>
<w id="17.21" pos="NN1">afoul</w> <w id="17.22" pos="PRF">of</w>
<w id="17.23" pos="AT0">the</w> <w id="17.24" pos="NN1">regime</w>
<w id="17.25" pos=".">.</w>
```

Figure 3: *Example for text input, tagged for POS*

3. Further research

There are other important and interesting aspects to examine when using corpus methods on weblog data. Studying variation within communities of practice and against that community's network formation allows the researcher to examine language, not only in its social context, but also in relation to its domain of use. Maurizio Gotti, in relation to his work on using corpora to study the effects of globalization in specialized discourse states that 'The process of internationalization of English offers a topical illustration of the interaction between linguistic and cultural factors in the construction of discourse, both within specialized domains and in wider contexts. This process is most evident in domains of use… where the socialization/textualisation of knowledge plays a crucial cohesive role [3]. Using corpora in combination with social network analysis of the COP allows the researcher a glimpse of the socialization/textualisation relationship.

The first step in making use of this data is to complete the pilot study. Pronouns, nouns, and noun phrases will be examined in order to measure referring expressions. Considering the difficulty of following discourse over different weblogs, establishing patterns here can be of interest. The informational/interpersonal focus of the texts can also be accessed through the syntactic and semantic characteristics of the noun phrases.

After completion of the pilot study, the variables will be reassessed and applied to COP's generated from analysis of outbound links. These studies will be compared across weblog COP's in order to identify possible generalizations about the 'relationship between network placement and variation' [5].

This poster is meant to serve as a brief description of the methodology/thought process behind the creation of a linguistic, weblog corpus. Current information about the progress of this project can be found on the author's weblog, sumofmyparts.org.

Acknowledgements

My sincere thanks to Johan Lindskog of HUMlab for his time and coding abilities.

References

[1] Bergs, A. Analyzing online communication from a social network point of view: questions, problems, perspectives. http://icame.uib.no/ij24/atwell.pdf

[2] Eckert, P. Linguistic Variation as Social Practice. Oxford:Blackwell, 2000.

[3] Gotti, M. Creating a corpus for the analysis of identity traits in English specialized discourse. The European English Messenger, 15.2. 2006.

[4] Efimova, L. Hendrick, S. & Anjewierden, A. In search for a virtual settlement: An exploration of weblog community boundaries. https://doc.telin.nl/dscgi/ds.py/Get/File-46041 . 2004.

[5] Ide, N. & Romary, L. XML Support for Annotated Language Resources.http://www.ldc.upenn.edu/exploration/expl2000/papers/ide/ide.pdf (2000).