Representativeness in Corpus Design

|  |  |
| --- | --- |
| First Author | Second Author |
| Affiliation | Affiliation |
| email@domain | email@domain |

Introduction

Conventional corpus building efforts have focused primarily on sampling language as an abstract, distinct and persistent entity. Though the sampling methods first used in the Brown/LOB corpus family have yielded many valuable results, and will doubtless continue to do so, this strategy has led to a number of pragmatic and scientific issues that continue to limit corpus linguistics.

Many of these issues have been framed in terms of improved description of language variation, most notably by Biber (1993) who assesses the variability within various language categories (e.g. preposition and relative clause usage) in order to determine the extent to which we may, probabilistically speaking, extract further useful meaning by sampling further. He goes on to conclude that we cannot know, *a priori*, the relative proportions of register usage within a language, and that we must go about sampling in an iterative manner, learning from our findings in order to improve further efforts. Although the iterative sampling method is widespread in other disciplines, it has not been discussed much within corpus linguistics, other than by Evert (ref to CL2009? paper).

Biber’s call for multi-phase sampling stems from a number of unknowns regarding language: we simply do not know what language is used on a daily basis, by whom, or for what purpose, with sufficient resolution to construct a demographically- and linguistically-balanced sample of language use with any certainty. Some of this uncertainty is, in part, down to the pragmatic decision to sample primarily in terms of language as a separate entity, making attempts to balance demographics by using proxy variables such as relative popularity of a text, formality and other socio-cultural external descriptors. As discussed in Leech (2006), a prime example of this confusion is the selection of language according to production or consumption, something only vaguely addressed by many corpora yet particularly important to the findings of many studies. Notably Leech calls for a re-inspection of proportionality, saying “the representation of texts should be proportional not only to their initiators, but also to their receivers”.

Since corpus studies are necessarily concerned with the social demographics underlying language (either directly or by assumption), there is a need to improve the quality of social and demographic information in corpora, and its relationship to real population distributions.

One way to examine these contextual and demographic markers is to construct a personal corpus—a record of all language used by a single person. This sample would represent a trade-off: accurately sampling text type proportions whilst sacrificing power in describing a wider population.

Sampling in this manner makes explicit many of the assumptions made during conventional corpus sampling, such as those involving variables with difficult-to-enumerate populations:

* Temporal information, such as the age of a document when used;
* The relative properties of language produced and consumed, and frequency of re-read material;
* Detailed demographic information surrounding those using the language;
* Proportionality of language categories used;
* Representative sampling of ephemeral and short texts (e.g. greetings, flyers, advertisements).

Such a sample may be seen as a first effort towards a model of language proportionality that is not based upon proxy variables such as setting, formality or plurality.

In this study, we describe the process of designing and building a personal corpus using techniques derived from life-logging. Ultimately, we aim to:

* develop a process of recording, digitising and coding multi-modal linguistic data in-the-field;
* identify proportions of language used for a single demographic, and compare these to existing corpora for an initial indication of possible improvements;
* develop methods for resampling corpora to align them in terms of a personal corpus, including relation of existing corpus genres to those commonly seen;
* develop a methodology that may be used to reweight corpora with less intrusion upon a given subject, through digital methods (such as logging) and conventional methods borrowed from the social sciences.

It is hoped we will be able to work towards an answer to some of the more persistent questions of corpus design: sample size, variability in real-world contexts, and stratum proportions, and, in so doing, assist those attempting to infer information from existing corpus resources by augmenting their meta-data.

Literature

One of the major challenges in capturing language use comprehensively is sampling small texts (such as greetings or brochures). These features are only sufficiently represented by verbatim recording methods, such as those used in the life-logging community. Though life-logging is commonly accomplished with video recording systems, making the data hard to summarise and digest without manual review, there have been some efforts to use textual and multi-modal input.

Many of these systems have been designed to augment human memories with verbatim accuracy: Microsoft's MyLifeBits project (Gemmell *et al*., 2002) uses, amongst other things, their SenseCam device (Hodges, S. *et al.* 2006) to incorporate documents and other ephemera, using their metadata to inform memories and records and integrating this with calendar software.

Combining the life-logging concept with that of learner corpora, Roy's work on the 'Human Speechome' project (2006) follows the linguistic development of his own son over a three year period. His methods centre around continuous video recording and automated processing.

Efforts in low-impact personal archives have produced methods for recording “high-level” details about spoken interaction (Lee and Ellis, 2006): the number of speakers, duration, length of interaction, etc. These methods, combined with those designed for other “real-world” tasks such as automotive voice recognition and mobilephone noise cancellation, may be used to reduce the load of processing long recordings.

Capture

One other primary consideration when sampling language is how to retain comparability with other corpora. This is especially challenging in the case where data sampled may fall beyond the bounds of existing taxonomies. Since the utility of a personal corpus is damaged by any disruption caused by its sampling methods, any metadata stored on texts must therefore be succinct, easily related to other corpora, and 'evocative' (such that a researcher may interpolate their data at the end of a day).

With this in mind, we have decided to sample genres, contexts and purposes using semi-free coding—the researcher will be aware of existing genre distinctions, but will insert others where they most accurately describe the use of the text.

Sampling methods have been selected to cover most common forms of interaction with language, with a focus on capturing those that may have 'slipped through the cracks' of other corpora. These include:

* Manual annotation, to record long or inconvenient texts (such as books) and metadata on the following;
* Continuous audio recording, to capture speech, radio and short interactions;
* Photography, for billboards and posters;
* Digital logging of web use, keyboard input, and online chat;
* Web storage for digital documents encountered elsewhere;
* Physical record-keeping.

Each of these presents a number of ethical and practical issues, which we will discuss in detail in the presentation.

Use

Though the statistical ideal would be to have many personal corpora, and combine them into a linguistically-balanced single sample, a single of these data-points yields useful information when combined with existing data.

We may use demographic information to locate the subject (or subjects) within the distribution of a conventional corpus, allowing us to relate the language proportions observed to a wider context with a high degree of accuracy. This is especially useful for establishing what text types are missing altogether.

Reversing this process allows us to re-sample existing corpora using the identified personal proportions of text types, in order to produce an augmented corpus that is balanced for a given demographic, yet large enough to train existing NLP systems. This process is also the key to easing the ethical concerns surrounding such invasive recording, as it makes possible the elicitation of language proportions through the use of questionnaires or irreversibly-hashed recording techniques, as proposed in (Ellis and Lee, 2004). This, in turn, would allow for the application of current general-purpose corpus data to targeted research questions.

Conclusion

We believe that the concept of a personal corpus forms a complementary sampling strategy to that of conventional corpora. The rise in digital language consumption, along with recent advances in speech recognition and portable device power, allow us to start the process of re-examining and improving our language resources in order to better align them both to the ground truth, achieving better levels of representation, and to the aims of linguistic studies, making them both easier to use and more fruitful.

References

Biber, D 1993 “Representativeness in corpus design” *Literary and linguistic computing* (8), p243-257

Ellis, D.P.W. And Lee, K. 2004 “Minimal-impact audio-based personal archives” *Proceedings of the 1st ACM workshop on Continuous archival and retrieval of personal experiences* p39-47

Gemmell, J. and Bell, G. and Lueder, R. and Drucker, S. and Wong, C. 2002 “MyLifeBits: fulfilling the Memex vision” *Proceedings of the tenth ACM international conference on Multimedia* p235- 238

Hodges, S. and Williams, L. and Berry, E. and Izadi, S. and Srinivasan, J. and Butler, A. and Smyth, G. and Kapur, N. and Wood, K. 2006 “SenseCam: A retrospective memory aid” *UbiComp 2006: Ubiquitous Computing* p177-193

Lee, K. and Ellis, D.P.W. 2006 “Voice activity detection in personal audio recordings using autocorrelogram compensation” *INTERSPEECH 2006: ICSLP: Proceedings of the Ninth International Conference on Spoken Language Processing: September 17-21, 2006, Pittsburgh, Pennsylvania, USA* p1970-1973

Leech, G. 2006 “New resources, or just better old ones? The Holy Grail of representativeness” *Language and Computers* (59) p133-149

Roy, D. and Patel, R. and DeCamp, P and Kubat, R. Fleischman, M and Roy, B. and Mavridis, N. and Tellex, S. and Salata, A. and Guinness, J. and Levit, M and Gorniak, P. 2006 “The Human Speechome Project.”Proceedings of the 28th Annual Cognitive Science Conference*.*