REQUIREMENTS DEFINITION AND DESIGN CRITERIA FOR TEST CORPORA IN INFORMATION SCIENCE

by
W. John MacMullen

School of Information and Library Science
CB#3360, 100 Manning Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360

SILS Technical Report 2003-03

School of Information and Library Science
University of North Carolina at Chapel Hill

Requirements Definition and Design Criteria for Test Corpora in Information Science

Abstract

This paper argues that structured collections of data and information (“corpora”) are needed for research in information science, and to measure the validity, accuracy, and effectiveness of tools, methods, and systems. It examines the needs and uses of corpora, and describes some specific examples from a variety of domains. The paper explores the relationship of scientific methods to corpora design, and then enumerates and discusses a variety of design criteria, primarily from the corpus linguistics literature.

Contents

1 Introduction

To be a science, the study of information must share the fundamental attributes of other sciences; for example, it must have testable theories and hypotheses; its practitioners must conduct experimentation; and those experiments must be reproducible by other investigators. One area for application of these ideas is formal and quantitative evaluation of software and information systems. This paper examines one specific class of tools used to facilitate evaluation: *test corpora*, defined here as structured collections of data and information used to measure the validity, accuracy, and effectiveness of tools, methods, and systems.

In a variety of disciplines and domains, activities that can broadly be termed "information discovery"[1] use software tools and structured collections of information to perform experiments, make comparisons, draw inferences, and extract meaning. Standardized collections of data and information are also used to test and validate software and information systems. Despite their importance and wide range of applications, little has been written directly about design principles for these collections from the perspective of information science. Nearly thirty years ago, Spärck Jones (1975) argued that "one major problem in experimental information retrieval is the lack of yardsticks representing good performance for test collections." This is largely still true today, although some exceptions will be reviewed below. Even publications that ostensibly discuss corpora construction in computational linguistics (the discipline doing a large amount of research using corpora) frequently do not explicitly characterize their underlying design methodology. This is quite different from published research in the sciences, where publications have a "methods" section that describes in detail the techniques used, composition of materials, characterization of samples, control of variation and error, and normalization of results.

Disciplines and domains other than information science have investigated corpus construction methodologies and enumerated design criteria for the collections used for their specific purposes. Examples from several domains are reviewed below to provide context and derive general principles that might be reused in information science.

2 Corpora Needs and Uses

The relatively recent availability of large quantities of digitized text and other data is changing the way many disciplines, from linguistics to biology, are thinking about and practicing scientific research. The names of a variety of disciplines can be substituted into Tognini-Bonelli's (2001:1) statement that "[w]hat we are witnessing is the fact that corpus linguistics has become a new research enterprise and a new philosophical approach to linguistic enquiry" as a result of new data. "It is strange to imagine that just more data and better counting can trigger philosophical repositionings, but […] that indeed is what has happened" (48). Empirical data provide context and the ability to confirm or deny what until then may have been only hypothesized. Hockey and Walker (1993:236) note that the computational linguistics community recognized that "researchers [had] been severely hampered by the lack of appropriate materials" for research, "specifically by the lack of a large enough body of text on which published results can be replicated or extended by others", and so subsequently established the ACL Data Collection Initiative. Similar ideas drove biologists to collaborate on the sequencing of whole genomes.

There is, in addition to corpora for research, a need for consciously created and organized collections of data and information that can be used to evaluate the performance and effectiveness of knowledge discovery tools. Extant collections have a variety of names that vary based upon the discipline of interest. One major distinction, then, can be made between research corpora and test corpora. As used in this paper, *research corpora* are collections of authentic data used to perform experiments to advance knowledge, while *test corpora* are collections of authentic or invented data used for testing, evaluating performance, and calibrating the tools used in experimentation. Names for research corpora include the general *corpus/corpora*, *monitor corpus*, *text collections*, and *data sets*, while test corpora names include *test collections*, *training sets*, *experimental retrieval collections*, and *test suites*. There may be sub-classes of each type for specific purposes; for example, Tognini-Bonelli (2001:8-9) describes collections in corpus linguistics for studying (among other things) translation, the language of learners, and varieties of language for specific purposes (LSP). From the linguistic perspective, the idea of a "general purpose" corpus for research seems difficult to create, although some researchers believe they would be valuable (Zampolli, 1995:59). One might imagine a "meta-corpus", i.e., a corpus of corpora, organized or annotated in such a way that extractions could be made for specific uses. (The benefit of this model would be the application of an extensible classification framework shared by all sub-corpora to facilitate ease of use, cross-corpora analysis, creation of standard analysis tools, etc.)[2]

Atkins, Clear, and Ostler (1992) provide a hierarchy of text collections (Figure 1), which can be composed of any type of content. A literary anthology can be called a corpus, as can a collection of laws; one of the key factors in design and use, therefore, is the purpose for which the corpus has been constructed (Tognini-Bonelli, 2001).

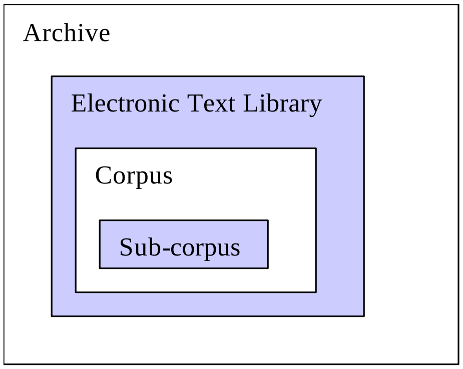

Figure 1. A hierarchy of text collections.

Within either research or test corpora, a distinction between "opportunistically" collected corpora (e.g., the Oxford Text Archive), and carefully designed, systematic collections (e.g., the Brown and SEU corpora) (Leech, 1991). The rest of this paper will explore the idea of systematically designed corpora. McEnery and Wilson provide what they call a "prototypical" definition of a systematically designed and collected corpus: "a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" (2001:32).[3] The content in these collections can be in a variety of forms: bibliographic citations, full texts of various types and lengths, transcriptions of speech, algorithms and rules, biological sequence data, etc. The common denominator is not the content, but the fact that the content has been systematically "engineered" to be representative of the particular problem domain.

3 Domain- and Task-Oriented Examples

As noted above, corpora as sources of empirical data are critical to both research and evaluation in a variety of disciplines. This section provides a brief survey of some use of corpora in several domains to demonstrate that investigators have independently arrived at the conclusion that corpora are beneficial to their research.

3.1 Linguistics

Many sub-disciplines in linguistics are using corpora for research: computational linguistics, lexicology and lexicography, communication theory and practice, language teaching, and computer-based training among them (Zampolli, 1995), as well as studies of speaking and the human voice (Perks and Crichton, 2000). Translation (research and practice) is another area of active study. Lindquist (1999:182) describes two types of corpora used in multilingual translation: parallel corpora and translation corpora. Parallel corpora consist of source texts and similar or related texts in target languages, while translation corpora are source texts and their translations into one or more target languages. Lindquist argues that the parallel corpora model is especially powerful for translation because the translator can see "the words and collocations in actual use in the appropriate type of text", and thus the resulting translation "is likely to sound more natural than it would have done otherwise". A variety of alignment mechanisms are used for multilingual translation (Véronis, 2000). Soler (1993) aggregates the importance of corpora to linguistics into three groups: a way to study real language in actual use; as "test beds for studying natural language products"; and "basic resources upon which to develop natural language software". There are several large-scale corpus linguistics initiatives underway, including the Network of European Reference Corpora (NERC) and the TSNLP project (Test Suites for Natural Language Processing) project (Calzolari, Baker, and Kruyt, 1995; Oepen, Netter, and Klein, 1998, respectively).

3.2 Information Science

The information science domain has for decades used the concept of a *test collection* to measure information retrieval system performance, beginning with the Cranfield experiments in the late 1950s (Robertson and Walker, 1997). Typical test collections in information science are composed of "documents" (in the form of titles, abstracts, and / or full text articles), a set of standardized queries or questions, and a set of "relevance judgments", typically made by experts, as to which documents are most appropriate for retrieved by each query. The standard measurements used to judge performance of IR systems are *recall* (the probability a relevant item will be retrieved) and *precision* (the probability that a retrieved item will be relevant).[4] Well-known test- and text collections in information science include the Cranfield collections (Salton, 1971; Robertson and Walker, 1997; Van Rijsbergen and Croft, 1975); the Reuters test collection (Sanderson, 1996), the Cystic Fibrosis (CF) Database (Shaw, et. al., 1991; Shaw, 1994), OHSUMED (Hersh, et. al., 1994),[5] and the TREC collections (TREC, 2002). These types of experiments have problems from an empirical perspective: while relevance is typically a binary determination, the fact that a document is relevant doesn’t guarantee that it is useful, or the answer a user is looking for. Many IR tools and Internet search engines determine relevancy largely based upon term frequency; but as Oepen, Netter, and Klein note, “a relevant phenomenon need not necessarily be a frequent phenomenon” (1998:35). There are questions as well about the relationship of the single-iteration query / relevance judgment model to real-life settings and systems. There are questions of accuracy and performance as well: whether the (typically) small collections used for testing are representative of real collections, and whether under real conditions a system that performs well on a collection of size *x* will function equally well on a collection of, say, size 100*x* or x^2 (Ledwith, 1992). Other related areas to information retrieval that use corpora include knowledge discovery in databases (KDD), data mining, information extraction, and latent semantic indexing.[6]

3.3 Bioinformatics

As more whole genomes are sequenced and the protein products of genes determined, the biology and bioinformatics communities will in essence be building their own versions of “parallel” and “translation” corpora in order to do comparative genomics. For that research, investigators need to see sequences DNA, RNA, and proteins of interest aligned with those that are (potentially) comparable in other organisms, as well as the context in which they appear, in order to determine degrees of similarity, amount of conservation, and time since evolutionary divergence. Often in biomedical IR and experimentation, multiple heterogeneous corpora or data sets are integrated or mined concurrently (Raychaudhuri, et. al., 2002; Tanabe, et. al., 1999). Examples of corpus-related research in bioinformatics are in the next section.

3.4 “Challenge” Competitions

While there may be an ostensible “winner”, for the most part these competitions are friendly and conducted in a spirit of collaboration and information sharing in order to advance their respective fields. The TREC conferences are one example of this, but the “knowledge discovery in databases” (KDD) and machine learning communities have their own challenge contests, which seem to focus almost exclusively on “real-world” data sets, which in many cases are unstructured (KDnuggets, 2002). The 2002 KDD Cup is a data mining competition held in conjunction with the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.[7] The somewhat similar “information extraction” (IE) domain, mentioned in 3.2, deals with finding specific answers to questions or identifying and extracting specific facts from collections.[8]

The life sciences have a variety of challenge competitions. In protein structure prediction there is CASP (the Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction) (CASP, 2002). For functional genomics there is CAMDA (the Critical Assessment of Microarray Data Analysis) (Johnson and Lin, 2001). Others exist for genomics (GASP, 2002), statistical genetics (GAW, 2002), and computational toxicology (Helma, 2001).

3.5 Overlap

In the above activities there are many areas of potential and actual overlap. Some examples include:

- Translation and cross-language information retrieval – TREC Cross Language track; Cross-Language Evaluation Forum (CLEF); NTCIR Asian Language Evaluation; other non-English language IR (Moukdad and Large, 2001; and Wu, 1999)
- Information retrieval, machine learning and biomedical disciplines – Proposed TREC pre-track on genomics
- Information retrieval, text mining, and literature-data integration in biomedical research (Raychaudhuri, 2002; Kim and Wilbur, 2001; De Looze and Lemarié, 1997; Tanabe, et. al., 1999)

We will undoubtedly see more as each domain matures. An interesting variation of the challenge competitions are the distributed- or grid-computing programs emerging that use the excess computing power of individuals and organizations to either generate data sets or determine answers to novel problems.[9]

3.6 Software Verification, Validation, and Testing (VV&T)

FDA guidance on software validation (FDA-US, 2002:5.2.5) states that *validation* is “confirmation by examination and provision of objective evidence that software specifications conform to user needs and intended uses, and that the particular requirements implemented through software can be consistently fulfilled”. […] “Software testing entails running software products under known conditions with defined inputs and documented outcomes that can be compared to their predefined expectations”. To achieve these objectives implies standardized test data sets, test cases, real-life use cases, etc. How much testing is enough to generate a sufficient confidence level that the software is valid? Testing using test corpora is only one method of software evaluation; there are automated programs that test code validity, for instance.

4 Scientific Foundations of Corpus Design

Tognini-Bonelli (2001) claims that while "it has been argued that corpus linguistics is not really a domain of research but only a methodological basis for studying language", one can in fact use corpora as part of an empirical approach to language study. Noting that corpus linguistics studies authentic data, she describes a general model consisting of observation of language facts, the formulation of hypotheses and generalizations based on patterns in data, and the subsequent derivation of theoretical statements (Figure 2).

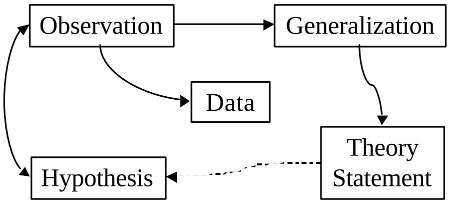

Figure 2. A generalized representation of Tognini-Bonelli's empirical model.

This is clearly a general model of empirical inquiry, and can be applied to other disciplines. However, there are some assumptions implicit in this model: the data are accurate and representative of the population under study; experiments can be reproduced; and generalizations can be made. Assuming those are true, much of the burden lies upon the design and creation of a representative corpus. *Representativeness* is the major issue in corpus design, and is driven by the identification of a specific population or focal point of study. Representativeness refers to "the extent to which a sample includes the full range of variability in a population" (Biber 1993).

4.1 Representativeness

Figure 3 illustrates the decision points surrounding representativeness: based on the desired function of the corpus, a strategy for sampling real-world data is developed; the data are selected and reviewed for representativeness, after which more sampling may be needed; a level of confidence or error is applied; the data are normalized; and the corpus is created. As in Tognini-Bonelli's model, reproducibility and generalizability are dependent upon the degree of representativeness. Engwall (1994) notes that availability of resources is a key constraint.

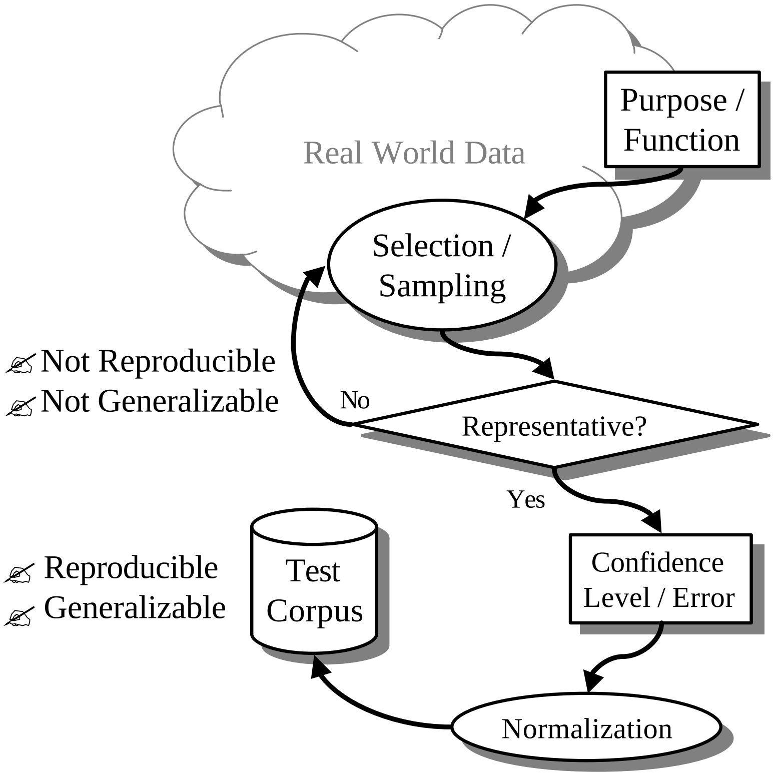

Figure 3. The centrality of representativeness.

Representativeness is a difficult question for linguistics – how much language is adequate to represent all language? Which words could be omitted? – but is a problem in many domains as well. Imagine you are building a corpus of seismic data; this data is months or years of data in a tight range, randomly punctuated by significantly outlying data. How can you select samples of data that are "representative" of all seismic data? In other domains such as molecular biology, this is potentially less difficult, because in linguistic terms there is a finite grammar: there are limited alphabets (5 nucleic acids, 20 amino acids) and their valid combinations are constraining factors. Rather than 4^2 combinations of nucleic acids, there are only two (plus their complements), for example.

Representativeness is multifaceted, in that some texts, for example, are selected from the universe of all existing texts (which itself represents a subset of all possible text), and then in many cases, smaller samples are derived from those texts. Not only is the representativeness of the entire language important, but how representative the sources are from which the samples are drawn.

In information retrieval research, another aspect of representativeness is the relationship of the queries (questions, topics) to reality. Tague-Sutcliffe (1992:476) says that “[a] query is a verbalized information need”. How realistic are the queries that are constructed? Are they invented, or perhaps derived from actual queries collected from real users in the domain of interest? And how representative are those of questions in that domain in general? Some question whether corpora by their nature can in fact be designed to be representative, or classified into certain categories (Spärck-Jones, 1973; Atkins, Clear, and Ostler, 1992).

4.2 Sampling and Statistics

> “Although statistics are often used to test a particular hypothesis […], statistics can also be used to *explore* the space of possible hypotheses, or to *discover* new hypotheses […]
>
> (Church and Mercer, 1993).

Zampolli, speaking for NERC, (1995:59) says that there is an urgent need for “multifunctional general language corpora” covering “many domains, registers, and comunicative situations”, and organized in a balanced, representative way. It is difficult to understand how one would operationally determine representativeness in a general language corpus. Many in the linguistics community, including Engwall (1994), and Atkins, Clear, and Ostler (1992:4, 6), are skeptical about the ability to do valid statistical sampling of linguistic text due to ill-defined populations and unit criteria. They are critical of the idea that a “balanced” corpus is essential prior to beginning research, arguing that representativeness is an iterative process (as does Biber, 1993:256) that gets optimized based on feedback over time as the corpus is used. They also support the position taken by Woods, Fletcher, and Hughes (1986) that when the particular sampling method of a corpus is unknown, linguists should assume “sampling had been carried out in a theoretically ‘correct’ fashion”. This is antithetical to the scientific method, where populations are rigorously defined in advance and methods are disclosed fully. Kretzschmar, Meyer, and Ingegneri (1997) encourage skepticism on the part of researchers using corpora whose design is unknown, and encourage developers of corpora to use probability sampling in their design and construction activities. “Random” sampling can be problematic; Tognini-Bonelli (2001) notes that “few linguistic features of a text are distributed evenly throughout”. Biber (1993) suggests that, in general, stratified sampling results in a more representative corpus than proportional sampling. He also notes that from an analysis perspective, much corpus-based research is univariate in nature, and suggests that multivariate techniques such as factor analysis and cluster analysis are useful in meta-analysis of corpus representativeness. Some of these ideas are discussed in greater detail in Oakes (1998).

4.3 Reproducibility

Experimental reproducibility is a hallmark of scientific inquiry. The ability to reproduce one's own (or others') experiments is important not only from the standpoint of validation of results, but in the general advancement of a discipline. However, in information science, as well as other disciplines, "individual projects typically work with their own data […]. This makes it extremely difficult to compare the results obtained by different projects […]; so cumulative progress in understanding how retrieval systems work, through the correlation of a range of results, is low" (Spärck Jones and Van Rijsbergen, 1976). In the case of the TREC conferences, having a standardized test corpus allows a relative level of comparability among results, even thought the specific approaches by different investigators vary. Reproducibility and generalizability are dependent upon representativeness. Conversely, there is the question of adaptability to environmental change and the evolution of knowledge (i.e., the need to add new observations and to change the physical structure due to improved tools (relational versus flat file, etc.) (Chafe, Dubois, & Thompson, 1991).

4.4 Error and Normalization

There are a variety of systematic errors that can be introduced into corpus experiments: theoretical, instrument (e.g., calibration), and operator, but given the centrality of representativeness, the most important is probably sampling error. Sampling errors can be due to chance or rarity of the instance that is selected, or to more serious problems resulting from non-systematic (opportunistic) corpus construction. The estimated error is inversely proportional to the sample size (i.e., the smaller the sample, the less representative it is of the population as a whole, and thus the greater chance for error) (Friedman and Wyatt, 1997; see also Atkins, Clear, and Ostler (1992:4-5). Normalization of data is necessary in the construction phase with regard to sample format as well as other sample attributes (length, for example, if doing a stratified sample). Error impacts all facets of the experimental process: validity, reliability, and efficiency Tague-Sutcliffe, 1992).

5 Design Desiderata and Criteria

As section four argued, all corpora need to be "designed" to a certain extent in order to be useful and valid; this is most important in test corpora, as they are used to validate functionality and accuracy of tools. Comparisons of tools are carried out to assess relative performance and relative advantage (performance comparison on the same task versus suitability to a particular task).

Spärck Jones and Van Rijsbergen (1976) refer to an “ideal collection” for information retrieval as “satisfying the requirements for commonality between projects, hospitality to projects, adequacy for projects, and convenience in projects”.[10] Implicit in this conception is the idea of reuse of corpora, both for multiple experiments and by other investigators. Hockey and Walker (1993:235) argue that “[t]o be truly reusable, a corpus needs to be considered in the light of multifunctionality, polytheoreticity, acquirability, intellectual property rights, representativeness, standardization, availability, and evaluation”. This is often difficult given the variation in both corpora as well as the types of experiments investigators are interested in exploring.

Spärck Jones and Van Rijsbergen (1976) identified variation among seven collections extant in 1975 as varying in size; subject matter; indexing source (i.e., extent of sample, ranging from title to full text); number and types of indexing language (i.e., ranging from title words to controlled subject headings); and variation in treatment of relevance. They concluded that the collections were incomparable, since they shared no common variables with the same values across all collections. They also distinguished types of bounded populations that could be represented in corpora, including text style, document type, subject, source, origin, citation, request, (user) need, user type, and vocabulary.

Spärck Jones (1975) and Spärck Jones and Van Rijsbergen (1975, 1976) set forth some design criteria for ideal test collections, particularly with respect to control of variables, saying that collections should be both variable and homogeneous with regard to: content, type, source, origin, time range, and language. In information retrieval research, where systems are evaluated for their effectiveness at meeting the needs of a certain population of users, there are a number of variables to consider; Tague-Sutcliffe (1992:471) provides four classes: type of user; context of use; kinds of information needed; immediacy of information need. Below are several design criteria, identified mainly in the corpus linguistics literature.

- **Function** – A well-formulated idea of the purpose(s) to which the corpus will be put is very important (Tognini-Bonelli, 2001: 55; Zampolli, 1995; Oepen, Netter, and Klein). This can be viewed as a part of the overall experimental design; even if an experiment is using an existing corpus, this step is important to insure the results are appropriate. While there is discussion of needs for multi-functional corpora, it is not clear that a corpus of that composition would be representative or yield valid results for any of the tasks.

- **Representativeness** – As discussed above, this is a critical and multifaceted criterion, as representativeness of a corpus "determines the kind of research questions that can be addressed and the generalizability of the results of the research" (Biber, Conrad & Reppen, 1998). Data should be authentic and reflective of reality, whether it is natural language in actual use, valid DNA sequences, etc. (Tognini-Bonelli, 2001:55; McEnery & Wilson, 2001). Test data doesn't mean false or invalid data; "dummy" data should be real.[11]

- **Sampling** – There are many methods for valid sampling based upon the content in question (proportional, stratified, based on a defined typology; and impacted by length considerations) (Hockey and Walker, 1993; Biber, 1993; Biber, Conrad, and Reppen, 1998), but sampling in general should be maximally-representative of the desired variety of content (McEnery and Wilson, 2001).

- **Size** – Very large corpora are not necessarily more representative; they can in fact cause problems in finding less abundant but more important items. Biber, Conrad & Reppen (1998) say that "size cannot make up for lack of diversity". But the corpus must be large enough to yield statistically significant results. Perhaps we can say that corpora should be "as large as necessary, but as small as possible". With the exception of special cases such as monitor corpora, corpus size should be fixed – changes in size and content in mid-experiment impact the reliability of quantitative data and comparability (McEnery & Wilson, 2001).

- **Scope** – Scope and diversity are related to function and representativeness, ranging from content varieties (e.g., register variation, dialect, subject matter) to time periods (synchronic [a defined period] versus diachronic [measuring change over time]) (Atkins, Clear, and Ostler, 1992; Biber, Conrad and Reppen, 1998).

- **Availability and Feasibility** – Availability should not be a driving factor in design because bias is introduced and the corpus moves away from representativeness. Nonetheless, content availability, intellectual property issues, and cost all affect design decisions. Rigorous investigators such as Tague-Sutcliffe (1992) might counsel against doing an experiment at all if a valid corpus cannot be built.

- **Reusability** – “To be truly reusable, a corpus needs to be considered in the light of multifunctionality, polytheoreticity, acquirability, intellectual property rights, representativeness, standardization, availability, and evaluation” (Hockey and Walker, 1993:235). Since they are so difficult, time consuming, ad expensive to construct, the assumption is that any particular corpus is “the standard” for the domain it represents, and should be made available if possible to other researchers who are working in that area. “The advantage of a widely available corpus is that it provides a yardstick by which successive studies may be measured” and that “a continuous base of data is being used and thus variation between studies may be less likely to be attributed to differences in the data being used, and more to the adequacy of the assumptions and methodologies contained in the study” (McEnery & Wilson, 2001:32). Reusability across different applications and extensibility are open questions (Oepen, Netter, and Klein, 1998).

6 Conclusions

Corpora are “reservoirs of evidence” (Tognini-Bonelli, 2001:55) that can be used in the scientific study of natural phenomena, phenomena ranging from natural human language to natural genetic language. This paper has tried to set out some criteria that would be useful to investigators designing, constructing and using corpora for a variety of purposes in information science research. While discussing concrete details, the criteria have been kept general because as argued above, optimal corpora are those that are designed for specific purposes. As Ostler (1993) suggests, it would be difficult, if not impossible, to make very detailed decisions about corpora design without knowing for what purpose a particular corpus is being used. This paper has also focused on the intentional, systematic design of corpora. While it may be true that, as Knowles (1996:36) argues, “[a]n important consequence of handling large amounts of data is that it enforces rigour and discipline in data organization”, this in itself is no guarantee that the data are representative.

To facilitate understanding about appropriate use, corpus constructors should communicate or make available the specific details of the makeup of the corpus (sources, source populations, samples, sample sizes, etc.). Figure 4 presents an evaluation process for an information system or tool using a test corpus.

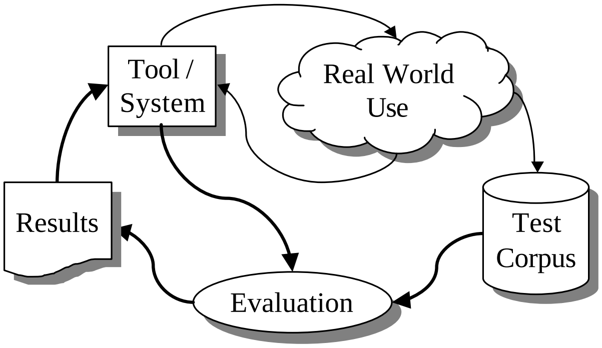

Figure 4. A general evaluation process for test corpora

6.1 Other factors

Annotation schemes should be a factor in corpus design because they impact collection and implementation (Ide and Priest-Dorman, 2000; Oepen, Netter, and Klein, 1998; Hockey and Walker, 1993; Leech, 1993; Shaw, 1993); however, they are not discussed in detail in this paper. Any interesting area for future study would be whether a resource description scheme would enable easier corpus design and construction if the attributes of a resource could be analyzed in an automated fashion. This is a growing upstream problem; Hockey and Walker (1993) note that the number of electronic texts that are available is rapidly growing, but there aren't "consistent guidelines and procedures for documenting, storing, and maintaining" them.

Another factor that may impact corpus design is the decision of which structural or database model to use when implementing the corpus (Nerbonne, John, ed., 1998; Knowles, 1996; Oepen, Netter, and Klein, 1998). Knowles also suggests research into a relational model of language that can be used to draw inferences and make generalizations. There are a host of other data manipulation issues to address when moving into the collection and implementation stages, addressed by Thompson (2000) and others.

Some of the ideas discussed here could be applied more broadly to information and library science research, particularly representativeness: what is a representative sample of search logs, or circulation records, or spam vs. non-spam email? Surveys are used very heavily in ILS research, but discussions of the representativeness of their populations are infrequent.

References

Armstrong, S. (1995). Using large corpora. *Information Processing and Management* 31:5:785.

Atkins, Sue, Clear, Jeremy; Ostler, Nicholas (1992). Corpus Design Criteria. *Literary & Linguistic Computing* 7:1:1-16.

Biber, Douglas (1993). Representativeness in Corpus Design. *Literary & Linguistic Computing* 8:4:243-257.

Biber, Douglas; Conrad, Susan; and Reppen, Randi (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Calzolari, Nicoletta; Baker, Mona; and Kruyt, Johanna G., eds. (1995). *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study*. Pisa: Giardini.

CASP (2002). The Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. http://predictioncenter.llnl.gov

Chafe, Wallace; Dubois, John W.; and Thompson, Sandra A. (1991). Towards a New Corpus of Spoken American English. In Aijmer, Karen and Altenberg, Bengt (eds.): *English Corpus Linguistics, Studies in Honour of Jan Svartvik.* London, Longman; pp. 64-82.

Church, Kenneth W. and Mercer, Robert L. (1993). Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19:1:1-24.

De Looze, Marie-Angèle and Lemarié, Juliette (1997). Corpus relevance through co word analysis: an application to plant proteins. *Scientometrics* 39:3:267-80.

Engwall, Gunnel (1994). Not Chance but Choice: Criteria in Corpus Creation. In Atkins, Sue B.T., Zampolli, Antonio (eds.) *Computational Approaches to the Lexicon*. Oxford: Oxford University Press 49-82.

FDA-US (United States Food and Drug Administration), Center for Devices and Radiologic Health (2002). General Principles of Software Validation; Final Guidance for Industry and FDA Staff. http://www.fda.gov/cdrh/comp/guidance/938.html

Friedman, Charles P. and Wyatt, Jeremy C. (1997). *Evaluation Methods in Medical Informatics*. NY: Springer. (Computers in Medicine series)

GASP (Genome Annotation Assessment Project)
http://www.fruitfly.org/GASP1

GAW (Genetic Analysis Workshop)

http://www.sfbr.org/gaw/

Helma, C.; King, R. D.; Kramer, S.; and Srinivasan, A. (2001). The Predictive Toxicology Challenge 2000–2001. *Bioinformatics* 17:107-108. http://www.informatik.uni-freiburg.de/~ml/ptc/

Hersh, W.; Buckley, C.; Leone, T.J.; and Hickman, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In Croft, W. Bruce and Van Rijsbergen, C. J., eds., SIGIR '94: International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 192-201. New York: ACM. http://www.ohsu.edu/bicc-informatics/hersh/sigir94.pdf

Hockey, Susan and Walker, Donald (1993). Developing Effective Resources for Research on Texts: Collecting Texts, Tagging Texts, Cataloguing Texts, Using Texts, and Putting Texts in Context. *Literary & Linguistic Computing* 8:4:235-242.

Ide, Nancy and Priest-Dorman, Greg (2000). The Corpus Encoding Standard. http://www.cs.vassar.edu/CES/

Johnson, Kimberly F. and Lin, Simon M. (2001). Call to work together on microarray data analysis. *Nature* 411:885. http://bioinformatics.duke.edu/camda/CAMDA00/paper.asp

KDnuggets (2002). Data Mining Competitions. http://www.kdnuggets.com/datasets/competitions.html

Kim, Won and Wilbur, W. John (2001). Corpus based statistical screening for content bearing terms. *Journal of the American Society for Information Science and Technology*. 52:3:247-59.

Knowles, Gerry (1996). Corpora, databases, and the organization of linguistic data. In Thomas, Jenny and Short, Mick, (eds.) *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*. London: Longman.

Kretzschmar, Jr., William A.; Meyer, Charles F.; and Ingegneri, Dominique (1997). Uses of Inferential Statistics in Corpus Studies. In Ljung, Magnus (ed.) *Corpus-based Studies in English: Papers From the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17) Stockholm, May 15-19, 1996*. Amsterdam: Rodopi.

Krieger, Elmar and Vriend, Gert (2002). Models@Home: Distributed computing in bioinformatics using a screensaver based approach. *Bioinformatics* 18: 315-318.

Ledwith, Robert (1992). On the difficulties of applying the results of information retrieval research to aid in the searching of large scientific databases. *Information Processing and Management* 28:4:451-455.

Leech, Geoffrey (1993). Corpus Annotation Schemes. *Literary & Linguistic Computing* 8: 4:275-281. (See also “Leech's Maxims of Annotation”, http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2maxims.htm)

Leech, Geoffrey (1991). The State of the Art in Corpus Linguistics. In Aijmer, Karen and Altenberg, Bengt (eds.): *English Corpus Linguistics, Studies in Honour of Jan Svartvik*. London, Longman; pp. 64-82.

Lindquist, Hans (1999). Electronic corpora as tools for translation. In Anderman, G. & Rogers, M., eds. *Word, Text, Translation*. Clevedon, England: Multilingual Matters.

McEnery, Tony and Wilson, Andrew (2001). *Corpus Linguistics: An Introduction* (2nd Ed). Edinburgh: Edinburgh University Press.

Moukdad, H. and Large, A. (2001). Information retrieval from full-text Arabic databases: can search engines designed for English do the job? *Libri* 51:2:63-74.

Nerbonne, John, ed. (1998). *Linguistic Databases*. Stanford: CSLI.

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh : Edinburgh University Press.

Oepen, Stephan; Netter, Klaus; and Klein, Judith (1998). TSNLP – Test Suites for Natural Language Processing. In Nerbonne, John, ed. *Linguistic Databases*. Stanford: CSLI.

Ostler, Nicholas (1993). Introduction to Part One [of two special sections on corpora]. *Literary & Linguistic Computing* 8:4:221-223.

Perks, R. and Crichton, C. (2000). The Millennium Memory Bank: a test case in archival collaboration. *IASA Journal* 15:40-9.

Raychaudhuri, Soumya; Chang, Jeffrey T.; Sutphin, Patrick D.; and Altman, Russ B. (2002). Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. *Genome Research* 12:203-214.

Robertson, S.E. and Walker, S. (1997). Laboratory experiments with Okapi: participation in the TREC programme. *Journal of Documentation* 53:20-34.

Salton, Gerald (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Sanderson, M. (1994). The Reuters test collection. In Leon, Ruben, ed. *Proceedings of the Sixteenth Research Colloquium of the British Computer Society Information Retrieval Specialist Group, Drymen, Scotland, 22-23 Mar 94*. London: Taylor Graham, 1996, p.219-27.

Shaw, William M., Jr. (1994). Retrieval Expectations, Cluster-Based Effectiveness, and Performance Standards in the CF Database. *Information Processing & Management* 30:5:711-723.

Shaw, William M., Jr. (1993). Controlled and Uncontrolled Subject Descriptions in the CF Database: A Comparison of Optimal Cluster-Based Results. *Information Processing & Management* 29:6:751-763.

Shaw, William M., Jr., Wood, Judith B., Wood, Robert E., & Tibbo, Helen R. (1991). The Cystic Fibrosis Database: Content and Research Opportunities. *Library and Information Science Research* 13:347-366. ftp://ils.unc.edu/pub/research/cfdbase/

Soler, José (1993). Text Corpora: Meeting the Challenge of Information Excess. *Literary & Linguistic Computing* 8: 4:225.

Spärck Jones, Karen (1975). A performance yardstick for test collections. *Journal of Documentation* 31:4:266-272.

Spärck Jones, Karen (1973). Collection properties influencing automatic term classification performance. *Information Storage and Retrieval* 9:9:499-513.

Spärck Jones, Karen and Van Rijsbergen, C. J. (1976). Information Retrieval Test Collections. *Journal of Documentation* 32:1:59-75.

Spärck Jones, Karen and Van Rijsbergen, C. J. (1975). *Report on the need for and the provision of an 'ideal' information retrieval test collection*. Cambridge (UK), Cambridge University Computer Laboratory, Dec 1975. [unavailable; not reviewed]

Tague-Sutcliffe, Jean (1992). The Pragmatics of Information-Retrieval Experimentation, Revisited. *Information Processing & Management* 28:4:467-490

Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L., and Weinstein, J.N. (1999). MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling. *BioTechniques* 27:6:1210-1217.

Thompson, Henry S. (2000). Corpus Creation for Data Intensive Linguistics. In. Dale, R., Moisl, H. & Somers, H. (Eds.) *Handbook of Natural Language Processing*. New York: Marcel Dekker Inc. Ch. 16, pp. 385-401.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

TREC (Text REtrieval Conference), National Institute of Standards and Technology (NIST). http://trec.nist.gov/

Van Rijsbergen, C. J. and Croft, W. B. (1975). Document clustering: an evaluation of some experiments with the Cranfield 1400 collection. *Information Processing and Management* 11:171-182.

Véronis, Jean, ed. (2000). Parallel Text Processing: Alignment and Use of Translation Corpora. Boston: Kluwer Academic Publishers.

Woods, A.; Fletcher, P.; and Hughes, A. (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Wu, M.-M. (1999). Proposing a prototype for Chinese corpus test collection. [In Chinese] *Journal of Library and Information Science* 25:1:68-87.

Zampolli, Antonio (1995). Corpus Design Criteria. In Calzolari, Nicoletta; Baker, Mona; and Kruyt, Johanna G. (eds.) *Towards a Network of European Reference Corpora: Report of the NERC Consortium Feasibility Study*. Pisa: Giardini.

Notes

1. I use “discovery” here for its implication that in many cases, the knowledge or information that a person desires to find may not be known, and in fact the tools used to find this knowledge may be performing novel experiments. The term “retrieval” to me connotes more of a search for known items.

2. I suppose that the International Corpus of English is an example of this concept; the site says it has a common corpus design for each of its component corpora. http://www.ucl.ac.uk/english-usage/ice/design.htm

3. This definition excludes content such as audio and video, for example, but it is targeted toward linguistics, not general use.

4. These measures are only determinable in a controlled environment; in real-life collections such as the world wide web, these are impossible to determine because the collection composition is unknown, as well as dynamic in many cases. The assumption is that the more representative the test collection is, and the more realistic the questions or queries are, the better the performance of the system will be in real life. One could question this assumption since real life collections such as the web are more representative of certain types of materials than others.

5. OHSUMED -- "This test collection was created to assist information retrieval research. It is a clinically-oriented MEDLINE subset, consisting of 348,566 references (out of a total of over 7 million), covering all references from 270 medical journals over a five-year period (1987-1991)” (Hersh, et. al., 1994).

6. Examples:
   - *KD/KDD:* KD Nuggets (http://www.kdnuggets.com/datasets/)
   - *IE:* UC Irvine Knowledge Discovery in Databases Archive (http://kdd.ics.uci.edu/)
   - *LSI:* University of Tennessee Latent Semantic Indexing site (http://www.cs.utk.edu/~lsi/)

7. ACM KDD 2002
   - KDD Cup: http://www.biostat.wisc.edu/~craven/kddcup/
   - KDD 2002 http://www.acm.org/sigkdd/kdd2002/

8. Information Extraction
   - MUC conferences: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
   - RISE (Repository of Online Information Sources Used in Information Extraction Tasks): http://www.isi.edu/~muslea/RISE/
   - University of Sheffield NLP Group: http://gate.ac.uk/ie/

9. Distributed / Grid Computing
   - Genome @ Home: http://gah.stanford.edu/
   - Folding @ Home: http://fah.stanford.edu/
   - Krieger and Vriend (2002)

10. A wider factor here is how reflective of reality is the overall information discovery system; this has long been an issue in information science / IR, more so since the advent of interactive and web-based search engines. This is beyond the scope of this paper, but perspectives from practitioners such as Ledwith (1992) are instructive.

11. Their particular ideal collection never came to fruition, but many of their ideas, carried over from the Cranfield experiments, were integrated into what became the TREC conferences.