Linguistics in the Internet Age: Tools and the Fair Use of Digital Data

William D. Lewis
University of Washington/CSU Fresno
Scott Farrar
University of Washington
D. Terence Langendoen [1]
National Science Foundation

Abstract

The current work explores the fair use of linguistic data in the context of the Internet. It is argued that because of recent strides in Internet technology, empirical linguistics is now at a critical turning point with respect to the way data are reused and disseminated. As the use of the Internet becomes more and more commonplace, the possibility of data misuse is becoming more acute, not only because data are now broadly accessible to anyone who is on-line, but also because of the development of very precise search engines that may access and reuse data in an automated fashion. The issue of what constitutes fair use in this domain, especially when considering the boundaries established under copyright law, is neither clear nor well-defined. As a solution, a set of principles, called Principles of Reuse and Enrichment of Linguistic Data, or **PRELDs**, is proposed for data found on the Internet. Each of these principles is developed by considering examples from traditional print-based media that show how linguists have used, reused, and, at times, misused data. The principles are put to use by considering how automated linguistic tools, especially those that have the potential to stretch the limits of fair use, can be made to promote linguistics as a cutting-edge scientific enterprise, while preserving the rights of authors and respecting individual scholarship.

1 Introduction

The mid-19th century saw the advent of scholarly journals as the main way to promulgate scientific results in linguistics. Though mostly dedicated to historical and philological topics, these early journals allowed data that were at one time shared only among a very small community to be made widely accessible. For the first time, someone's analysis was in the public forum and available for scrutiny by other linguists. It is safe to say that the field has benefited enormously from this advancement, as the *verifiability* of data is one of the key principles of the scientific method. Today, the paradigm of the scholarly journal is beyond question. At the beginning of the Internet age, however, the field of linguistics is, along with every other field, entering a new era of scholarly communication. Its primary data are becoming accessible in a way that, just 15

[1] This material is based in part upon work supported while Langendoen was serving at the National Science Foundation. Any opinion and conclusions are those of the authors and do not necessarily reflect the views of the National Science Foundation.

years ago, was only hinted at in the pages of science fiction. With only a few minutes at the keyboard, anyone with an Internet connection can discover interesting facts even about the world's lesser-known languages. Searches can be performed, and actual example data, often with accompanying linguistic analysis, can be downloaded and stored away for a future research paper—the very model of progress for the empirically minded. Such a process is undertaken routinely and without question within our scientific community. At the same time, electronically available data can be manipulated in other ways, potentially with illegal, unethical or malicious intent. For example, the fruits of some scholar's hard work can be "borrowed" with little regard for copyright or for the context in which it was presented. With the development of ever more precise search engines, combined with a phenomenal increase in the number of people who are "connected", the problem of data misuse will become even more acute. In this paper we explore this rapidly emerging issue and propose solutions that are intended to serve the individual scholar while promoting linguistics as a cutting-edge scientific enterprise. Though some aspects are certainly unique to the age of the Internet, the issues involved are remarkably similar to those faced when data first became available in published journals. We argue that the availability and searchability of massive amounts of linguistic data will ultimately benefit and perhaps revolutionize the field, if only certain caveats are heeded.

2 Linguistic Data and the Scientific Enterprise

Progress in linguistics is driven by both data and theory: data helps us to build accurate pictures of language universals and human cognition, and theory helps us make predictions about data that we have not yet encountered. Should we discover data that are not compatible with our theory, we revise and rebuild it, or we throw it away and start anew. Crucial to the linguistic enterprise is the reuse of already published data.

2.1 The reuse of data

Linguistically relevant data exist in a variety of forms, from raw unanalyzed language data that a field linguist might record, to the same data that have been analyzed and subsequently annotated. What gets analyzed and what gets annotated is subject to significant variation across the field, highly dependent on research tradition, language family and sub-discipline. However, it is relatively uncommon for raw, unanalyzed language data to get disseminated. Rather, chunks of analyzed and annotated data find their way into publication, most often as part of some larger analysis, perhaps as part of a grammar, a lexicon, or, most commonly, as part of the analysis presented in a journal article. Once published, these snippets of data become available to the discipline as a whole, and are often recycled and reused. The act of "borrowing" data is accepted practice in the field, since it is not possible for every linguist to engage in fieldwork to locate primary data. Such second sourcing of data is the only way that the discipline can continue to grow and prosper, so as a practice, it is perfectly acceptable to borrow data found elsewhere, the only proviso being that the source of said data should be adequately cited.

As an example of how data are used and reused in print media, consider (1) which shows an example of interlinear data for Bulgarian that has been borrowed extensively. The source document for this example is Rudin (1988), *On Multiple Questions and Multiple Wh Fronting.* Bailyn (2003), Dayal (2003), Richards (1997), Richards (1999) and Stjepanovic (2003), among others, all use this example directly, changing only the example number. All cite Rudin. With the exception of Bailyn (2003), all are scholarly papers dealing with multiple-Wh, with Bailyn (2003) being course material on the same topic. Karimi (1999) also cites Rudin, but her source was Richards (1997), so she actually cites *both* papers as sources for her data, indicating first Richards (1997), followed by Rudin (1988). (The chain actually grows: we first discovered this example in Karimi (1999), so our presentation here represents the third layer of reuse for this particular datum in this chain!)

(1) Koj kogo vižda
Who whom sees
'Who saw whom?'

Locating relevant data has always been one of the more difficult aspects of linguistic inquiry. Linguists' interests are typically very narrow (by necessity), and finding data that fit into some highly constrained analytical framework can be difficult and time consuming. You might be interested in how quantifiers are distributed across sentences in a number of languages—how they appear in a sentence, whether or not they are fronted or exist *in situ*, how they interact scopally—and there might be particular configurations you want to observe in a large number of languages. It might be that across a variety of language data, you will discover that crucial datum that will refute your theory, or it might be the sheer volume of particular forms that will confirm it. Scouring papers, grammars, and texts for data of relevance is the tried and true method that linguists employ. Likewise, the habit of copying down examples, squirreling them away for some later date, and then eventually reanalyzing and/or republishing them is equally tried and true. No one sees anything wrong in the method. Although a particular reanalysis might be disputed, it is rarely the case that the "reuse" itself is disputed. It is what we do.

2.2 Reuse of digitally accessible data

But now, the field stands on the edge of a major revolution in how we can locate and reuse our data. Since so much data are being published in electronic form, whether in electronic journals or posted directly to the Web, much of this data can now be automatically discovered by machines. The method of scouring documents in the library for those precious little pieces of data is now being replaced by automated means that do the same thing. It is the rare linguist that has not Googled at some point in his or her career. If you want to find data on Awa-cuaiquer, for instance, an endangered language spoken in Columbia and Ecuador, the first stop might just be Google, or maybe even some linguist-friendly search engine such as the OLAC search facility.[2] Even queries for obscure or little known languages can return hundreds of hits, some containing linguistic data. With some clever refinements the hit rate can often be increased. For example, by specifying terms frequently used in linguistic annotation, such as *ERG*, *PERF*, *PAST*, *DUAL*, in combination with the language name, can significantly improve the quality of the results.

Taking this a step further, a search engine built for linguists could be designed to distinguish what is linguistically relevant from what is not, and return only those documents that are the most salient. One step beyond this would be the independent display of the actual linguistic data of relevance, that is, separate from the actual documents, though perhaps with links to them.

At a fundamental level, this modern scenario differs little from what has been done by generations of linguists. What has changed, however, is the sheer volume of data that can be retrieved and processed. For instance, a search engine that is trained to retrieve snippets of **interlinear glossed text** (**IGT**) could extract large numbers of examples from source documents for any given language. If we were looking for generic examples of data for Hausa, as an example, and our search engine came across Abdoulaye's dissertation (Abdoulaye 1992) on Hausa, we would find ourselves confronted by 853 instances of Hausa data extracted from this document. Similarly, if we were to look for data about the structure of noun phrases cross-linguistically, we might find Dryer's *Noun Phrase Structure* (Dryer 2006), and be presented with 177 examples from 74 different languages. In both cases, we could have found all of these examples manually. We could have copied them down, stored them away, and republished one or another of them at some later date, all the while citing Abdoulaye or Dryer when we did. Nothing has changed, except for the tools we use to find the data.

In such a world, what then constitutes fair use? If a search engine returns a list of data extracted from one or more linguistic papers, how should that data be presented? How should the presentation and query honor the authors' intentions? Even more important, how should the linguist who crafted the query—who can now merely cut and paste results into a file on his computer—be beholden to the source of the data?

The brave new world of the Web is now forcing us to ask these and other questions. The solutions can borrow from existing standards and mechanisms, but must be modified considering the differences between the world of print and the World Wide Web. In the following sections we present the problems posed by legacy[3] data reuse, specifically within the electronic domain, and suggest

[2] OLAC, the Open Language Archives Community, is dedicated to collecting information on languages and making it available to search. OLAC can be searched through an interface at LINGUIST List (http://www.linguistlist.org/olac/index.html) or at the Linguistic Data Consortium (http://www.language-archives.org/tools/search/). See Simons & Bird (2003).

[3] *Legacy* refers to the fact that most of the data currently on the Internet is in a format intended for display purposes only (e.g., PDF or HTML) and not necessarily intended to be

solutions to these problems in the form of a set of general principles. We then discuss these principles in light of tools that search for and manipulate linguistic data, and how and to what extent these tools must observe the principles.

3 Fair Use and Linguistic Data

What constitutes fair use, especially within the boundaries of copyright law, is neither clear nor well-defined, neither by the statutes nor by case law (Liberman 2000). Liberman (2000) points to the following reference in U.S. Copyright Law, Section 107[4] about fair use of copyrighted material:

> Notwithstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—
>
> 1. the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> 2. the nature of the copyrighted work;
> 3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> 4. the effect of the use upon the potential market for or value of the copyrighted work.

Since most data reuse within linguistics is done for research and educational purposes, the law would appear to be fairly lenient as to how material can be reused. As Liberman (2000) points out: "Taken as a whole, this in fact suggests a wide latitude for fair use in language documentation, since the purpose is typically 'teaching ... , scholarship, or research', the use is typically nonprofit, and the 'copyrighted work' is typically not a work for which there is otherwise a commercial market." The other provisions of the section should not be taken lightly, however; fair use of other's material should always carefully weigh the four factors outlined in Section 107. [5]

machine-readable.

[4] http://www.copyright.gov/title17/92chap1.html#107

[5] Section 107(4), in particular, has been invoked in two recent lawsuits involving Google Book Search: *Author's Guild v. Google Inc.* filed September 20, 2005, and *McGraw Hill Companies, Inc., Pearson Education, Inc., Penguin Group (USA) Inc., Simon & Schuster, Inc. and John Wiley & Sons, Inc. v. Google Inc.* filed October 19, 2005, the latter filed by the American Association of Publishers (AAP) on behalf of the plaintiffs (all are members of the AAP). The lawsuits challenge Google's effort to create digitized copies of 15 million books in

Nevertheless, we argue that fair use in the field of linguistics, as it is currently understood, is actually more restrictive than in copyright law; it is governed not only by law, but by years of uncodified, common practice in the field. We have distilled this common practice into a set of principles, each of which we call a Principle of Reuse and Enrichment of Linguistic Data, or **PRELD**. PRELDs take on added significance considering the changes in scholarly communication described above, and must take into account the fact that electronic tools for dissemination, search, and interoperation currently exist or are being constructed. Furthermore, tool makers and users should continue to observe long-established traditions and customs of the field, while at the same time taking advantage of the significant advantages provided by the rich and highly distributed nature of the Web. These principles are identified and summarized here. Each are discussed in more detail in the following sections.

- **Attribute Fully**—Indicate the full citation trail for a given datum, preferably to its source.
- **Honor Author Intent**—Avoid changes to data or analysis that mislead the reader from the original author's intent, while at the same time allow for changes that permit differences in terminology, presentation format or analyses.
- **Acknowledge Ownership**—Determine the limits of ownership of data, and identify the point at which changes or enrichments to source data reflect new ownership.
- **Allow Data to be Sheltered**—Provide a means in the electronic environment for an author to disallow use of his or her data or allow for the retraction of deprecated analyses.

the Harvard, Michigan, Oxford and Stanford libraries, and the New York Public Library and the Library of Congress as part of a new on-line service that provides unprecedented access to our written cultural and scientific heritage. The facility allows the user to search the contents of any of the scanned books using a front end already familiar to Google users. Depending on the level of access granted to a particular book, the user can then display the entire book, get a limited preview of the book, get a snippet preview (i.e., a few lines surrounding the relevant text), or get its author and title.

Rulings against Google in these cases could have particularly worrisome consequences for data reuse in scientific disciplines. The lawyers for the plaintiffs in both cases argue that Google must seek permission to copy and display any portion of the works still under copyright, arguing that Google's use of such works (scanning and displaying portions of the documents) without the authors' consent violates the publishers' right under Section 107(4) to protect the market value of those works (Goldstein 2005). Google's position is that the fair use provisions of Section 107 allow them to display small portions of books that are out of print but still under copyright (the so-called "snippet preview") without having to obtain authors' consent. Google points out that it will be impossible for them to request permission for a large percentage of these works since there is no clear contact information. A victory for the Author's Guild or AAP would not only undermine the utility of Google's service, but could also drastically limit the availability of data on the Web as a whole. The free and fair use of data is an essential part of the scientific enterprise generally, and of linguistics in particular. Reusing others' data, as we show in this paper, is common practice, and we see every reason that the practice should expand as we adopt more of the tools and benefits provided for us by the Internet. It would be onerous and likely impossible for permission to be obtained in most cases of data reuse, and to require it would have a chilling and counterproductive affect on our discipline. As Lessig (2004) observes, "The opportunity to create and transform becomes weakened in a world in which creation requires permission and creativity must check with a lawyer." A loss for Google will be a loss for us all.

- **Seek Permission when Appropriate**—Determine the limits of fair use, and the point at which permission should be sought from the authors or providers of source data.

4 The Principles of Reuse and Enrichment of Linguistic Data

In this section we present each of the Principles of Reuse and Enrichment of Linguistic Data. First, we define what we mean by *data*, *reuse* and *enrichment*.

- **Data** means any form of linguistic material, from phonetic transcriptions to highly enriched data that are accompanied by analyses. Generally, linguistically-relevant data include some form of analysis. Even transcriptions contain implicit analysis, since in transcribing sound recordings, the linguist must segment the input signal and map perceived sounds to phonetic symbols. Thus, we see a distinction between **raw language data** and **annotated language data**, after Bird & Liberman (2000), where the latter bundles the former with analyses. In the former category we include various kinds of audio and video recordings. In the latter category—the category, which, due to its form, we see as the most likely to be reused or republished—we include all forms of linguistic data and analyses, including transcriptions, interlinearizations, syntactic analyses (including, but not limited to, trees), phonological transformations, and even theory specific artifacts, such as optimality theoretic tableaux.

- **Reuse** means the republication of data from another source. Reuse involves taking an example datum from another source, and including it, with or without modification, in a new presentation.

- **Enrichment** simply means the inclusion of additional content or annotation to raw or annotated data.

4.1 Attribute fully

Providing full attribution is perhaps the most important principle we discuss. *De facto* standards in linguistics require the citation of any data that are reused, in particular, on an instance by instance basis. Thus, as shown in Section 1, the source for (1) is provided in each described document by a reference to Rudin (1988) and by the inclusion of the associated bibliographic entry.

Minimally, the same rules should apply on the Web. Any document posted to the Web that includes data from another source should include full citation information on that source and, if citing another on-line document, its URL. In addition to author, title and year, it is customary to include the source of the example and page numbers in a citation, which helps future users of the data to locate the source. Although not common practice, Walker & Taylor (1998)[6] recommends including the date of access or download for on-line bibliographic entries. Due to the transitory nature of documents on the Web, we see such a practice as a rudimentary means of versioning.

We envision, however, that this minimally compliant model can be extended further. We can think of the citation information as metadata, a means by which we can reconstruct the association of a particular datum with its source. Because of the well-defined and rich hyperlink structure on the Web, we have the capability of identifying not only the **provenience** of a particular datum, but its full **provenance**. A datum's provenience refers to where it is found on-line, i.e., the URL of the document containing it, while a datum's provenance is the full citation trail, starting with its URL and a citation chain leading back to the source. As long as each step in the trail contains citation metadata, the full provenance of a particular datum can be reconstructed. Thus, the full citation trailed exhibited in Karimi (1999) becomes the standard rather than the exception.[7]

Full data provenance gives us two things. First, if the date of collection is associated with each datum, we can establish the currency of the data we have discovered, just as the standard citation methodology establishes the year of publication for a particular document. Second, and more importantly, provenance can give us information about how data might have changed over time, and when and where changes were introduced. Changes can take two forms, either copying errors or purposeful changes. The latter variety occurs when a particular use requires reformatting or altering the data, perhaps due to a particular reanalysis, or the needs for a particular presentation (e.g., reformatting for display on a Web page). We will discuss the latter variety in more detail in Section 4.2. The former could occur when a linguist, or a tool that a linguist uses, inadvertently corrupts or alters the data accidentally during the act of copying. The reuse of (1) in Karimi is an example of just such a copying error. (2) is a verbatim copy of the Karimi example, which contrasts with Rudin (1988) by the absence of a diacritic. Although a rather harmless error, it does change content, and subsequent copies could preserve this error. In fact, without full provenance, we would not know where the error occurred. It could easily be the case that Karimi herself did not introduce the error, but rather Richards did in his dissertation (Richards 1997). Without full provenance, we would not be

[6] Walker & Taylor (1998) is an excellent resource for on-line citation style. We have adapted the practices outlined there in constructing the bibliography for this paper.

[7] We have the capability of discovering not only the source for a particular datum, but all instances of citation and reuse as well. Such discovery would be a little more difficult to accomplish, since citation records are unidirectional. However, an automated tool such as the Citeseer citation index (http://citeseer.ist.psu.edu/) (Lawrence *et al.* 1999) is a model of building a service that does just that. Citeseer brings order to the distributed nature of documents on the Web and the citation information they contain, allowing users to find where particular papers are cited, by whom, and even provides links to the papers themselves. Citeseer, however, only allows users to search the citation trail for documents, not for individual data reused within a document. A tool that provides the latter is not possible given the current structure of the Web nor standard practice for data citation.

able to establish this fact. As it turns out, the error was in fact introduced with Karimi (1999).

(2) Koj kogo vizda
Who whom sees
'Who saw whom?'

Although we see the example citation trail given in Karimi (1999) as the "best practice" for data citation, we recognize that it is not possible or realistic for an author to provide full provenance for every datum that he or she reuses. We also recognize that an author is relying on the accuracy and completeness of the citation information that is provided to him or her by the source document. In light of these facts, we recommend minimally that the source document be referenced, as is current standard practice, and, for the convenience of the reader and future potential users of the data, that the example number in the source be given if relevant or available.

4.2 Honor Author Intent

As noted earlier, linguistic data that appear in publication are not raw data; in nearly all cases the data have undergone some degree of annotation. As such, any linguistic data that are reused borrow something from the source annotation, even if only the transcription. The principle of honoring author intent that we outline here defines the limits that the source author's intent and analysis should impose on reuse. In other words, it defines the degree to which the author's original intent should be honored in the act of borrowing. We see the goal of this principle as a very simple one: no changes should be made to the author's original data and analysis that are not essential for the new analysis or presentation.

The Leipzig Glossing Rules (Bickel *et al.* 2004) define what we take to be standard practice in the borrowing of linguistic data, albeit specific to interlinearized texts: "A remark on the treatment of glosses in data cited from other sources: Glosses are part of the analysis, not part of the data. When citing an example from a published source, the gloss may be changed by the author if they [sic] prefer different terminology, a different style or a different analysis." The limits the Leipzig Glossing Rules define on altering analysis are clear: the analysis may be changed by the author if he or she *prefers* different terminology, a different presentation style, or even a wholly different analysis. This does, in fact, appear to approximate standard practice, especially with respect to IGT: if annotated data are altered, it is generally the analysis, as captured in the gloss line, that is changed. As an example of this, consider the following Choctaw example given by Whaley (1997:p. 48) but credited to Davies (1986), indicated by "(Adapted from Davies (1986:p. 14))":

(3) Hilha-li-tok
dance-1S-PST
I danced.

Whaley uses (3) to illustrate the phenomenon of pro-drop. In order to do this, however, she changes the original analysis to bring out the number feature of the verb. Consider Davie's original example, repeated here as (4):

(4) Hilha-**li**-tok.
dance 1NOM PST
I danced.

That is, Whaley changes the analysis of the *li* morpheme from *1NOM* to *1S*, where '*S*' refers to singular. The original example was meant to illustrate that certain predicate classes are signaled by nominative agreement, and thus glosses the *li* morpheme as *NOM*. Borrowed data are nearly always adapted to fit some new analysis, making alterations to the source analysis a necessary part of linguistic scholarship. However, it is linguistic scholarship that dictates that the changes be justified within this context. Thus, the change that Whaley makes to Davies' analysis was a necessary part of her analysis; without the change her larger analysis would not have made sense. What is clear in Whaley's changes is that they were not a matter of *preference* on her part, but rather, the changes she made were essential adaptations needed for her analysis. We take this as a crucial part of the current principle: alterations to the analysis should not be *ad hoc*, but rather should be motivated by the current analysis. The source author's intent is honored inasmuch as the new analysis permits.

It is also common practice to alter an analysis for the purposes of adapting it to a different theoretical framework, or where the markup vocabulary requires significant changes to the source. Such alterations can involve changes in terminology or the addition of analytical content. Note for instance the changes made by Richards (1995) to an example borrowed from Kroeger (1993), shown in (5) and (6), respectively. Since topicalization in Tagalog was central to his analysis, Richards abandons the markup vocabulary used by Kroeger in favor of that used in by Schacter (1976): *AT* for Actor-Topic, *T* for Topic, *A* for Actor, etc. Further, since arguments and control are also central to his analysis, he adds the dummy PRO in the adjunct clause *nang nagiisa*, and shows the co-index relations with the potential antecedents *Juan* and *hari*.[8]

(5) Bumisita si Juan$_i$ sa hari$_j$ [nang nagiisa PRO$_{i/*j}$]
AT-visited T Juan L king Adv AT-one
'Juan visited the king alone.'

(6) Bumisita si=Juan sa=hari nang nagiisa.
AV.PERF-visit NOM=Juan DAT=king ADV AV.IMPERF-one
'Juan visited the king alone. (Juan is alone)'

Richards also omitted the clitic delimiters ("=") used by Kroeger in his analysis. It is not clear in Richards' analysis whether he intended this modification—in other words, that he felt that *si* and *sa* were not actually clitics—or if the changes were made by accident. Although the removal of the delimiters was

[8] See also the notational change made by Martínez Fabián (2006) to Yuasa & Sadock (2002) in Figures 1 and 2 in the next section. Although the language data was changed, the notational differences are of a similar variety to those described here.

a less significant change than the others that he made, their removal lacks the legitimacy of the other changes since it does not contribute to his analysis, nor does it appear to be motivated by presentation constraints. Thus, we see it as a violation of the principle of honoring author intent.

Finally, authors sometimes change the granularity of the original analysis by removing or adding detail. Such changes are often motivated by the requirements of a specific analysis, where the data copied are adapted to that analysis. Consider (7) where Payne (1997:p. 118) has altered the Bella Coola example (8), from Fasold (1992:p. 84), which had included markup of the third person plural, *3PL*.

(7)	staltmx-aw	wa-ʔimlk
	chief-INTR	PROX-man

“The man is a chief.”

(8)	staltmx-aw	wa-ʔimlk
	chief(3PL.)(INTR.)	(PROX.)man

“The man is a chief.”

Thus, Payne reduces the granularity of the example, since the focus of his discussion is on the verbalizer; the associated *3PL* gloss was irrelevant, so he dropped it. While a perfectly reasonable omission in a scholarly work where brevity is important, we suggest an acknowledgment indicating that the example has been adapted, i.e., “adapted from Fasold (1992:p. 84)”.

4.3 Acknowledge Ownership

On the surface, there would appear to be little difference between the PRELDs Attribute Fully and Acknowledge Ownership. After all, a citation is an acknowledgment of ownership, and providing an attribution indeed defines that ownership. However, we cast the ownership issue as a separate principle precisely because ownership issues are not always clear, even when a citation is provided. Acknowledging ownership means acknowledging not only the intellectual property of the source, but also recognizing ownership of the changes and subsequent reanalyses. The question is: When linguistic data—either raw or annotated—are reused, who owns the copies? The tradition of including a citation record with copied data would suggest that the source author is the owner. However, as shown in the preceding sections, not all copies are made unaltered. Ownership of altered data, even if the source is acknowledged, is not clear. Here is an example cited by Mercado (2004:p. 101) to illustrate the point:

(9)	[Noong Lunes]$_i$	ay	ipinagbili	ng	mama	ang	kalabaw	niya	t$_i$.
	Last Monday	AY	buy.BT.CAUS.PRF	CS	man	SBJ	water buffalo	3SG.CS.CL	

‘The man sold his water buffalo last Monday.’

Example (9) was adapted from Schachter & Otanes (1972:p. 488) in which the original sentence contained the gloss *he* instead of *the man.* As another example, consider (10) given by Dalrymple & Nikolaeva (in press:p. 28):

(10) Marija zadumalas' ob ostavlennyx / *ostavlennom
Maria thought about left.behind.LOC.**PL** / left.behind.LOC.MASC.**SG**
muže i dočeri.
husband.LOC.**SG** & daughter.LOC.**SG**
'. . . Maria thought about her [husband and daughter] (who had been) left behind.'

Here, we see an adapted example in which the starred word, **ostavlennom*, was not present in the original from Corbett (1983:p. 20). These examples illustrate how the authors of secondary sources often change the data themselves, thus blurring the issue of ownership.

Ownership is clearer in the syntactic example borrowed by Martínez Fabián (2006:p. 175) shown in Figure 1. In this example, Martínez Fabián gives an analysis of a Yaqui sentence where, because Yaqui is OV, the tensed finite clause is in final position. Martínez Fabián (2006) cites Yuasa & Sadock (2002) as his source; the relevant example is shown in Figure 2. Although Martínez Fabián (2006) borrows the basic idea from Yuasa & Sadock (2002), the example contains notational variants, is missing the second tree, and has been adapted to language data from Yaqui as opposed to the original Japanese. Given these changes, most especially the last, and the fact that "ideas" are not subject to strict copyright laws (except possibly in cases of intellectual property rights), ownership of Figure 1 clearly rests with Martínez Fabián rather than with Yuasa and Sadock.

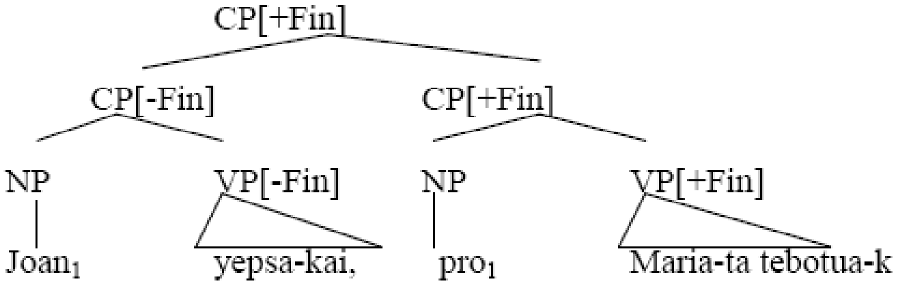


Figure 1: Yaqui example as presented in Martínez Fabián (2006)

Citations obviously indicate an acknowledgment of ownership, either of the data or the "idea" encapsulated in the analysis. Clearly, significant changes in data or analysis, or even a complete reanalysis, requires a careful examination of the resulting data, and a clarification of ownership. Martínez Fabián (2006) attempts to clarify ownership by including "adapted from Yuasa & Sadock (2002)" with the data in Figure 1. With this statement, he clearly acknowledges that he borrowed material from Yuasa & Sadock (2002), but at the same time indicates that he made alterations to it. Thus, we see his use of *adapted from* with the accompanying citation as a means to adhering to the good scientific practice of fully acknowledging the source, while at the same time providing the reader

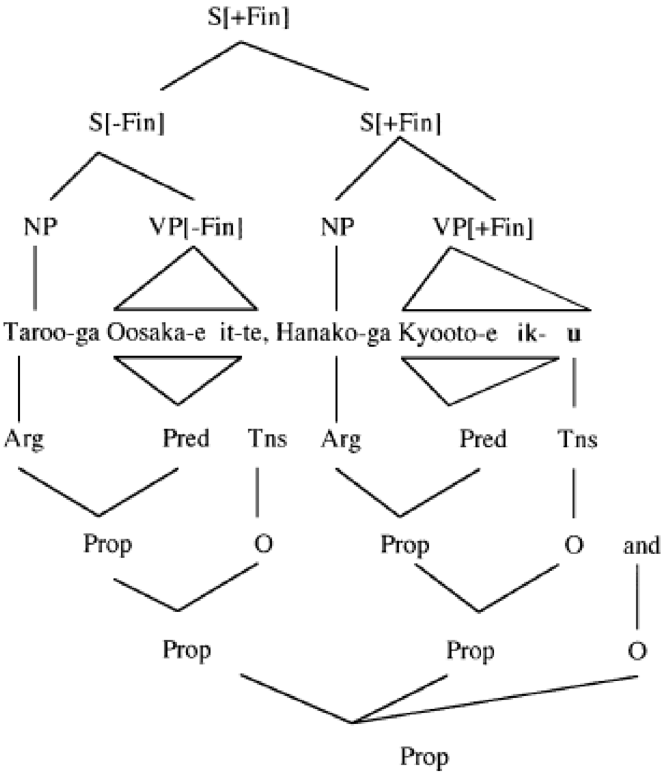


Figure 2: Japanese example as presented in Yuasa & Sadock (2002)

with information indicating new content and potentially new ownership.

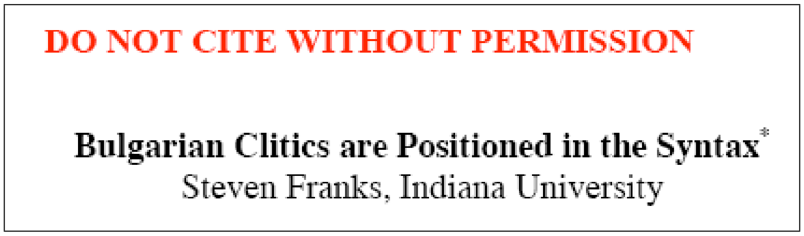
Citations are more than just a means for attribution, as discussed in Section 4.1: they are themselves acknowledgments of intellectual property and act as licenses for use. However, if data that are reused have undergone significant alterations from the source, good science requires that we acknowledge these changes, as Martínez Fabián (2006) has done above. But it should also be acknowledged that significant alterations to source data results in *new* data, hence new intellectual property, with the resulting change of ownership. It is important to recognize that such transformative uses of other's material are fair use and not violations of copyright, especially if value has been added (i.e., new analyses or changes in data) or the changes are substantial enough to suggest a different work (such as with Martínez Fabián (2006)). If substantial enough they may not even require acknowledgment. However, we feel that linguistics custom would be violated if such acknowledgment is not given.

Equally important as acknowledging the ownership of source data is making sure that your own data are citeable. In the digital environment, most specifically on the Web, it is common practice to post documents without clear titles, authorship or copyright information, making accurate citation difficult if not impossible (cf. Bird & Simons (2003)). Acknowledging the status of your own intellectual property is as important as acknowledging that of others. All documents posted to the Web that are meant for wider consumption should clearly identify ownership, and should provide a clear means for citation.[9]

4.4 Allow Data to be Sheltered

Honoring the intellectual property rights of a fellow researcher is important to doing good science, and important to the growth and prosperity of the field. In that light, language data, whatever its form (analyzed or raw), are often viewed as the intellectual property of the linguist who analyzed and disseminated it. It is important to recognize, however, that there are cultural sensitivies at play within indigenous language communities who sometimes consider data on their language as part of their cultural heritage (Barnhart 2002). Although copyright provisions do not extend to communal ownership—its provisions extend only to the intellectual property rights of a particular individuals—it is important for linguists to recognize the rights and sensitivities of indigenous communities, whether codified into law or not. Likewise, restrictions on access could originate from funding agencies or institutional review boards (Bird & Simons 2003). Further dissemination could even be forbidden, as part of binding ethics agreement with a funding institution or research body, for example, as with the DoBeS project (see Wittenburg (2005)). It is therefore reasonable and appropriate that certain data be "sheltered" or prohibited from use if that requirement is clearly stated with the data.

It is also possible that a linguist might wish to provide data and analysis to the linguistics community, but does not wish it to be reused or cited. In this event, it is common practice in the field to include an explicit note such as "DO NOT CITE" as part of the disseminated document. As an example, see Figure 3, a snapshot of the title taken from Franks (2005)[10]. Note that Franks clearly states "DO NOTE CITE WITHOUT PERMISSION" with the title of the document, and includes the same statement on each page of the document.



DO NOT CITE WITHOUT PERMISSION

Bulgarian Clitics are Positioned in the Syntax*

Steven Franks, Indiana University

Figure 3: Do not cite example from Franks (2005)

[9] See, for an example, the Ethnologue site (http://www.ethnologue.com/) where clear instructions for how to cite the website and materials contained within it are given, including a citation record itself (Raymond G. Gordon 2005).

[10] Permission to cite this paper was granted on 2006-May-25.

Because data are often placed on the Web for dissemination purposes, secondary use is often not considered. It is essential that information regarding how the data should be used or regarding restrictions on further dissemination or analysis be included with the data, just as clear information about citation should be included (as discussed in Section 4.3). Authors and data providers can also restrict access to documents and pages on the Web by adopting the Robots Exclusion Protocol, which we discuss in Section 5.4.

4.5 Seek Permission when Appropriate

It was indicated in Section 3 that, since most data reuse in linguistics is done for research and educational purposes, reuse in such a context constitutes fair use. On one hand, this may certainly be true for reuse where only a small portion of the source is copied. On the other, copying and distributing a *substantial* portion of a work would certainly not constitute fair use (Section 107, factor 3), even in the research and educational context. This is especially true if such use can be argued to affect the actual or potential market value of the work (Section 107, factor 4). Unfortunately, copyright law is noticeably vague in its definition of *substantial*, leaving its definition in practice to be decided either locally, e.g., through guidelines published by institutions and business entities, or through the results of litigation.

When the boundary of fair use is crossed, it is essential that permission be requested. Many academic institutions provide guidelines for dealing with the fair use provision of copyright law, and some even give specifics on what constitutes "substantial". A common measure is 10% or more, although in some cases, such as where the "creative core" of a work is copied, substantial could be far less than this figure.[11] In cases where it is unclear, it is important that local guidelines be consulted if for no other reason than to lessen the exposure to litigation.[12]

Nonetheless, there are cases where it is clear when permission should be sought:

- Where the intended use is for commercial purposes, as noted in Factor 1 of Section 107 of Copyright Law: Commercial uses include the sale or distribution of the material, or any uses that fall outside of the research or educational limitations.
- Where the source data are restricted or specific permission is required for use: Examples include on-line databases and resources that have clear license arrangements (such as those through the Linguistic Data Consortium, LDC), or clearly state that permission is required (such as with Franks (2005)).
- Where the intended use involves a substantial portion of the source work, where substantial is clearly a large portion of the source work, such as the work in its entirety, or in the case of linguistics, where a large number of linguistic examples are used.

[11] See, in particular, the provisions published by the University of North Carolina, Chapel Hill and the University of Washington.

[12] We will return to the discussion of fair use in the context of the Internet in Section 5.5.

5 PRELDs in Action: The World of Automated Linguistic Tools

In this section we address each of the PRELDs in light of automated tools that can search for and interoperate across documents and data on the Web.[13] We will address each of the PRELDs in turn and discuss their relevance with respect to automated tools while focusing on the responsibilities of tool developers to observe copyright law and linguistic custom.

The first generation of tools that allows search specifically for linguistic data on the Web already exists. The OLAC search facilities at LINGUIST List and the LDC are notable examples. With the OLAC search facility, users can locate documents and resources on thousands of the world's languages using either language names or Ethnologue language codes. LINGUIST List also provides supplementary search facilities beyond OLAC search, allowing users to locate additional information on languages, including websites, documents, resources, and even the names of linguists who work on specific languages. Further, this facility provides the means to search for resources on extinct languages and to search by language family.

Overall, these search tools allow users to locate resources or pages much like a library catalogue system, but do not operate at the level of data. When a resource is found, the user is provided with a URL to the resource which they can then explore further. Data-centric search facilities, however, are now starting to become available. A number of tools are being built whose intent is to harvest data from language-bearing documents on the Web. Of particular note is the Langgator project, a joint project among the LDC, the University of Melbourne, and Rosetta[14], which is being built to locate language documents and pages using sophisticated data aggregators and language identification algorithms. The documents and pages located will be searchable through OLAC.

Likewise, the Linguist's Search Engine (LSE)[15] allows users to search the LSE Web Collection, composed of a corpus of 3.5 million English sentences and a smaller corpus 100,000 aligned Chinese/English sentences, all collected from the Internet Archive[16]. LSE enables searches for specific syntactic structures across the collection, and includes a fairly sophisticated parser and syntactic tree query editor.

Another data-centric tool is the Online Database of INterlinear text, or ODIN (Lewis to appear)[17], which was designed specifically to find instances of interlinear glossed text (IGT) embedded in scholarly linguistic documents. Operating as an OLAC provider, ODIN allows users to locate documents and

[13] Our discussion in this section is noticeably Web-centric. However, our discussion applies to tools and uses for any linguistic data that exist in a large, distributed form, such as archives of linguistic documents (e.g., the Rutger's Optimality Archive, http://roa.rutgers.edu/), multilingual databases distributed on CD-ROM, etc.

[14] http://www.rosettaproject.org/

[15] http://lse.umiacs.umd.edu:8080/

[16] http://www.archive.org/

[17] http://www.csufresno.edu/odin/

resources by language name and code. In addition, it also provides a search facility that allows users to look within instances of IGT themselves, and provides a search vocabulary normalized to a common form, namely to concepts defined in the General Ontology of Linguistic Description (GOLD) (Farrar & Langendoen 2003)[18]. For example, if a researcher wished to look for instances of past tense morphemes contained in IGT, he or she could specify the GOLD concept *PastTense* as the query term, rather than the myriad of notational variants used in IGT (e.g., *Past*, *PST*, *RemotePast*, *HodiernalPast*, etc.). Additional search facilities include the ability to look for linguistically salient constructions contained within language data, constructions such as passives, conditionals, possessives, etc., all of which may or may not be directly encoded in the source examples. Future facilities will include search over associated text in addition to contents of IGT examples themselves.

On the one hand, the move towards data-centric tools from those that operate at the document or page level does not require any changes to tool design. For example, search results that resolve to URLs require little on the part of tool developers with respect to fair use. As long as the contents of the documents or pages referenced by the URL remain external to the tool or to the site that houses the tool, no questions of fair use are relevant. However, tools that display data extracted from pages, or allow manipulation of that data, start to move into an area where fair use comes into play. Just as linguists must observe fair use principles and linguistic custom, so too should tool developers. In other words, tool developers should be beholden to the same principles of use and enrichment that govern linguists. Thus, more savvy, data-centric tools will ultimately require a different design with respect to fair use.

In the remaining sections we discuss each of the PRELDs and how they affect tool development, specifically that of data-centric tools. Since the PRELDs affect the design of tools in specific ways, we ground our discussion in a specific, albeit hypothetical, tool—a fictitious linguistics search engine that allows the user to locate linguistic data embedded in other documents. Our discussion, then, focuses on the specifics of how the tool would be affected by each of the PRELDs. We call the fictitious search engine the Linguistics Query Program (LQP), which can be viewed as a prototype for tools in general, not just for search engines, since the PRELDs would likely have similar affects on the design of other types of tools. For the sake of brevity, we will not discuss all the features and design characteristics of a search engine, but rather concentrate on the issues relevant to how the PRELDs would affect the design and output of such a tool.

5.1 Attribute Fully

In scholarly linguistic discourse, the manner of citation usually involves the inclusion of the author(s) and year of the source document, which is included in the vicinity of the data that are borrowed. This citation record is co-indexed with a full citation record included as part of the bibliography. In the Web

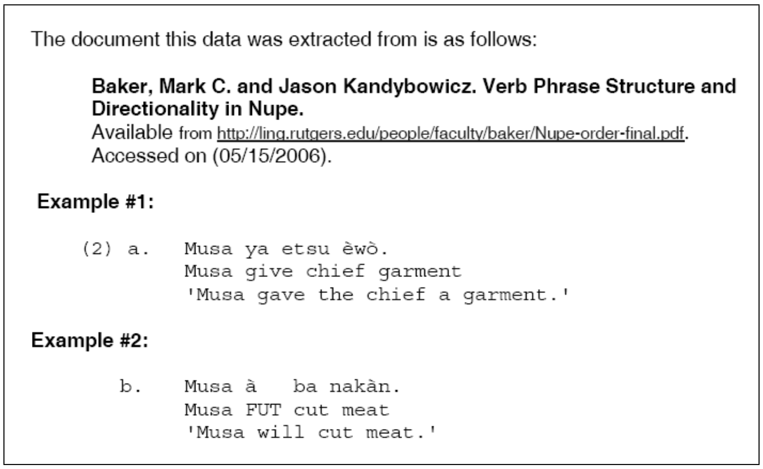
[18] http://www.linguistics-ontology.org

environment, the tradition and format can be modified somewhat, since the manner of display may not be the same as in a scholarly document, nor might the pertinent material required for a full citation be available. However, in the design of a tool, every reasonable effort should be made to include as much citation information as possible with each datum that is reused. If a linguistic search engine provides the facility to extract and examine data from documents discovered on the Web, the citation information from these documents should be provided with the data. It is only in this manner that the linguist who uses the tool can appropriately cite the source. And it is only in this manner that the owner of the data is given his or her due. Should a linguist fail to adequately cite source documents, it is his or her own failing, not the failing of the search engine or tool used to discover the data.

We prioritize the kinds of information that can be harvested from source pages and be included in a citation. These are labeled as to what is required in all cases, required if available, and what would be desirable to include if available:

1. Source URL—required
2. Date source was accessed—required
3. Source author(s)—required if available, but declare "not available" where appropriate
4. Source title—required if available, but declare "not available" where appropriate
5. Year—if available and easily harvested
6. Where else published (journal, volume, etc.)—if available and easily harvested

If we use the LQP to search for data on Nupe, a Niger-Congo language spoken in Nigeria, it is likely that it would find the paper *Verb Phrase Structure and Directionality in Nupe* written by Baker & Kandybowicz (2003). This paper contains numerous examples of Nupe language data, most in interlinearized form. If the LQP had the capability of extracting these examples and could display them as part of the results from a query, we would want to include as much of the attribution information as contained in the Web copy of Baker & Kandybowicz (2003) as possible. The Web copy of the paper lists the title and the names of both authors, so these would be fairly easy to extract. The year is not listed anywhere in the paper, so we cannot reasonably be expected to provide it. Although the paper did eventually appear in an anthology, we cannot know this from the Web document, and therefore cannot reasonably be expected to provide it either. Since the URL and date of access are given to us for free, we can include these in the citation record as well. We can thus display a fairly complete citation record with the source data as shown in Figure 4.



The document this data was extracted from is as follows:

Baker, Mark C. and Jason Kandybowicz. Verb Phrase Structure and Directionality in Nupe.
Available from http://ling.rutgers.edu/people/faculty/baker/Nupe-order-final.pdf.
Accessed on (05/15/2006).

Example #1:

```
    (2) a.   Musa ya etsu èwò.
             Musa give chief garment
             'Musa gave the chief a garment.'
```

Example #2:

```
        b.   Musa à   ba nakàn.
             Musa FUT cut meat
             'Musa will cut meat.'
```

Figure 4: Sample output from an LQP query for the language Nupe with data taken from Baker & Kandybowicz (2003)

Furthermore, it may be possible that LQP will discover example data that have no citation information available. Perhaps the material discovered is a chapter from a book where the author and title are not identified locally, or perhaps it is a Web page used for classroom instruction on a particular language. Also, it could easily be the case that our tool just cannot discern who the author or what the title is. We can recover the URL and the date of access easily, and can provide access to the document via its URL. Displaying any data from this document without additional citation information is problematic, however. It could be argued that since no additional information could easily be recovered, displaying just the URL and date of access with extracted data is the most we could be expected to do. However, since automated tools for extacting data from free text are fallible, there is no way in an automated environment to ensure that the failure to find relevant citation information is a failure of the tool, or that it results from the absence of the relevant citation information in the source document. Since displaying another's data without attribution is clearly a violation of linguistic custom as encapsulated in the PRELD Attribute Fully, the tool developer should weigh carefully the precision of his tool before displaying data that are not fully cited. In other words, in the absence of near certainty, it is probably best to display just the URL and date of access, rather than to overstep the bounds of fair use.

5.2 Honor Author Intent

The PRELD of Honor Author Intent requires that no unjustified changes be made to borrowed data. In linguistic scholarly discourse, changes that are made to copies are usually done based on the requirements of a particular presentation or analysis. An author might require, for instance, that a particular borrowed datum be modified to suit a particular line of argumentation. In some cases, changes to the source are so significant that it is difficult to determine what similarities exist between the source and its copy (as with the Martínez Fabián (2006) example discussed in Section 4.2).

With automated tools, however, even more significant changes can be made to the source, either to accommodate certain display requirements or to enrich content to facilitate other types of search. Further, whatever changes that are made can be done *en masse*, affecting not just isolated examples, but all the data that are displayed or manipulated by the tool. Still, it is important in an automated environment to preserve the source author's original analysis, which should be recoverable from the modified output.

Suppose the LQP can discover IGT in queried documents, as illustrated earlier. For purposes of retrieving specific morphemes and their glosses, it is necessary to align the language line and the gloss line so that the language morphemes align correctly with their glosses. In the standard IGT presentation, this alignment is implicit: linguists decipher the alignment by counting delimiters (hyphens, spaces, tabs) in both the language and gloss lines, a task that would be time consuming for an automated tool to do online. Facilitating faster search might involve re-encoding the discovered data into a more structured form, say in XML, and providing the facility either to display the XML in its raw form, or providing scripts that render it into some common IGT display format. An XML data format, such as that proposed in Hughes *et al.* (2003)[19], would look nothing like its source. An example XML representation of the Japanese example in (12) from Ogihara (1998) is shown in (11). However, using XSL stylesheets, it is possible to render the XML in (12) into a form that is either identical or very similar to (11).

(11) Taroo-wa ima ie-o tate-te iru
Taro-Top now house-Acc build-TE IRU-Pres
'Taro is now building a house.'

(12)

```
<exampledoc>
   <sourceURL>http://faculty.washington.edu/ogihara/papers/teiru.pdf</sourceURL>
   <sourceAuthor>Ogihara, Toshiyuki</sourceAuthor>
   <sourceTitle>The Ambiguity of the -te iru Form in Japanese</sourceTitle>
   <example id="1131">
      <language code="JPN">Japanese</language>
      <interlinear-text>
         <phrases>
            <phrase>
               <item type="gloss">'Taro is now building a house.'</item>
```

[19] The XML example we show here is truncated due to space constraints. We also deviate somewhat from the model described in Hughes *et al.* (2003), and include additional content beyond that shown in (11). Note, in particular, the addition of enriched content, namely the new attributes *pos* and *morphtype*, and their respective values: *pos* marks the likely part of speech for the corresponding word or morpheme, and *morphtype* indicates the type of morpheme, (e.g., suffix, prefix or root).

```
<words>
   <word>
      <morphemes>
         <morph>
            <item type="text">Taroo</item>
            <item type="gloss" pos="noun" morphtype="root">Taro</item>
         </morph>
         <morph>
            <item type="text">wa</item>
            <item type="gram" morphtype="suffix">Top</item>
         </morph>
      </morphemes>
   </word>
   <word>
      <morphemes>
         <morph>
            <item type="text">ima</item>
            <item type="gram" morphtype="root">now</item>
         </morph>
      </morphemes>
   </word>
   <word>
      <morphemes>
         <morph>
            <item type="text">ie</item>
            <item type="gloss" pos="noun" morphtype="root">house</item>
         </morph>
         <morph>
            <item type="text">o</item>
            <item type="gram" pos="" morphtype="suffix">Acc</item>
         </morph>
      </morphemes>
  </word>
  ...
```

If we use the LQP to discover information not directly encoded in the language data, but discernable from it, we might add additional content to the instances of data that are discovered. For example, even though it is possible to deduce the boundaries between morphemes using the standard morpheme delimiters, it is not possible to determine from any IGT presentation what are roots and what are affixes, nor is it possible to further delineate prefixes from suffixes. If we build a tool that pre-processes each instance of IGT that we discover and determines morpheme type for us, we can add this additional information to the content of the example we display (as shown in 12), and allow users to include this information in their queries. The additional content we add in this manner should be clearly identified in the presentation or accompanying materials.

5.3 Acknowledge Ownership

If we build our search engine to add significant content to data, as described in Section 5.2, it is important that users of our tool recognize what additional content and modifications our tool made. This is important not just for attribution, but also necessary in the event our tool introduces error, most especially if others make use of the potentially flawed data. It is also important to recognize our own intellectual property in the modifications our tool makes. In addition to documenting the additional content, we should include citation information that points to the LQP and its website. Following and adapting from Walker & Taylor (1998), we recommend a citation record that looks like (13). This citation record recognizes the source, Baker & Kandybowicz (2003), while at the same time identifies our tool's additional contribution. Full data provenance is preserved, and this citation information is available to subsequent users of the data.

(13) Baker, M. & Kandybowicz, J. (2003) Verb Phrase Structure and Directionality in Nupe. Linguistics Query Program. http://lqp.washington.edu/query.php?doc=1101 (May 15, 2006).

5.4 Allow Data to be Sheltered

In rolling out the LQP, it is possible we will discover authors and resource providers who do not wish their data to be accessed. It is perfectly reasonable for them to restrict access to their data, and we should accept any restrictions without question. Fortunately, there is an automatic means to restrict access to Web sites: if a Web host wishes to prevent access to his or her site and any material contained on the site, all that needs to be done is to follow the **Robots Exclusion Protocol**[20]. The Robots Exclusion Protocol is recognized by most search engines and basically involves creating a special file, at a designated URL, called `robots.txt`. This file contains the following information: an indication of who can and cannot access the site and what they can and cannot do on the site. In the design of the LQP we will want to ensure that we observe this protocol, and design our tool to look for the robots.txt on any website we access. The example robots.txt file in (14) disallows the Google Image Crawler from crawling the site altogether, and restricts access for all others to anything on the site outside the paths `/pdfs` and `/cgi-bin`.

(14)

```
User-agent: *
Disallow: /pdfs/
Disallow: /cgi-bin/
User-agent: Googlebot-Image
Disallow: /
```

[20] The protocol is described in detail at http://www.robotstxt.org/. Also see Wong (1997) for detailed instructions on how to create a `robots.txt` file.

We will also want to create a means whereby researchers can *retract* examples that they no longer wish to have accessed. Perhaps our search engine locates syntactic trees, and finds examples of Government and Binding style trees in several documents by a particular author. The author later discovers these while using the tool, and indicates that he prefers us use more up-to-date trees contained in other documents on the same site. It is important that we provide a means for this kind of retraction, whether that facility is provided as a manual or automated service.

Since the Web is constantly changing, and pages appear and disappear often, dead links present another challenge for any kind of tools that searches over the Web. There are no rules as to what should be done with data discovered on pages that are no longer active, and there are no real-world equivalents to dead links—except, perhaps, lost or destroyed manuscripts to which all that remain are references. We could treat dead links as retractions, but that is probably excessive since there are many reasons a link could go dead. And, many have nothing to do with whether a particular author wants his or her material accessed. We recommend that tools be designed to recognize when links have gone dead. Further, it is relatively straightforward for tools to search for moved documents and pages, thus repairing deadlinks by referencing updated URLs. In light of the fact that there are no clear rules for dealing with deadlinks, we leave the details of implementation to the individual tool developers.

5.5 Seek Permission when Appropriate

Automated tools present a challenge to the attribution and fair-use principles thus far described, especially since they could conceivably extract and display large collections of data from source documents. An example of such a scenario is described in Section 2.2, in which a potential query extracts 853 examples of interlinear text examples on Hausa from Abdoulaye (1992). Copyright law notwithstanding, fair use gets stretched thin in such a scenario, even if the data are fully cited, and stretched even thinner if the entire document is copied. Would it be acceptable for a linguist to do the same thing, essentially to republish a large portion of someone else's work, even with adequate citation? The answer to the latter question is likely no. But, here we think a compromise should be made in the electronic realm. Although we can see the tools of the future as *users* of data, just as much as human linguists are currently, we should also recognize that they are not themselves linguists. *They are tools that serve linguists.* If a linguist jots down or photocopies a large number of examples from Abdoulaye's dissertation with the intent of using just a small subset of these examples in his own work at some point in the future, should he be prevented from copying more than he needs just in case? Likewise, if a search engine returns far more examples than any given linguist would likely need, should the results be trimmed back in the name of fair use? The results that are returned are transitory and ephemeral; they exist only for a moment separate from their source. In this case, they should only be viewed as results returned by a query, and should not themselves be viewed as a publication. They are merely one step towards some future publication.

That said, however, given the highly public nature of content on the Web, we suggest tempering the output of particular tools, especially in cases where the output represents a significant portion of the source (we can follow the 10% rule here, as described in Section 4.5). In such an event, it may be more constructive to provide a link to the source document, rather than to redisplay a large portion of its content. Likewise, users could be asked to refine their query such that only the specific examples relevant to the refined query are displayed, thus reducing the number of examples presented to the user. In any event, if a given query would result in the display of a signficant portion of any document, it is probably best to seek the permission of the source's owner. In the absence permission in such a scenario, displaying the source URL and relevant citation information is probably the best course of action.

6 Conclusion

In this document, we have discussed a number of issues relevant to the fair use of linguistic data, especially in the context of the Web. The manner in which linguists can do their work is changing dramatically, and as more data makes its way to the Web, the more dramatic these changes will become. We envision a future where linguists will be able to query and interoperate over data from thousands of languages, and where queries based on linguistically relevant phenomena inferred from source data will be possible. However, crucial to such a future environment is the continued observation of the basic principles of fair use, as determined by copyright law and linguistic custom, which we have codified into the Principles of Reuse and Enrichment of Linguistic Data.

References

Abdoulaye, Mahamane L., 1992. *Aspects of Hausa morphosyntax in Role and. Reference Grammar*. State University of New York dissertation. http://linguistics.buffalo.edu/people/students/dissertations/abdoulaye/hausadiss.pdf (2006-May-17).

Bailyn, John F., 2003. MGU spec-kurs: Russian syntax (course materials). http://www.sinc.sunysb.edu/Clubs/nels/jbailyn/WH.MGU.2003.pdf (2006-May-17).

Baker, Mark, & Jason Kandybowicz. 2003. Verb phrase structure and directionality in nupe. In *Linguistic Typology and Representation of African Languages, 1-22*, ed. by John M. Mugane. Africa World Press.

Barnhart, Leslie, 2002. Intellectual property rights: A reference paper for protecting language and culture in the digital age. www.indigenous-language.org/files/intellectual_property_issues.pdf (2006-May-17).

Bickel, Balthasar, Bernard Comrie, & Martin Haspelmath. 2004. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses (revised version). Technical report, Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig. http://www.eva.mpg.de/lingua/files/morpheme.html (2006-May-17).

Bird, Steven, & Mark Liberman. 2000. A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Computer and Information Science, University of Pennsylvania.

——, & Gary F. Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79.

Corbett, Greville G. 1983. *Hierarchies, Targets, and Controllers: Agreement Patterns in Slavic*. London: Croom Helm.

Dalrymple, Mary, & Irina Nikolaeva. in press. Syntax of natural and accidental coordination: Evidence from agreement. *Language* .

Davies, William D. 1986. *Choctaw Verb Agreement and Universals*. Dordrecht: D. Reidel Publishing Company.

Dayal, Veneeta. 2003. Multiple wh questions. In *Syntax Companion 3*, ed. by M. Everaert & H. van Riemsdijk, chapter 44. Oxford: Blackwell Publishers. http://www.rci.rutgers.edu/∼dayal/Chapter44.PDF (2006-May-17).

Dryer, Matthew S. 2006. Noun phrase structure. In *Complex Constructions, Language Typology and Syntactic Description, Vol. 2.*, ed. by Timothy Shopen. Cambridge University Press, 2nd edition. http://linguistics.buffalo.edu/people/faculty/dryer/dryer/DryerShopenNPStructure.pdf (2006-May-17).

Farrar, Scott, & D. Terence Langendoen. 2003. A linguistic ontology for the Semantic Web. *GLOT International* 7.97–100. http://www.u.arizona.edu/∼farrar/papers/FarLang03b.pdf.

Fasold, Ralph W. 1992. *The Sociolinguistics of Language: Introduction to Sociolinguistics*, volume II. Oxford: Blackwell, 15th edition.

Franks, Steven, 2005. Bulgarian clitics are positioned in the syntax. http://www.cogs.indiana.edu/people/homepages/franks/Bg_clitics_remark_dense.pdf (2006-May-17).

Goldstein, Michael. 2005. Google's literary quest in peril. *Intellectual Property and Technology Forum at the Boston College Law School* 110301. http://www.bc.edu/bc_org/avp/law/st_org/iptf/articles/content/2005110301.html (2006-October-29).

Hughes, Baden, Steven Bird, & Cathy Bow. 2003. Interlinear text facilities. In *E-MELD 2003*, Michigan State University. http://emeld.org/workshop/2003/baden-demo.html (2006-May-17). See also http://www.cs.mu.oz.au/research/lt/emeld/interlinear.

Karimi, Simin. 1999. Is scrambling as strange as we think it is? *MIT working papers in Linguistics* 33.159–189. http://minimalism.linguistics.arizona.edu/AMSA/PDF/AMSA-172-0900.pdf (2006-May-17).

Kroeger, Paul. 1993. *Phrase Structure and Grammatical Relations in Tagalog*. Stanford: CSLI Publications.

Lawrence, Steve, C. Lee Giles, & Kurt Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer* 32.67–71. http://citeseer.ist.psu.edu/aci-computer/aci-computer99.html (2006-May-17).

Lessig, Lawrence. 2004. *Free Culture*. The Penguin Press.

Lewis, William D. to appear. ODIN: A Model for Adapting and Enriching Legacy Infrastructure. In *Proceedings of the e-Humanities Workshop*, Amsterdam. Held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing.

Liberman, Mark. 2000. Legal, ethical, and policy issues concerning the recording and publication of primary language materials. In *Proceedings of the workshop on web-based language documentation and description*, ed. by Steven Bird & Gary Simons. http://www.ldc.upenn.edu/exploration/expl2000/ (2006-May-17).

Martínez Fabián, Constantino, 2006. *Yaqui Coordination*. University of Arizona dissertation. http://dingo.sbs.arizona.edu/∼langendoen/martinez/yaqui-coordination.pdf (2006-May-17).

Mercado, Raphael. 2004. Focus constructions and wh-questions in tagalog: A unified analysis. *Toronto Working Papers in Linguistics* 23.95–118.

Ogihara, Toshiyuki. 1998. The ambiguity of the *-te iru* form in Japanese. *Journal of East Asian Linguistics* 7.87–120. http://faculty.washington.edu/ogihara/papers/teiru.pdf (2006-May-21).

Payne, Thomas E. 1997. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge, U.K.: Cambridge University Press.

Raymond G. Gordon, Jr. (ed.) 2005. *Ethnologue: Languages of the World*. Dallas: SIL International, 15 edition.

Richards, Norvin, 1995. Another look at tagalog subjects. Manuscript, MIT.

——, 1997. *What Moves Where When in What Language?*. MIT dissertation. https://dspace.mit.edu/handle/1721.1/10236 (2006-May-17).

——. 1999. Dependency formation and directionality of tree construction. *MIT Working Papers in Linguistics* 33. http://web.mit.edu/norvin/www/papers/top-down.pdf (2006-May-17).

Rudin, Catherine. 1988. On multiple questions and multiple wh-fronting. *Natural Language and Linguistic Theory* 6.445–501.

Schachter, P., & F.T. Otanes. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.

Schacter, Paul. 1976. The subject in Philippine languages: topic, actor, actor-topic, or none of the above? In *Subject and Topic*, ed. by Charles Li, 491–518. New York: Academic Press.

Simons, Gary, & Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing* 18.117–128. http://www.arxiv.org/abs/cs.CL/0306040 (2006-May-17).

Stjepanovic, Sandra. 2003. Multiple wh-fronting in serbo-croatian matrix questions and the matrix sluicing construction. In *Multiple Wh-fronting*, ed. by Cedric Boeckx & Kleanthes Grohmann. Amsterdam: John Benjamins. http://people.umass.edu/roeper/711/STJEPANOVIC.pdf (2006-May-17).

Walker, Janice R., & Todd Taylor. 1998. *The Columbia Guide to Online Style*. New York: Columbia University Press.

Whaley, Lindsay J. 1997. *Introduction to Typology: The Unity and Diversity of Language*. Thousand Oaks, CA: SAGE Pulications.

Wittenburg, Peter, 2005. Legal and ethical documents (dobes-led-v1). http://www.mpi.nl/DOBES/ethical_legal_aspects/le-documents-v1.pdf (2006-May-17).

Wong, Clinton. 1997. *Web Client Programming with Perl: Automating Tasks on the Web*. Sebastopol: O'Reilly. http://www.oreilly.com/openbook/webclient/ (2006-May-21).

Yuasa, Etsuyo, & Jerry M. Sadock. 2002. Pseudo-subordination: a mismatch between syntax and semantics. *Journal of Linguistics* 38.87–111.