Opening plenary

Doing linguistics in the 21st century:

Interoperation and the quest for the global riches of knowledge

Gary F. Simons
SIL International and
Graduate Institute of Applied Linguistics

This paper gives the full content (with added references) of the opening plenary address for the TILR workshop. It follows the presentation slides which are posted at http://linguistlist.org/tilr/papers/TILR Plenary Slides.pdf; they should be consulted to see the referenced diagrams. In this paper, numbers in square brackets provide synchronization with the slides; thus, [n] *marks the end of the discussion of slide* n *and is the point at which the presentation should be advanced to the next slide.*

Abstract

Doing linguistics can be likened to a quest for riches—the riches of knowledge about language in general and about thousands of languages in particular. During the 20th century, linguists were limited in that quest to resources in their local library, to data collected in their own field work, and to collaboration with personal acquaintances. The Internet is changing all that. During the 21st century, the practice of linguistics could cast off the limits imposed by countless, fragmented hoards of local knowledge and begin to

exploit the riches of a shared, ever-growing, collaboratively-developed, integrated treasury of global knowledge. But this transformation can happen only if the community agrees to work with shared standards and thereby embrace the practice of interoperation.

The presentation begins by offering an illustration from everyday life of how standards work and how they will help our community to interoperate. It then develops a vision for interoperation within the language resources community in the 21st century that is expressed as a twelve-point prescription for a cyberinfrastructure for linguistics. The first eight points—aggregator, metadata standard, submission protocol, harvesting protocol, gateways, spiders, subcommunities, and datum-level search—deal with building an infrastructure for amassing and searching the global riches of linguistic knowledge. The final four points—open repositories, collaboration with amateurs, matching supply with demand, and measuring authority and reputation—deal with exploiting the web as a new playing field for global collaboration to partner with the speakers of the world's languages in order to actually fill that treasury of knowledge.

About the author

Gary F. Simons currently holds the position of Associate Vice President for Academic Affairs with SIL International (Dallas, TX). He is also Adjunct Professor of Language Development at the Graduate Institute of Applied Linguistics (Dallas, TX). He has recently contributed to the development of cyberinfrastructure for linguistics as co-founder of the Open Language Archives Community (http://www.language-archives.org/), co-developer of the ISO 639-3 standard of three-letter identifiers for the known languages of the world (http://www.sil.org/iso639-3/), and executive editor of the *Ethnologue* (http://www.ethnologue.com/). Before taking up his current role in 1999, he served 15 years as director of SIL's Academic Computing Department where he oversaw the development of language software tools like *IT* (the Interlinear Text processor) and *LinguaLinks*. Prior to that he did linguistic fieldwork with SIL in Papua New Guinea (1976) and Solomon Islands (1977-1983). In 1979, he received a Ph.D. in general linguistics from Cornell University (with minor emphases in Computer Science and Classics).

1. Introduction

Doing linguistics can be likened to a quest for riches—the riches of knowledge about language in general and about thousands of languages in particular. [2] During the 20th century, linguists were limited in that quest to resources in their local library, to data collected in their own field work, and to collaboration with personal acquaintances. The Internet is changing all that. During the 21st century, the practice of linguistics has the potential to cast off the limits imposed by countless, fragmented hoards of knowledge stashed away in local storage silos and to begin exploiting the riches of a single, ever-growing, collaboratively-developed, and integrated treasury of global knowledge that is open to all. In a word, this is the transformation from an economy of *scarcity* to an economy of *abundance*. [3]

This notion of an economy of abundance is a theme that is being taken up by authors who are highlighting the fundamental changes that the Internet is bringing to the world of business (Anderson 2006:18*ff.*; Wang 2006:11). In the 20th century, business operated in a physical world in which there was finite shelf space. Today, business can operate in a digital world where there is virtually infinite shelf space. In the physical world, there was significant cost to produce inventory along with limited capacity to carry inventory. In the digital world, there is little cost to reproduce inventory with virtually limitless capacity to carry inventory. In the physical world, publishers controlled content, while in the digital world users generate content. Because of all the constraints of the physical world, there was limited choice and suboptimal matching of supply with demand. Now, when one goes to a web-based business like Amazon or Rhapsody or Netflix or eBay there is seemingly unlimited choice along with optimal matching of supply and demand. [4]

But abundance has its dark side. The problem with unlimited choice is that it breeds overwhelming confusion (Wang 2006:28). The way to address this problem as our community embraces the economy of abundance is for us to voluntarily adopt constraints that will limit our choices and thereby promote interoperation within the community.

Interoperability, which is the theme of our workshop, can be defined as the ability for two or more systems to exchange information or services and for each to make satisfactory use of what is exchanged. [5]

This solution statement encapsulates an overview of this talk. With respect to adopting constraints that will limit our choices, I begin by offering an illustration from everyday life of how standards work and how they will help our community. With respect to promoting interoperability, I then propose a vision for interoperation within the language resources community in the 21st century; it is expressed as a twelve-point prescription for a cyberinfrastructure for linguistics. [6]

2. Understanding standards

Standards play an indispensable role in the interoperation we take for granted in everyday life. Why then are standards such a hard sell for academics? In pondering this problem, I've been inspired by George Lakoff's (1995) work on metaphor and its application to contemporary politics, in which he points out that conservatives (with their "strict father" model) have gotten the upper hand over liberals (with their "nurturant parent" model) through the deft use of metaphor in the public discourse. How might metaphor be used to inform the discourse about standards among linguists? What metaphors might cast standards in a nurturant light rather than a strict one?

I've identified two that might work. The first is "linguistics as community," which highlights the role of standards in allowing linguists to function as a community. The second, "development as freedom," builds on Nobel economist Amartya Sen's (1999) book by the same name, and highlights the role of standards in freeing the community to develop the riches of knowledge it is seeking. These ideas are developed in an essay (Simons, forthcoming) written to honor the career of Terry Langendoen with whom I've collaborated since 1989 in developing information standards for linguistics. [7]

One passage from that essay I will share here. It offers an analogy from everyday life that serves well to illustrate what we are trying to achieve in establishing standards for language resources. It has to do with the standardization of time:

> In medieval times (and earlier), time was regulated by the position of the sun in relation to the individual's position on earth. Noon was defined as the moment when the sun was directly overhead. This standard worked fine as long as wind and muscle power constrained the distance that could be traveled in a day. But the advent of train travel in the nineteenth century changed all of that. In the most populated latitudes of North America, the earth rotates at a rate of twelve and a half miles a minute (Blaise 2000:30). Thus rail passengers could journey 100 miles in a couple hours, only to find that their pocket watches were eight minutes off when they arrived.
>
> Today it is hard for us to imagine life without standard time, but just 150 years ago in North America there were 144 official times based on local solar noon (Blaise 2000:34). The rail network grew up in this context; each railroad set and published its schedules in terms of the official time of its headquarters, rather than of the city in which the train was stopping. Thus in a station that serviced more than one railroad, it was simultaneously a different time on each road (Blaise 2000:70). Rail passengers had to travel with a big book of time conversions in order to plan their connections, and it was still easy to miscalculate and miss the train. What's worse, sharing the same tracks among railroads employing different official times could lead to disastrous results—train wrecks were a daily occurrence (Blaise 2000:72).
>
> These problems were finally solved in 1884 when the nations of the world gathered at the Prime Meridian Conference. In addition to establishing the zero meridian at Greenwich, this conference established the International Date Line and the system of universal time with 24 standard time zones stretching around the globe. For the first time in history, it was possible to answer the question "What time is it?" with a single global answer, rather than with a myriad of local answers. [8]

The chart on slide 9 summarizes these two approaches to local time. With solar noon, local time is a continuous function; there are an unlimited number of noons. With time

zones, local time is a step function; there are exactly 24 noons. As far as interoperation, any east-west travel under the solar noon approach requires time conversion by a number of minutes to be looked up. By contrast, the time zone approach converts only by increments of hours and such conversion is required only when traveling over a long distance. The solar noon approach is optimized for convergence with physical reality (which is, incidentally, one way of defining the notion of *truth)*. As a result, interoperation is possible, but it is very cumbersome and error prone. The time zone approach, on the other hand, is optimized for interoperation and this is the system of temporal reckoning that we now take for granted. [9]

The challenge for linguists is that in order to achieve interoperation, we must go beyond establishing the local reality of our work and begin, as well, to place our work in the right "time zone." One example of a system of linguistic time zones is the ISO 639-3 standard for the identification of languages. For instance, the identifier [eng] does not represent a precise point in linguistic space; rather, it is a zone covering all the local varieties of English. Another example of time zones for linguistics is GOLD, the General Ontology for Linguistic Description (Farrar and Langendoen 2003; see also http://www.linguistics-ontology.org/). For instance, the GOLD identifier PastTense is a zone covering all local varieties of past tense. Thus annotating something as PastTense is not saying that it is exactly the same as everything in all other languages that have been labeled in that way, but just that it is in the same zone as the others for purposes of cross-linguistic search and comparison. [10]

Recall that the two main perils of train travel in the absence of temporal interoperation were the missed train and the train wreck. These same perils plague travel in cyberspace. Users miss the information train altogether when the same thing has many different local names; in this case, a query in terms of one name fails to retrieve all the relevant resources that use one of the other names. Users experience an information wreck when many different things have the same local name; in this case, a query using that name returns what the user is really looking for plus the co-mingled results for all the other things that have the same name.

More than thirty years ago, Joseph Grimes encountered these problems of the missed train and the train wreck when he was developing a database for cataloging the world's known living languages. As described in the final report for the project, his solution was to develop a standard for identifying languages:

> Each language is given a three-letter code on the order of international airport codes. This aids in equating languages across national boundaries, where the same language may be called by different names, and in distinguishing different languages called by the same name. (Grimes 1974:i)

That scheme has served as an in-house standard within SIL International ever since its development (Simons 2002). It proved so useful to others after its publication on the web that the International Organization for Standardization invited SIL to submit the scheme to its standards process. This was done after incorporating codes for 600 extinct and constructed languages that were developed by Linguist List and then reconciling the differences between that combined set of codes and the existing ISO 639-2 standard that had codes for almost 400 languages. The process came to a climax in February 2007 with the formal publication of *ISO 639-3, Alpha-3 Code for Comprehensive Coverage of Languages* as an international standard (ISO 2007). It includes 7,500 codes for all known human languages, past and present, and can serve the language resources community as a foundational scheme for interoperation. The establishment this year of ISO 639-3 has the potential to do for languages what the establishment of the Prime Meridian and time zones did for time123 years ago. [11]

With this background, I am ready to begin presenting a twelve-point vision for interoperation in linguistics. Part 1 covers "Global search." The first eight points deal with interoperating over passive content; they address the question, "How, in a world of unlimited choice, can we ever find what we're looking for?" Part 2 deals with "Global collaboration." The last four points deal with interoperating among agents who are active in creating and using language resources. They address the question, "How, once we realize that the treasury of linguistic knowledge is not very full, can we match supply with demand to fill it?" [12]

3. Toward global search: Interoperation over passive content

The first part of the vision for doing linguistics in the 21st century pertains to the interoperation of information. In order to knit the language resources of the world into a single virtual storehouse for global search and comparison, it must be made interoperable. [13]

The size of today's public Internet is estimated to be 100 million distinct web sites displaying 30 billion unique web pages (Pandia 2007). When faced with such abundance, a typical reaction is, "The language resource I'm looking for is probably out there somewhere, but which site should I look on and how do I find the exact resource once I get to the right site?" It can be like searching for a needle in a haystack. [14]

The general strategy that has been developed for solving this problem on the web is to *aggregate* and *filter*. In the value chain of Internet publishing, a producer publishes content by placing it on a web site. An aggregator then discovers that content and inserts it into a single index of all known web content. This solves the "Which site do I look on?" problem by creating a single place to look. A filter addresses the "How do I find it?" problem by showing only the resources that match the user's search criteria. The results of such a query are displayed in a browser through which the consumer is then able to access the published content. This, of course, is the model of Google and all the other web search services. The search engine with which end users interact is the filter; but, behind the scenes, the key to global search is the aggregator that amasses all known web resources into a single collection. [15]

We can distinguish two kinds of interoperation in web search: shallow and deep (Simons and Dry 2006:27*ff.*). Shallow interoperation is generic to all problem domains. It aggregates everything that is reachable via the ubiquitous HTTP infrastructure of the web and filters on the surface content of plain text. By contrast, deep interoperation is built for a specific problem domain. It uses a domain-specific protocol to aggregate only what is relevant to the domain community. It also uses domain-specific markup and vocabularies to filter on the underlying concepts and structures of the domain. [16]

Both approaches involve good news and bad news. Regarding shallow interoperation, the good news is that it already exists on a global scale (through services like Google and Yahoo!) and it is easy to support and use. But the bad news is that it gives poor results for language resources. This is because query results have lots of drop out when the words used in queries have synonyms and translations and those were the terms actually used in the pages one is looking for; this is the missed train problem (which information retrieval specialists call *low recall*). Queries also return lots of noise when the words used in queries have many other senses of meaning so that many irrelevant pages are returned; this is the train wreck problem (which information retrieval specialists call *low precision*).

Regarding deep interoperation, the good news is that it gives both high recall and high precision. But the bad news is that it is more work to achieve because information providers must follow domain-specific standards. But this is what we must do if we are going to achieve the kind of interoperation we are looking for in our domain community. [17]

The first element in the twelve-point vision for a cyberinfrastructure for the language resources community is that it must be anchored by an **aggregator**:

(1) The community needs an aggregator for language resources to anchor its cyberinfrastructure.

Such an aggregator would provide a single authoritative inventory of every resource in the treasury of linguistic knowledge. If we follow the riches analogy, it is like amassing the community's assets into a single fund. This is desirable because the larger the fund grows, the greater the possible return on investment for the community. Not only is there the search engine that allows community members to find resources they want to borrow from the fund (without actually needing to pay them back), but also there is an open web services API (or "application programming interface") that allows members of the community to invest information resources and earn interest by building services that add value for the community. Most of the points in the envisioned cyberinfrastructure are instances of such services. [18]

The second element of the cyberinfrastructure is a **metadata standard**:

(2) The community needs a metadata standard that describes resources in a way that will support its filtering needs.

Metadata is "data about data." It is like the library catalog cards that describe the resources in a library. The digital library community has developed a generic metadata standard for the purpose of describing web resources, namely, the Dublin Core (DC) metadata element set (DCMI 2003). The language resources community can simply augment the generic standard, and indeed, this is what the Open Language Archives Community (OLAC, www.language-archives.org) has done (Bird and Simons 2004). [19]

The basic DC metadata standard has fifteen metadata elements: Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, and Type. OLAC adds five extensions specific to our community. The first, for Language Identification, uses ISO 639-3 as a controlled vocabulary to achieve precise identification. The remaining extensions use controlled vocabularies that have been developed by the community: Linguistic Data Type, Linguistic Field, Participant Role, and Discourse Type (Bird and Simons 2003). [20]

The third element of the cyberinfrastructure is a **submission protocol**:

(3) Institutions with language resources to share need an open protocol for submitting metadata to the aggregator.

An open protocol is one that is freely published so that any interested party can learn all the details and can be freely subscribed to so that anyone can implement the protocol if they so choose. The digital library community, through the Open Archives Initiative (OAI), has developed such a protocol for catalogs that describe resources using Dublin Core metadata (Lagoze and others 2002). Our community can simply adapt it, and indeed, this is what OLAC has already done (Simons and Bird 2003a). [21]

The OLAC protocol for submitting metadata specifies an exact syntax for describing individual resources (Simons and Bird 2003b) and the protocol for publishing and registering a complete repository catalog (Simons and Bird 2003c). Slide 22 shows a sample metadata record in the OLAC format; it is the description of a Shoebox-format lexicon of the Ega language (Côte d'Ivoire) housed in the University of Bielefeld Language Archive. Note the use of domain-specific code values on the Subject, Type, and Language elements. These are the key to high recall and precision in search. [22]

Slide 23 gives a complete list of the institutions that are currently sharing their language resource catalogs through the OLAC protocol. The participants include 34 institutions from seven nations. [23]

The fourth element of the cyberinfrastructure is a **harvesting protocol**:

(4) Institutions who want to provide filtering services need an open protocol for harvesting metadata from the aggregator.

The OAI (Lagoze and others 2002) and OLAC (Simons and Bird 2003c) protocols referenced above also include a harvesting protocol which allows institutions to retrieve metadata records from the aggregator and download them into their own database. It is an open protocol because any institution that wants to build a value-added search service over the metadata records aggregated from the participating archives is free to harvest the records and do so. [24]

OLAC currently aggregates around 30,000 records from the 34 participants. Two institutions implement search services over the complete set: Linguist List (http://linguistlist.org/olac/) and the Linguistic Data Consortium (http://www.ldc.upenn.edu/olac/search.php). Slide 25 shows a sample from the Linguist List service; it is the user-friendly view of the Ega lexicon record shown in slide 22. [25] The next slide gives a sample from the LDC service; it is the first screen from a search for Potawatomi showing nine matches from four archives. [26]

Slide 27 gives a diagram of the cyberinfrastructure that results from combining the four elements discussed thus far. It shows the aggregator and its metadata standard in the

center (labeled "1,2"), with a number of institutions (labeled "3") submitting their metadata on the left, and two institutions (labeled "4") harvesting metadata to provide search services. This is, in fact, a representation of the existing OLAC infrastructure in which there are currently 34 metadata providers and two search providers. The architecture is open and extensible so that any institution can join as a metadata provider or a search provider simply by implementing the appropriate protocol. Indeed, OLAC invites participation from any institution that has language resources or services to share. [27]

Now the vision for cyberinfrastructure goes beyond what we have already. The fifth element is **gateways**:

(5) The cyberinfrastructure needs to include gateways to language resources catalogued by other communities.

General academic catalogs contain language resources. Aggregators for those exist, such as WorldCat for research libraries (http://www.oclc.org/worldcat/) and OAIster for institutional repositories (http://oaister.umdl.umich.edu/). Our infrastructure needs gateways to such aggregators that will discover the language resources in those collections and "crosswalk" their descriptions to our metadata standard. The cyberinfrastructure diagram represents this as an external aggregator on the left that feeds a gateway (symbolized as a door labeled "5") that in turn submits metadata for the discovered language resources into our community's aggregator on the right. All of the infrastructure elements shown in slide 27 are gold, while the door representing gateways is gray. The color distinction in the diagrams is used to distinguish existing cyberinfrastructure elements (in gold) from projected elements (in gray). [28]

The sixth element of the cyberinfrastructure is **spiders**:

(6) The cyberinfrastructure needs to include spiders that discover language resources on the Web.

Uncatalogued language resources abound on the web, such as in academic papers and web sites, language community blogs and web sites, and even pages in minority

languages (which are a kind of primary language data). Our infrastructure needs spiders that will crawl the web to find such resources and then report them to our aggregator. The cyberinfrastructure diagram represents the Internet as a cloud on the left. Information about discovered resources flows through the spiders (labeled "6") that in turn submit metadata for the discovered language resources into our community's aggregator on the right. In this case, the top spider is shown in gold (since the first one exists, as described in the next paragraph) with more spiders behind it in gray (since the infrastructure needs more of them). [29]

The first linguistic spider exists, namely, ODIN — the Online Database of Interlinear Text (Lewis 2006; see also http://www.csufresno.edu/odin/). The methodology is to seed Google searches with abbreviations for glosses commonly used in text annotation (like SG, PL, NOM, FEM). An algorithm then scans each returned page for instances of three consecutive lines that match the pattern of text-gloss-translation typical in text glossing. If a match is found, it is assumed to be an interlinear text example and a language name is found in the preceding context. ODIN currently reports more than 41,000 instances of interlinear glossed text examples from over 700 different languages in more than 2,900 different linguistic documents. [30]

Slide 31 shows a screen shot of what a user of the service sees. In this case, the user has selected Aceh as the language of interest and ODIN reports that it has found three examples in the two documents to which links are given. [31]

But this is not the only interface that ODIN implements. If it were, ODIN would be just one more information silo in linguistic cyberspace. ODIN also implements the OLAC metadata submission protocol so that the aggregator knows that this page about Aceh exists. Thus anyone who searches for Aceh in the search service at Linguist List or the Linguistic Data Consortium will be directed to this page on the ODIN site. Slide 32 shows the metadata record generated by ODIN; note that the URL for the page shown in slide 31 is given in the metadata as the value of the `<dc:identifier>` element. [32]

The seventh element of the cyberinfrastructure is **subcommunities**:

(7) Subcommunities need to establish specialized conventions within the community's metadata standard in order to support specialized filtering services, gateways, and spiders.

The subcommunities within our larger community will always have more specialized filtering needs than the community-wide OLAC standard supports. Just as OLAC was immediately successful because it specialized the existing OAI and DC standards, our subcommunities can do the same by specializing the OLAC standard. Then they can build services that exploit their specializations of the metadata. [33]

One example of a subcommunity is those doing geocoding of language resources. The LL-MAP project at Linguist List (http://linguistlist.org/llmap/) is using extensions to the DC Coverage element for specifying geospatial coordinates. For instance, the following element in a metadata record pinpoints the location of Perth, Western Australia:

```
<dc:coverage xsi:type="dcmi:point">name= Perth, W.A.;
east=115.85717; north=-31.95301</dc:coverage>
```

LL-MAP can specify the use of markup like this within its subcommunity and then build a service that harvests all records with this type of content and plots them on a dynamically-generated map. Building on this example, the cyberinfrastructure diagram represents specialized search services for subcommunities as a terminal (like the generic search of element 4) that shows the graphic of a globe and is labeled "7".

Another example of a subcommunity is the July 14 workshop at the LSA summer institute organized by Barbara Lust and Suzanne Flynn on *Applying Cyberinfrastructure to the Language Sciences: A Case Study for Language Acquisition* (http://linginst07.stanford.edu/workshops/cyberinfrastructure/workshop.html). The workshop explored how the language acquisition subcommunity could build a cyberinfrastructure by developing subcommunity-specific refinements of the OLAC scheme.

Establishing a subcommunity can be as simple as defining a fixed metadata value for a particular metadata element. For instance, one of the working groups in this workshop explored the suitability of using version 1.0 of LIFT, an XML-based lexical interchange format (Hosken 2007), as a standard within the language resources community. All it would take to create a subcommunity focused on interoperation among resources that follow that format would be for members of that subcommunity to agree that they will always mark such resources by means of a DC Format element with the following fixed value:

```
<dc:format>LIFT 1.0</dc:format>
```

That subcommunity could then build a service that harvests all records containing such a metadata element. [34]

The eighth element of the cyberinfrastructure is **datum-level search**:

(8) Subcommunities need to establish content standards in order to support datum-level search over relevant resources harvested from the aggregator.

Searching for resources as a whole is not the only thing we want to do. We also want to search across resources for matching data within their content. Deep search like this requires that a subcommunity interested in a particular content type must establish standards for XML markup or for controlled vocabularies used within data content or for both.

GOLD, the General Ontology for Linguistic Description (Farrar and Langendoen 2003; see also http://www.linguistics-ontology.org/), is an example of a specialized standard that was developed by the E-MELD project. One outcome of the initial E-MELD workshop in 2001 was a consensus that XML markup represented best practice for language resources, but that it would not be feasible to prescribe any one schema for markup as best practice. The solution arrived at was thus to develop an ontology that would standardize at the level of the concepts behind the markup rather than on the markup itself; by transforming the original resources into their interpretation in terms of the concepts in GOLD, it would be possible to query across the resources in an

interoperable way. Proof-of-concept implementations were made of a service that queried across lexicons from three different languages that were encoded with three different markup schemas (Simons and others 2004b) and another service that queried across interlinear glossed texts from seven languages that were originally encoded with different markup schemas (Simons and others 2004a). [35]

Implementing datum-level search amounts to implementing the aggregate and filter strategy at a more granular level. The approach begins with the aggregator for the specialized content (which is represented in the cyberinfrastructure diagram as a pair of gears labeled “8”). The specialized aggregator begins by querying the community aggregator for all of the resources that are of interest to the specialized subcommunity. It then harvests those resources themselves and parses their content into a content-specific database. The subcommunity service then implements a filter (that is, a specialized search service) over that database. The subcommunity service also submits metadata descriptions of the reports it generates back to the aggregator for the community at large as we saw with ODIN in slide 32. In this way the larger community will be directed to the service provided by the specialized subcommunity when it would be relevant to a query entered on a community-wide search service. [36]

Slide 37 gives a diagram of the cyberinfrastructure that results from combining the eight elements discussed thus far. Recall that gold color represents elements that exist already while gray represents elements that have yet to be developed. Note, too, that while an element may be shown in the diagram as only a single icon, in fact the icon represents a type of which there would be multiple instances in a mature cyberinfrastructure. The focus of the first eight elements has been global search that interoperates over data. [37] The focus now turns to supporting collaboration among all the people involved in the development and use of language resources. [38]

4. Toward global collaboration: Interoperation among active agents

Today our big challenge is to pull together a cyberinfrastructure that amasses all known language resources. Once we get it in place, we will realize how empty the global

treasury of linguistic knowledge still is. We will find that, "The information I'm looking for isn't there!" Our next big challenge will be, "How do we fill the huge gaps in the global treasury of linguistic knowledge?" We will quickly realize that filling that treasury is a goal that the community of professional linguists, working individually as they have in the past, will never be able to achieve. The second part of the vision for doing linguistics in the 21st century thus deals with developing new habits of global collaboration in order to support the interoperation of services rendered by linguists, both professional and amateur alike. [39]

The book that first got me thinking about global collaboration was *The World is Flat*, by Thomas Friedman (2005). In setting up the premise of the book, he observes that in 1492, when Columbus acted on his conviction that the world was round by sailing west to reach "the countries of India," he thought he had reached part of the Indies, but in fact he had run into America. Friedman (2005:3–5) recounts how in 2004 he flew east to make his own voyage of discovery to Bangalore, the "Silicon Valley" of India. When he actually got to India, he was surprised to find parts of America—billboards touting American companies, software firms using American business techniques, people using American names and American accents at large call centers. It dawned on him that the world isn't round anymore; it's been flattened. [40]

Friedman (2005:9–10) observes that there have been three stages of globalization. In Globalization 1.0, which began with the voyage of Columbus, a handful of countries drove global integration as they sailed the seas to establish colonial empires. Globalization 2.0 began in the 19th century and was powered by the Industrial Revolution with its falling transportation costs. During that era, organizations and companies were able to globalize. The globalizing organizations amassed the global riches of knowledge in world-class libraries. Globalization 3.0 was ushered in at the turn of the 21st century when the Information Age, through the Internet, created a flattened world in which individuals can globalize. Today the global riches of knowledge are in every web browser. A peasant logging in from an Internet cafe in the South has access to the same resources as a professor logging in from a prestigious university in the West. [41]

Friedman talks about three things that converged to bring about the flat world. The first is the playing field. Innovations in the last 15 years have produced a global, web-enabled playing field for collaboration. The second is processes. The flattening did not happen as soon as the Internet became available; rather, it took a decade or so for business processes to change in order to achieve the great productivity breakthroughs made possible by the new technologies. The third is people. By the time the new processes were in place, three billion people (of China, India, Russia, Eastern Europe, Latin America, and Central Asia) who had been frozen out of the playing field only two decades ago, found themselves with the potential to plug into the field and play with the rest of the world. [42] Friedman calls this the "triple convergence" and concludes:

> It is this triple convergence—of new players, on a new playing field, developing new habits and processes for horizontal collaboration—that I believe is the most important force shaping global economics and politics in the early 21st century. … The scale of the global community that is soon going to be able to participate in all sorts of discovery and innovation is something the world has simply never seen before. (Friedman 2005:181–182) [43]

One example of changing processes is a paradigm shift that is in progress in the field of knowledge management (Kuhlen 2006). The traditional approach to knowledge management is that experts collect existing tacit knowledge and transform it into explicit knowledge that they represent formally and organize into a knowledge base. The emerging approach recognizes that knowledge is a common good and that it should be collaboratively produced. It is decreasingly produced individually, but increasingly by distributed and often virtually organized groups. The paradigm shift is from expert-focused knowledge warehouse management to community-focused collaborative knowledge production. The four elements of the cyberinfrastructure that remain to be presented are in support of this latter approach. [44]

The ninth element of the cyberinfrastructure is **open repositories**:

(9) Individuals need an open method for submitting language resources into repositories that are plugged into the aggregator.

We need to remove the archiving bottleneck by setting up self-service digital archives for language resources. Institutional repositories, in which faculty members self-archive their scholarly work (Johnson 2002), have become mainstream. For instance, DSpace (Smith and others 2003; see also http://www.dspace.org/) is now in use at eight of the world's top 30 universities (as ranked by the *Times Higher Education Supplement*): University of Cambridge, Massachusetts Institute of Technology, Cornell University, Australian National University, National University of Singapore, University of Melbourne, University of Toronto, and University of Michigan. [45]

The self-archiving workflow proceeds as follows. First, a contributor uses a web form to fill in a metadata template and to upload the language resource to the repository. Then a curator verifies the quality of the metadata and ensures that the submission falls within the collection policy of the repository. If all is okay, the curator accepts the submission into the permanent collection of the repository and the metadata for the resource is immediately available to the aggregator for discovery by the community. The cyberinfrastructure diagram represents these open repositories (labeled "9") with the same icon that is used for the metadata repositories for element 3. The difference is that these open repositories show an input coming from an icon that represents the collaborators who are contributing directly.

In order to foster global collaboration, we need more than just university faculty to be able to deposit language resources; anyone who is capable of creating a language resource should be able to contribute one. Thus our cyberinfrastructure needs for some institutions to host digital repositories that are open to deposit from members of the public. The DSpace system already supports the submission workflow described above; it also supports the OAI protocol for sharing metadata. What our community needs to do is to add the specializations for supporting OLAC metadata in order to be able to plug DSpace into the cyberinfrastructure for linguistics. [46]

The tenth element of the cyberinfrastructure is **collaboration with amateurs**:

(10) The cyberinfrastructure needs to include services that support generation and refinement of language resources through collaboration with amateurs.

This is a growing trend in science, sometimes referred to as "citizen science" (http://en.wikipedia.org/wiki/Citizen_science). In a recent essay in *Nature* about the growing use of such approaches in science, consulting editor Philip Ball observes that achieving the common good lies at the heart of this movement:

> Can you get thousands of people to work for you, generating a high quality product, without paying them? Conventional economics would answer: don't be silly.
>
> The trick is simple enough. The work must not, in fact, be for 'you' but for 'the public good'; there should be no top dog who rakes in profits, financial or otherwise. The corollary is that the fruits of all this labor must be freely available. If, furthermore, the goal is a worthy one, then people will flock to offer their time and effort for free. (Ball 2004)

A well-known example that Ball cites is Project Gutenberg (http://www.gutenberg.org/) in which thousands of volunteers over a span of twenty years have produced a freely downloadable library of over 20,000 books that are out of copyright. Another example is Wikipedia (http://www.wikipedia.org/) in which thousands of volunteers in just a few years have produced an encyclopedia with ten times the coverage of the *Encyclopedia Britannica*. Another is the NASA Mars Clickworkers project (http://clickworkers.arc.nasa.gov/) in which thousands of volunteers in just three months located and classified 200,000 craters on the Martian surface by clicking on photos presented on their web browsers. [47]

Astronomy is a field that has really embraced citizen science. The universe is just too big for professional astronomers to be looking everywhere all the time. Thus they have increased their coverage an order of magnitude by embracing collaboration with amateurs. *The Long Tail* (Anderson 2006), in a chapter with the intriguing title of "The new producers: Never underestimate the power of a million amateurs with keys to the factory," documents this trend toward professional-amateur collaboration in astronomy. The author quotes a 2004 industry report as concluding:

> Astronomy is fast becoming a science driven by a vast Pro-Am movement working alongside a much smaller body of professional astronomers and astrophysicists. (Anderson 2006:60)

Anderson goes on to observe:

> Over the past two decades, astronomy has become one of the most democratic fields in science, in part because it's so clear what an important role the amateurs play. (Anderson 2006:61)

Surely the documentation and preservation of endangered languages would qualify as a worthy goal in the eyes of the public. It seems that the great challenge for 21st century linguistics will be to launch the next great Pro-Am movement in science. It turns out that everyone on earth has personal, linguistic knowledge that is of interest to professionals in our community. Can we exploit the global collaborative playing field to enlist everyone in the quest for that knowledge? [48]

Some possible applications of Pro-Am collaboration in linguistics might be: (1) Performing the language documentation workflow, which begins when someone creates a recording, that someone else could in turn transcribe or translate, and still others could annotate, evaluate, or refine. If the primary recordings and all the derivative works could be deposited and withdrawn from open repositories, then people anywhere could contribute to the process. (2) Linguists and native speakers could collaborate in authoring a dictionary using a system like Wiktionary (http://www.wiktionary.org/). (3) Linguists could develop and post elicitation schedules that anyone would be invited to fill in for the language they know best. (4) Members of the public could be invited to verify (and correct as needed) the output of automated services that perform tasks like tagging parts of speech, parsing words or sentences, identifying the language of a text sample, and finding language resources on the web.

In each case we would have collaborators using a web-based service to create or refine resources. The cyberinfrastructure diagram represents a collaboration service as a set of pages (labeled "10") that receive input from the collaborators icon. Any such

collaboration service should also submit metadata for the created resources to the community-wide aggregator so that they will be found in searches. [49]

The eleventh element of the cyberinfrastructure is **matching supply with demand**:

(11) The community needs an infrastructure for matching the demand for work to be done with the supply of people who could do it.

Harnessing the collaborative workforce is a matter of matching supply and demand. The interactive web has excelled at matching consumers to niche products, for instance, Amazon.com for books, Rhapsody.com for music, and Netflix.com for videos. The interactive web has also excelled at matching individual buyers and sellers, for instance, with eBay.com for online auction and eHarmony.com for online dating. [50]

Can we put some of those approaches to work for the language resources community? We have a huge demand to perform tasks on language resources like gather, convert, transcribe, translate, segment, annotate, analyze, describe, and evaluate. There is also a potentially huge supply of people who could get involved. We need an infrastructure for expressing this demand and fitting it to the supply of volunteers.

The cyberinfrastructure diagram represents a matching service as a group of interlocking puzzle pieces (labeled “11”) that receive input from the collaborators icon. Such a service should query the community aggregator in order to make inferences about next steps that are needed in the language resources workflow. It should also submit metadata describing the work needing to be done so that people making queries would learn of contributions they could make that are in line with their expressed interest. [51]

The twelfth and final element of the cyberinfrastructure is a means for **measuring authority and reputation**:

(12) The community needs protocols for measuring the authority of resources and the reputation of contributors within the cyberinfrastructure.

In the economy of scarcity, the authority of a work was conferred by the prestige of the journal or the publisher or the author’s institution. It was also conferred by the author’s

degree and tenure status. In the economy of abundance, scholarship is changing. New metrics for the authority of resources and for the reputation of contributors are emerging from the "collective intelligence" of the web (Jensen 2007, Van de Sompel and others 2004). [52]

For measuring the authority of resources, mechanisms involving user feedback are common on the web today: user product ratings, user comments, and the percent who found something helpful. There are also a number of automated metrics like Google page rank, the number of page views, the number of outgoing and incoming links, and the rank of referred-to and referring pages. On a wiki-based site, the number of times a page has been edited and the number of people who have edited it can be taken as reflections of how authoritative it has become.

For measuring the reputation of contributors, an example of a well-known mechanism involving user feedback is the rating of buyers and sellers on eBay. Automated metrics for reflecting the reputation of a contributor are the number of contributions by that contributor and an analysis of the authority metric assigned to those contributions. These metrics may become important for the cyberinfrastructure for they will help to provide incentive and gratification. In the economy of scarcity the latter derive from financial recognition, while in the economy of abundance they derive from the satisfaction of serving the common good and the reputational recognition of one's contributions (Kuhlen 2006:47).

The cyberinfrastructure diagram represents a service for measuring authority or reputation as a meter (labeled "12"). It has the aggregator as a primary input for measuring both. Another important input for gauging the authority of resources is incoming references from the Internet at large. Another input for gauging level of contribution would be the service that matches supply with demand. [53]

5. Conclusion

Slide 54 gives a picture of the complete cyberinfrastructure that emerges when we put together all twelve elements discussed above. By agreeing to work with shared standards,

we could act in community to build such an infrastructure for searching the global riches of linguistic knowledge. And by exploiting the new playing field for global collaboration, we could partner with the speakers of the world's languages to actually fill that treasury of knowledge. [54]

References

Anderson, Chris. 2006. *The long tail: Why the future of business is selling less of more.* New York: Hyperion.

Bell, Philip. 2004. "The common good." News@nature.com. Published online: 20 August 2004; doi:10.1038/news040816-14. http://www.nature.com/news/2004/040816/pf/040816-14_pf.html

Bird, Steven and Gary Simons. 2003. "Extending Dublin Core metadata to support the description and discovery of language resources." *Computers and the Humanities* 37(4):375–388. Preprint: http://arxiv.org/abs/cs.CL/0308022

Bird, Steven and Gary Simons. 2004. "Building an open language archives community on the DC foundation." In Diane I. Hillmann and Elaine L. Westbrooks (eds.), *Metadata in Practice*, pages 203–222. Chicago: American Library Association. Preprint: http://www.sil.org/~simonsg/preprint/Metadata in Practice.pdf

Blaise, Clark. 2000. *Time lord: Sir Sandford Fleming and the creation of standard time.* New York: Pantheon Books.

DCMI. 2003. "DCMI metadata terms." Dublin Core Metadata Initiative. http://dublincore.org/documents/2003/03/04/dcmi-terms/

Farrar, Scott and D. Terence Langendoen. 2003. "A linguistic ontology for the semantic web." *GLOT International* 7(3):97–100.

Friedman, Thomas L. 2005. *The world is flat: A brief history of the twenty-first century.* New York: Farrar, Straus, and Giroux.

Grimes, Joseph E. 1974. *Word lists and languages.* Technical Report No. 2, Department of Modern Languages and Linguistics, Cornell University, Ithaca, NY.

Hosken, Martin. 2007. "Lexicon Interchange Format (LIFT): A description." Version 1.0. http://lift-standard.googlecode.com/files/lift_10.pdf

ISO. 2007. *ISO 639-3:2007: Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages.* Geneva: International Organization for Standardization. http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=39534. Registration authority: http://www.sil.org/iso639-3/

Jensen, Michael. 2007. "The new metrics of scholarly authority." *The Chronicle of Higher Education* (Section: *The Chronicle Review),* volume 53, issue 41 (June 15, 2007), page B6. http://chronicle.com/free/v53/i41/41b00601.htm

Johnson, Richard K. 2002. "Institutional repositories: Partnering with faculty to enhance scholarly communication." *D-Lib Magazine*, volume 8, number 11. http://www.dlib.org/dlib/november02/johnson/11johnson.html

Kuhlen, Rainer. 2006. "Change of paradigm in knowledge management: Towards collaborative knowledge management." *Second EPOS Forum on International Health Consultancy: Workshop on Applied Knowledge Management in the Health Sector,* 18 May 2006, Maritim Bad Homburg, Germany. http://www.inf-wiss.uni-konstanz.de/People/RK/Vortraege06-Web/PP-epos-kollaboratives_knowledge-management-epos180506.pdf

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. *The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0.* http://www.openarchives.org/OAI/openarchivesprotocol.html

Lakoff, George. 1995. Metaphor, morality, and politics, or, Why conservatives have left liberals in the dust. In *Webster's World of Cultural Democracy*. Seattle: The World Wide Web center of The Institute for Cultural Democracy. http://www.wwcd.org/issues/Lakoff.html

Lewis, William D. 2006. "ODIN: A model for adapting and enriching legacy infrastructure." A paper presented at *e-Science 2006,* 4–6 December 2006, Amsterdam. Preprint: http://faculty.washington.edu/wlewis2/papers/ODIN-eH06.pdf

Pandia, 2007. "The size of the World Wide Web." *Pandia Search Engine News.* http://www.pandia.com/sew/383-web-size.html

Sen, Amartya. 1999. *Development as freedom.* New York: Anchor Books.

Simons, Gary F. 2002. "SIL three-letter codes for identifying languages: Migrating from in-house standard to community standard." *Proceedings of the Workshop on Resources in Tools and Field Linguistics*, Third International Conference on Language Resources and Evaluation (LREC), 26–27 May 2002, Las Palmas, Canary Islands, Spain. Pages 22:1–8. Preprint: http://www.sil.org/~simonsg/preprint/SIL language codes.pdf

Simons, Gary F. Forthcoming. "Linguistics as a community activity: The paradox of freedom through standards." In William D. Lewis, Simin Karimi, Heidi Harley, and Scott Farrar (eds.), *Time and Again: Theoretical and Experimental Perspectives on Formal Linguistics. Papers in honor of D. Terence Langendoen.* Amsterdam: John Benjamins. Preprint: http://www.sil.org/~simonsg/preprint/standards.pdf

Simons, Gary F. and Steven Bird. 2003a. "Building an open language archives community on the OAI foundation." *Library Hi Tech* 21(2):210–218, special issue on the Open Archives Initiative. Preprint: http://arxiv.org/abs/cs.CL/0302021

Simons, Gary F. and Steven Bird. 2003b. "OLAC metadata." Standard, Open Language Archives Community. http://www.language-archives.org/OLAC/metadata.html

Simons, Gary F. and Steven Bird. 2003c. "OLAC repositories." Standard, Open Language Archives Community. http://www.language-archives.org/OLAC/repositories.html

Simons, Gary F. and Helen Aristar Dry. 2006. "Preservation, intelligibility, and interoperability: The E-MELD vision of digital language resources." *Proceedings of the E-MELD Workshop on Tools and Standards: The State of the Art*, 20–22 June 2006, Lansing, MI. http://www.emeld.org/workshop/2006/papers/aristar-dry-simons.pdf

Simons, Gary F., Brian Fitzsimons, D. Terence Langendoen, William D. Lewis, Scott O. Farrar, Alexis Lanham, Ruby Basham, and Hector Gonzalez. 2004a. "A model for

interoperability: XML documents as an RDF database." *Proceedings of the E-MELD Workshop on Linguistic Databases and Best Practice*, 15–18 July 2004, Detroit, MI. http://emeld.org/workshop/2004/simons-paper.pdf

Simons, Gary F. , William D. Lewis, Scott O. Farrar, D. Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez. 2004b. "Mapping legacy markup schemas to a common semantics." *Proceedings of the XMLNLP Workshop,* Association for Computational Linguistics, Barcelona, Spain, July 2004. Preprint: http://emeld.org/documents/SOMFinal1col.pdf

Smith, MacKenzie and others. 2003. "DSpace: An open source dynamic digital repository." *D-Lib Magazine,* volume 9, number 1. http://www.dlib.org//dlib/january03/smith/01smith.html

Van de Sompel, Herbert and others. 2004. "Rethinking scholarly communication: Building the system that scholars deserve." *D-Lib Magazine,* volume 10, number 9. http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Wang, Spencer. 2006. "The long tail: why aggregation and context and not (necessarily) content are king in entertainment." Bear, Stearns & Co. http://www.bearstearns.com/bscportal/research/analysts/wang/112706/Slide1.htm