

Language Digitization

Steven Abney

“Language Digitization” is not a standard term. I use it to mean language documentation and description, where the results are intended for machine consumption instead of human consumption. Essentially, instead of producing printed documents for human reading, one produces a large database that is suitable for automated processing.

The motivation is not merely the development of language technology. For software development, it is clear enough why one would want data that supports automated processing. But the aims of language digitization are primarily scientific: they represent the way a computational linguist would approach the question of universal grammar.

It turns out that there are interesting convergent trends in language documentation and computational linguistics, though also significant differences. Language documentation is familiar to most linguists, but a review will be useful here. Then we will turn to computational linguistics.

Himmelman (1992) is usually cited as defining the current approach to language documentation. He proposed separating language documentation from language description, and both from theoretical linguistics. The goal of documentation is the collection of primary data, including audio and video recordings, field notes, and indigenous written materials. Language description is concerned with the organization and annotation of the primary data. Traditionally, its products are interlinear glossed texts (annotation), and a lexicon and descriptive grammar (organization). Both documentation and description contrast with theorization. In the former two, objectivity is paramount (“just the facts, ma’am”), whereas in theorization, the goal is to construct a model of the workings behind the scenes that *explain* the observed facts.

Himmelman argued that documentation—primary data collection—is an important activity in its own right, with its own measure of quality. In the words of Himmelman:

The aim of language documentation is to provide a comprehensive record of the linguistic practices of a speech community. [**needs citation**]

In this, there is agreement between language documentation and computational linguistics: data collection is most useful when it is comprehensive and systematic.

Systematic data collection is diametrically opposed to the traditional model, in which data is collected in response to particular questions posed by descrip-

tive or theoretical linguists. There are strong reasons for preferring systematic documentation. For one thing, descriptive or theoretical linguists are not the only parties with an interest in language: good language documentation should also serve the needs of sociolinguists, anthropologists, discourse analysts, those interested in oral history, and the speakers of the language themselves.¹ And even theoreticians themselves are better served by systematic documentation, if we take the long view. The questions posed by current theories are likely to seem quaint or even incomprehensible to future theoreticians. To quote Ives Goddard:

[T]he linguist who has a philological [i.e., documentary] approach looks not only to the past but also to the future; he must be concerned with minimizing the problems which the documents he produces will cause his successors. **[needs citation]**

Language description adds a layer of annotation and organization to the primary documentation. Annotation typically takes the form of **interlinear glossed text (IGT)**, such as the following:

you yoTaa lobhi kukur-ko kathaa ho.
 PROX.L one greedy dog-GEN story be.3smL **[needs
 citation]**
This is a story of a greedy dog.

In addition to glossed text, the other major products of language description are a lexicon and descriptive grammar. Ideally, these merely organize information that is already present in the annotated text. For example, every dictionary entry should ideally be able to cite locations in the text where the lexeme occurs, and if the annotation is sufficiently rich, the information in the lexical entry should be transparently reconstructable from the citations.

Description contrasts with theory-construction in that it attempts to be objective and factual rather than explanatory. Its terms are not theory-internal, but observational, and a major desideratum is consistency: different observers should use terms in the same way. The majority of annotation efforts in computational linguistics are examples of language description. Interannotator agreement is a key measure of a well-defined annotation task, and it is achieved through the use of controlled vocabularies and rules of annotation, often called “stylebooks.” The International Phonetic Alphabet (IPA) is an example of a controlled vocabulary—it is a set of symbols for phonetic annotation with well-defined meanings. Ideally, two phoneticians listening to the same recording should transcribe it identically. Equally importantly, a future phonetician should be able to read an IPA transcription and know exactly how to interpret it. Computational linguists have developed similar controlled vocabularies for parts of speech and even for sentence structure. There are also proposed controlled vocabularies and annotation rules for IGT, such as the Leipzig Rules **[needs citation]**.

¹This is Himmalmann’s list. Computational linguistics is notably absent.

Unlike description, theory is not factual. To quote Einstein, “imagination is more important than knowledge” **[needs citation]**. Theory construction takes the descriptive facts as a starting point, and seeks a model of what is going on behind the scenes to give rise to the facts. Success depends on imagination and insight, not stylebooks and consistency.

Theory is important, but good theory must build on a good foundation of documentation and description. As Himmelmann points out, the structure has become dangerously top-heavy. For example, one would expect to have more text collections than lexica, and more lexica than grammars, but in fact the ratio is more like 1 : 3 : 10, by Himmelmann’s estimate. To rectify the situation, there needs to be increased respect for documentation, and increased appreciation of it as an undertaking with its own goals and measures of quality. In fact, since Himmelmann’s original paper, documentation has enjoyed dramatically increased attention and prestige. A major motivation has been alarm about language endangerment.

We are all familiar with the estimate that 90% of the world’s languages will disappear by the end of the century. That estimate comes from Krauss, and is based on the reasoning that a “safe” language is one with at least 100,000 speakers,² and that any language that is not safe is endangered. Half of the world’s 6,000 languages³ have fewer than 6,000 speakers, and only about 600 languages are “safe” by Krauss’s definition. **[needs citation]**

The UNESCO *Language Atlas* provides finer distinctions **[needs citation]**. They classify a language as vulnerable if it is spoken only under restricted circumstances (e.g., at home but not at work or school), definitely endangered if children no longer learn it as a first language, severely endangered if the youngest fluent speakers are elderly, and critically endangered if the only remaining speakers are elderly and less than fluent. Their current tabulation is as follows. Note that the *Language Atlas* does not distinguish between languages that are safe and those about which no data is available.

Category	In category	Cumulative
Extinct since 1950	4%	4%
Critically endangered	10%	13%
Severely endangered	9%	22%
Definitely endangered	11%	33%
Vulnerable	10%	43%
Safe or <i>no data</i>	57%	100%

A third of the world’s languages are no longer learned by children, and will be extinct within a few decades. Assuming that most of the vulnerable languages, and some portion of the languages for which there is no data, will become extinct as well, a conservative estimate places language loss at 50% by the end of the century.

²Or government support, though virtually all languages with government support have at least 100,000 speakers.

³According to the *Ethnologue*, there are 6,909 living languages, but *Ethnologue* tends to “split” rather than “lump” in uncertain cases, often treating as separate languages two varieties that others would classify as dialects of a single language.

In addition to endangered languages, we must also be concerned about endangered documentation. An untold quantity of primary documentation moulders in cabinets of individual researchers, to be thrown out when the researcher retires. The materials are typically in the form of poorly preserved audio tapes, or, if digital, in data formats or on media such as floppy disks that are no longer readable. The collection and dissemination of primary data was held in low regard for many years; and the idea of making data publicly, electronically available has been absent entirely. But without replication, and preservation by redundancy, not even archived material is safe. In May 2010, for example, an archive of unique recordings in Papua New Guinea was destroyed when, through a bureaucratic error, the building housing the archive was razed before the archive was moved [needs citation].

Attitudes toward data distribution differ significantly between documentary and computational linguistics. In computational linguistics it is standard practice to make data freely available.⁴ Research projects typically release their data publicly, and collaborative data collection and annotation projects are commonplace.

The lack of a similar public expectation of data-sharing in linguistics leads to a third kind of endangerment, what we might call **digital endangerment**. For computational linguistics, data that is not publicly available might as well not exist, and languages without publicly available data receive no attention. This is to the detriment not only of technology development in those languages, but also to the detriment of their effective preservation.

The number of **digitized languages**, in the sense of languages with significant publicly-available electronic resources, is painfully small: Maxwell and Hughes place it at about 30 [needs citation]. There are perhaps that many again where private resources exist. For example, Google currently translates 57 languages [needs citation]. But not only does that represent only 1% of the world's languages, all but 18 of them belong to the same language family (Indo-European), and *none* of them are endangered.

The situation is unlikely to change without a synthesis of computational linguistics and documentary linguistics. To put it bluntly, documentary linguistics is motivated to cover the world's languages comprehensively, but it does not understand language digitization. Computational linguistics does understand how to digitize languages, but it is interested only in languages that have commercial or intelligence value—essentially, the languages that it has already digitized. To make further progress, we must combine the motives of documentary linguistics with the know-how of computational linguistics.

Doing so is valuable not only for the preservation of primary linguistic data for future generations. The computational linguistic interest in digitization is predicated not merely on engineering, but on an approach to scientific inquiry that is characteristic of current trends in the natural sciences broadly. Integrating computational linguistics into linguistics has profound potential not only for

⁴Prominent data providers, such as the Linguistic Data Consortium (LDC), do charge a subscription fee, though some portion of the community considers even subscription fees to violate the prevailing spirit of data sharing.

documentary and descriptive linguistics, but also for theoretical linguistics.

Computational linguistics has grown up as a field separate from linguistics, but that was not the original intent. Computational linguistics was born from machine translation in the late 1960s. Computers were used for machine translation from the very first. The development of general-purpose computers was intertwined with codebreaking efforts in World War II, and initial efforts in machine translation used technologies developed for cryptography. Warren Weaver famously wrote,

When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode” [needs citation].

In 1966, an important government report, the ALPAC report, was published on progress in machine translation. It determined that machine-aided translation was both more expensive and qualitatively worse than using governmental translation services, and, as a consequence, funding for machine translation was deeply curtailed for many years. What is much less generally known are the actual recommendations of the report. These occupied a single page of the report. Expenditures were recommended, first and foremost, in

computational linguistics as a part of linguistics—studies of parsing, sentence generation, structure, semantics, statistics, and quantitative linguistic matters . . . Linguistics should be supported as science, and should not be judged by any immediate or foreseeable contribution to practical translation [needs citation].

The envisioned integration of “computational linguistics as a part of linguistics” never occurred. Linguistics moved away from parsing, sentence generation, and quantitative matters, and computational linguistics became disillusioned with symbolic grammars.

To become disillusioned, one must first put one’s heart into making something work. During the 1970s and 1980s, the central goal of computational linguistics was the development of rigorous generative grammars—rigorous in the sense that they support parsing and generation of natural text and speech. One could say that the focus was on two of Chomsky’s well-known three questions, namely, “What is language?” and “How is it processed?”

The efforts stalled in disappointment and frustration. There were four major issues: (1) how to resolve ambiguities in a principled way; (2) how to deal with noise, both because of the messiness of natural input and because, in Sapir’s words, “all grammars leak;” (3) how to handle the sheer magnitude of the problem; and (4) how to disentangle parts of the problem. On the latter point, it was common wisdom that to solve even simple problems like assigning parts of speech, one must solve the entire problem of AI. In evidence, examples were adduced like the following, that show that one can only determine the part of speech of “duck” by understanding the complete context:

when he began flailing about, he made her **duck**
when he invited her to dinner, he made her **duck**

The resolution of the frustration came by shifting attention to the third of Chomsky’s questions: How is language learned? The breakthrough came with the demonstration that part of speech assignment could be disentangled and solved as a separate task, with high accuracy on unrestricted natural text, by using probabilistic models. This addressed not only the entanglement and noise issues, but also supplied a principled basis for ambiguity resolution (probability theory), and addressed the issues of scale through the way in which such systems were constructed, namely, by training from large annotated resources instead of crafting grammars by hand.

In the new paradigm, nearly every linguistic problem is treated as a learning problem, and specifically, a problem of inducing a system from annotated data. This is the source of the premium that computational linguistics places on electronic resources—which is to say, language digitization.

The trend in computational linguistics is part of a larger trend in the natural sciences. Jim Grey, a fellow at Microsoft Research, identified four historical paradigms of scientific research [**needs citation**]. The “Empirical Paradigm,” which goes back at least to the Renaissance, is characterized by observation rather than appeal to authority. The “Theoretical Paradigm” adds to observation the formulation of simple, mathematical laws of nature, such as Maxwell’s equations. The “Computational Paradigm” allows scientific method to be extended to phenomena that are too complex to be governed by simple laws; it uses computer simulations of the workings of the system to account for observed properties of complex systems. The use of Monte Carlo methods in the development of the atomic bomb can be cited as an early instance, and it is currently used to understand everything from interactions between galaxies to the workings of the cell.

The fourth paradigm, the “Data Exploration Paradigm,” is nicely summarized by this quote from Gordon Bell:

In the 20th century, the data on which scientific theories were based was often buried in individual scientific notebooks . . . Such data, especially from individuals or small labs, is largely inaccessible. It is likely to be thrown out when a scientist retires, or at best it will be held in an institutional library until it is discarded.

In the 21st century, much of the vast volume of scientific data captured by new instruments on a 24/7 basis, along with information generated in the artificial worlds of computer models, is likely to reside forever in a live, substantially publicly accessible, curated state for the purposes of continued analysis [**needs citation**].

The similarities to our earlier comments concerning endangered documentation should be obvious.

Let us consider briefly what Bell means by the volume of data captured by new instruments. An excellent example is provided by the new Large Hadron Collider at CERN. According to the CERN web site, it “will produce 15 petabytes

(15 million gigabytes) of data annually” [needs citation].⁵ But what is even more interesting is what happens with that data. In order to make it available to the entire physics community, CERN has constructed a new international infrastructure for distributed computing and storage, called the Worldwide LHC Computing Grid (WLCG). Clearly, they are deeply committed to publicly accessible data.

Another example of relevance to linguistics is the Human Genome Project.

One of the greatest impacts of having the sequence may well be in enabling an entirely new approach to biological research. In the past, researchers studied one or a few genes at a time. With whole-genome sequences . . . and new high-throughput techniques, they can approach questions systematically and on a grand scale. They can study . . . how tens of thousands of genes and proteins work together in interconnected networks to orchestrate the chemistry of life [needs citation].

Large-scale, publicly accessible language digitization is not just about preservation: it is about transforming the practice of linguistics, and enabling an entirely new kind of research.

Here is another way to put it. Let us consider how a computational linguist would approach the question of universal linguistics. His or her first instinct would be to build an annotated dataset. In this case, the question of interest is learning entire languages, so the dataset would consist of digitizations of a large number of languages. Such a dataset is new both for linguistics and computational linguistics. Generalizing across languages is nothing new for linguistics, but systematicity and support for automated processing are. Large electronic resources are nothing new for computational linguistics, but computational linguistics focuses on “vertical” integration of resources at different levels of description for a single language. Very few existing datasets support “horizontal” processing, across many languages.

The proposal is to construct a publicly available, machine processable, annotated Universal Corpus. The purpose is to enable a new, systematic approach to Universal Grammar. Almost eighty years ago, Bloomfield wrote:


The only useful generalizations about language are inductive generalizations. . . . The fact that some features are, at any rate, widespread, is worthy of notice and calls for an explanation; when we have adequate data about many languages, we shall have to return to the problem of general grammar and to explain these similarities and divergences, but this study, when it comes, will be not speculative but inductive [needs citation].

It is, at long last, time to begin the endeavor.

⁵By comparison, even if we collect 10 million words of CD-quality speech for every language of the world (that is far more than we can actually hope for), the entire corpus will occupy only 2.1 petabytes.

What should go into such a corpus? How do we digitize a language? One way of thinking about it is to ask what it would take to resurrect a dead language—what record of a language is necessary for a human to learn to speak it? More ambitiously, what record would be necessary to create an artificial speaker?

At a minimum, we need the sound-meaning mapping. An imperfect, but practical, representation of the meaning of a sentence is its translation into a reference language, like English. So, at a minimum, we require a large number of examples of recorded speech with a transcription and a translation. In machine translation, the combination of text and translation is called a **bitext**, so let us call this **bitext-annotated speech**.

Audio: 

Text: ii haṇṇina maali maravannu hatti, aa, maavina haṇṇugaḷanna udaristirtaane

Trans: *This fruit farmer, having climbed the tree, is picking mango fruits.*

[needs citation]

That may do for documentation, but what might we need, if our goal is to develop a universal grammar in a computational linguistic fashion? Computational linguists use annotated corpora for training systems, and for testing them. The methodology is a version of the scientific method: a training set is used to develop a model (hypothesis), and it is evaluated by comparing its predictions to what actually occurs in a test set. We do not evaluate the model’s “predictions” on the training set—predicting the past is too easy. The test set consists of a fresh experiment: new items drawn after the model is settled on.

In the case of a rigorous grammar, a reasonable evaluation is how well it does at assigning the correct structure to sentences—that is, at parsing. The usual way to evaluate a parser is by building a treebank (a collection of manually parsed sentences). Treebanks exist for a handful of languages; creating them is an enormous amount of work. Creating treebanks for 6,000 languages is hopeless. Now, arguably, the reason for wanting the sentence structure is to enable interpretation, so instead of using the parse tree for evaluation, we could use the sound-meaning mapping. Unfortunately, if we use a logical calculus to represent meaning, we have gone out of the frying pan into the fire. But we could instead use a translation into a reference language to represent meaning. By this reasoning, we arrive back at the idea that transcription and translation—bitext annotation—may be sufficient, even if our goal is universal grammar. Incidentally, machine translation systems are standardly trained from nothing but bitexts, so the idea of using bitexts as a basis for grammar development is at least plausible.

One catch is the amount of bitext that may be needed. Estimates for the amount needed to train a machine translation system vary widely, but range from about two to ten million words. That is a good deal more than we are likely to acquire for any endangered language. The Bible, for example, represents a bitext collection (the verses permitting sentence-by-sentence alignment with a translation). It consists of 0.8 million words, but is available for less than 500

languages [**needs citation**]. The New Testament is available for over 1200 languages, but it is only 0.1 million words in size.

In part, current machine translation methods need so much data because they use brute force in place of linguistic sophistication. We can hope to do better, and developing data-lean methods for machine translation is an interesting computational challenge.

But there is no doubt that we need to increase the rate at which language data is collected. The best hope for that is by enlisting help outside of linguistics. After all, a linguist obtains translations simply by asking a native speaker for them. For languages whose speakers have internet access, crowd-sourcing is a promising mechanism for getting more people involved. For example, the web site dotSub allows people to post videos, transcribe what is said, and translate the transcriptions, so that speakers of other languages can view the videos with subtitles [**needs citation**]. The purpose is video sharing, but a side effect is the production of bitext-annotated speech on a large scale.

A documentary linguistic project of particular note in this connection is BOLD:PNG [**needs citation**]. It is a project headed by Steven Bird to collect ten hours of annotated speech for each of 100 languages in Papua New Guinea, relying entirely on speakers of the languages. At several universities in Papua New Guinea, Bird provided students with digital recorders and training, and the students returned to their home villages to make recordings. Each student collected 10 hours of speech, selected one hour for oral “transcription” and translation, and one-tenth of an hour for written transcription and translation. In “oral transcription,” the student repeats what the speaker said, slowly and clearly, onto a second recorder. Slow, clear speech allows a phonetician who does not speak the language to create a good transcription, and may in the future make machine transcription viable.

Both of these projects are examples of **empowering speakers** to participate in the documentation of their own languages. Much of what is currently in language archives remains locked away because the researchers who collected the data never got permissions to distribute it. In the traditional model, collecting language data was treated much like collecting botanical specimens. To increase the volume of documentation, and to make the data publicly available, it is essential to move to a model in which the primary data is produced by the speech communities themselves, as a species of publication or broadcast.

A second major source of materials is legacy documentation. There is an enormous quantity of documentation and description that exists in print form from which we would like to extract information to construct digital resources. Here the challenges are computational. One task is crawling the web to identify materials in different languages. An idea that we are pursuing at the University of Michigan is to collect a seed of material from each language, and use the common words of that language to search for more documents in the same language; the new materials can then be used to improve the seed. A second task is to scan printed materials, pass them through optical character recognition, and process the results to identify bitexts, in the form of interlinear glossed text, paradigms, or lexical entries. Existing digital collections are an excellent source

of materials.

Once we have materials, what sort of annotation do we aim for? Bitexts suffice for many purposes, such as the training of machine translation systems, but we would like to have richer annotation for at least some material. Standard IGT adds morphological analysis and glossing to bitexts:

Text: ii haṇṇina maali maravannu hatti, aa, maavina haṇṇugaḷanna udaristirtaane
Morph: ii haṇṇ-ina maali marava-nnu hatti maavina haṇṇu-ḡaḷanna udarist-irtaane
Gloss: this fruit-GEN farmer tree-ACC climb.PF mango fruit-PL-ACC picking-state
Trans: *This fruit farmer, having climbed the tree, is picking mango fruits.*
[needs citation]

We can view the word-level alignment, in part, as adding a degree of supervision to training a translator. In standard MT training, one first aligns the transcript to the translation, and on the basis of the aligned texts, one learns common word translations, and common re-ordering patterns between the languages. The English word glosses provide anchors to the translation, hence information about both word translation and re-ordering. Once an MT system is trained, we can imagine using it to automatically supply the “Morph” and “Gloss” lines.

Though we earlier dismissed treebank construction as overly ambitious, there is in fact a lightweight way of representing syntactic information, namely, dependency structure. The Perseus Project has had impressive success at enabling classics students to create treebanks of Latin and Classical Greek [needs citation]. A dependency structure fits nicely into traditional IGT:

Morph:	ii	haṇṇ-ina	maali	marava-nnu	hatti	maavina	haṇṇu-ḡaḷanna	udarist-irtaane
Gloss:	this	fruit-GEN	farmer	tree-ACC	climb.PF	mango	fruit-PL-ACC	picking-state
POS:	DET	N	N	N	V	N	N	V
Role:	SPC	MOD	SBJ	OBJ	MOD	MOD	OBJ	ROOT
Govr:	3	3	8	5	8	7	8	–

For example, the annotations on “haṇṇ-ina” indicate that it is a noun (N) that modifies (MOD) the third word, i.e., “maali.”

In short, we aim to construct a database of digital interlinear glossed text. IGT is deceptive in its simplicity—particularly when supplemented with dependency information, it provides the raw material for grammar development and evaluation at all levels of linguistic description.

Digital IGT can also be viewed as integrated documentation and description. For example, having the word *maali* annotated with part of speech N and glossed “gardener” provides a fragment of a bilingual lexical entry:

maali (n.) *gardener*

With digital IGT, one can construct the lexicon automatically, and provide links directly from words in context to lexical entries, and from a lexical entry to a concordance of all the places where it occurs in the corpus (as an exhaustive set of examples of use), and from the concordance back to locations in the text.

Such interactivity is of benefit not only for language documentation and description, but also for language instruction, with a particular eye to language

preservation. The Perseus website provides a version of such an integrated text and lexicon for students of Latin and Ancient Greek.

To increase the usefulness of language documentation and description, and to automate many parts of the process, to provide data security through redundancy, to support language instruction and preservation, and to enable a new large-scale, systematic approach to research in universal grammar, it is time for linguistics to undertake the construction of a community resource on the scale of the Large Hadron Collider or the Human Genome Project. Doing so will require us to at last fulfill the original intention for computational linguistics: that it truly be a part of linguistics. Embracing computational linguistics will also change linguistics, bringing parsing, learning, and quantitative models back into the purview of the field. It is a change that, I think, will be for the good.

I can think of no better conclusion than to quote Steven Bird, from an essay addressed to the computational linguistic community:

We live during a brief period of overlap between the mass extinction of the world's languages and the advent of the digital age . . . It's time that we focused some of our efforts on a new kind of computational linguistics, one that accelerates the documentation and description of the world's endangered linguistic heritage, and delivers tangible and intangible value to future generations [**needs citation**].

References

- [1] Mike Maxwell and Baden Hughes. Frontiers in linguistic annotation for lower-density languages. *Proceedings of the COLING/ACL 2006 Workshop on Frontiers in Linguistically Annotated Corpora*. 2006.
- [2] Warren Weaver. Translation. Typescript, July 1949. Reprinted in: Locke, William N., and A. Donald Booth (eds) *Machine translation of languages: fourteen essays*. Technology Press of the Massachusetts Institute of Technology. Cambridge, MA. 1955.