The establishment of Motorola's Human Language Data Resource Center: Addressing the criticality of language resources in the industrial setting

Jim Talley

Motorola Labs, Human Interface Laboratory
7700 W. Parmer Ln., MD: PL26; Austin, TX 78729; USA
James_Talley@email.mot.com

Abstract

Within the human language technology (HLT) field it is widely understood that the availability (and effective utilization) of voluminous, high quality language resources is both a critical need and a critical bottleneck in the advancement and deployment of cutting edge HLT applications. Recently formed (inter-)national human language resource (HLR) consortia (*e.g.*, LDC, ELRA,…) have made great strides in addressing this challenge by distributing a rich array of pre-competitive HLRs. However, HLT application commercialization will continue to demand that HLRs specific to target products (and complementary to consortially available resources) be created. In recognition of the general criticality of HLRs, Motorola has recently formed the Human Language Data Resource Center (HLDRC) to streamline and leverage our HLR creation and utilization efforts. In this paper, we use the specific case of the Motorola HLDRC to help examine the goals and range of activities which fall into the purview of a company-internal HLR organization, look at ways in which such an organization differs from (and is similar to) HLR consortia, and explore some issues with respect to implementation of a wholly within-company HLR organization like the HLDRC.

1. Introduction / background

Today, it is fairly uncontroversial[1] to assert that one of the most crucial requirements for fielding highly functional applications of human language technology (HLT) is having access to substantial, appropriate, high quality human language resources (HLRs) for development and testing. As this realization has become increasingly clear over the years, the HLT community, at large, has responded by forming national and international consortia and data resource centers to attempt to meet those burgeoning resource requirements. That such organizations have yielded enormous benefits in the advancement of HLT is, I think, undeniable. Yet, by their very nature, those organizations will always fail to meet the complete set of requirements for HLRs that commercial enterprises have. That is, if data (corpora, dictionaries, annotations,…), *i.e.*, HLRs, provide value added above and beyond that of innovative and well-engineered algorithms (and, I think, few would argue against that position), then such data resources can be appropriately considered an essential part of a company's competitive advantage. And, it follows that a company which acquires all of its HLRs from publicly available sources is giving up some of its capability to differentiate itself and lead the market in its chosen area of HLT activity.

It may seem, by stating the above, that we are, in effect, sounding the death knell for public, HLR producing / distributing consortia. …or, at least, predicting industrial disinterest in their activities and resources. But, that is not the case. The consortial entities, like the LDC[2] or ELRA[3], which have done so much in the in the way of improving the quality of HLT products and prototypes and lowering barriers to entry in the field, will continue to play a vital role in the production, standardization, and dissemination of HLRs. And, their efforts will continue to be of great interest to companies working towards commercial deployment of HLT applications. Resources supplied by the public consortia will (continue to) serve as the pre-competitive basis upon which companies can build as they target their particular markets and strive to differentiate themselves from their competitors. That is, with respect to HLRs, there is a valid, on-going role for both pre-competitive (consortial) and competitive (company-specific) production. Moreover, a company can potentially differentiate itself by virtue of the efficacy with which it utilizes language resources (both internally produced and publicly available).

In recognition of the criticality of effectively managing all aspects of our HLR activities, Motorola has recently formed the Human Language Data Resource Center (HLDRC) as part of Motorola Labs, Motorola's corporate research organization. It is a fairly novel industrial organization (though similar in some ways to open, consortial HLR organizations like the LDC). The charter of the HLDRC can be summarized in a nutshell as "making the acquisition, creation, management, distribution, and utilization of HLRs as efficient and effective as possible for Motorola."

Most language data of interest (at least given today's technological capabilities) are specific, or targeted, to the task domain under consideration. HLRs are also relatively expensive to collect and annotate (though the costs pale beside those of other areas such as marketing). In the context of corporate issues such as organizational divisions, local budgetary control and agendas, and the like, those two factors (specificity and expense) tend to impede sharing and reuse of HLRs. Cross divisional groups tend to focus on the short-term objectives of their immediate organization, often indifferent to similar efforts by other groups within a company. This, in the context of HLR production or procurement, has, historically, meant: 1) collecting data which satisfied the group's immediate requirements with little attention to how minor

[1] One would assume that such an assertion would be especially uncontroversial for attendees at this conference on language resources and evaluation!
[2] Linguistic Data Consortium (LDC)
[3] European Language Resource Association (ELRA)

adjustments or augmentations might make the results more useful to a wider range of potential consumers (and/or more useful over a longer period of time), 2) making little, to no, effort to disseminate information about the existence or content of collected data sets, and 3) having (group-based) ownership issues effectively choke off any attempt at sharing / collaboration which happened to overcome the first two of these barriers. The net result, historically, has been underutilization of (probably greater than necessary) investments in the area of HLRs.

At Motorola, we realized that this typical, evolutionary state of affairs was at odds with our quest to deliver market-leading, advanced natural language based technologies in future products. That realization (and recognition of the, undoubtedly quite common, lack of coordination of our HLR efforts) led to the formation of the Human Language Data Resource Center (HLDRC).

2. Range of Activities

The HLDRC, as we conceive it, is primarily a service organization with the goal of facilitating the work of all of the speech, handwriting, and other natural language resource (*i.e.*, HLR) consuming organizations throughout Motorola. That is, its *raison d'être* is to help provide leverage and efficiency (and expertise) in the creation and utilization of HLRs (whether those resources originate company-externally or -internally). There are multiple aspects to such service which we break (somewhat arbitrarily) into three groups: basic services, HLR production, and extended services.

2.1. Basic services

These are the collection of services (exclusive of production of new resources) which a HLR center clearly should be involved in. It was felt that an organization (the HLDRC) which *only* performed the "basic" functions of archiving and distributing HLRs (2.1.1), disseminating information regarding available resources (2.1.2), serving as the corporation's interface to external HLR-related entities (2.1.3), and fostering standardization and reusability (2.1.4) would be a wise investment for Motorola, yielding substantial benefits for the company.

2.1.1. Librarian services

The most basic of the "basic services" is that of serving as the company's archivist with respect to HLRs. That is, collecting the existing and future resources, organizing them, documenting them, standardizing them, and doing whatever else is necessary to make them easily available to Motorola consumers of HLRs – in short, being a usage facilitator and an institutional knowledge repository for those resources. Given the geographically dispersed population of users, the "librarian services" are, implemented so as to be completely accessible via the company intranet (on the HLDRC web site). Under this rubric we include most of the "passive" activities typified by traditional libraries – *i.e.*, preparation of resources which then wait for "pull" from a consumer who knows about the library-like repository and takes it upon him-/herself to come look for material of interest – and more proactive dissemination of information about archive contents.

2.1.2. Communication

Probably the most crucial of the "basic services" is that of communication. Besides proactivity with respect to informing people of the company's HLR "library" holdings, the center should inform interested parties of general HLR activities, HLT news of interest, HLR availability, and so on. For the HLDRC, this involves maintaining a web site which we hope will draw frequent visits from its "customer base" simply due to the one-stop utility that it provides. We also "push" potentially interesting information and announcements via distribution of a regular, electronic newsletter.

It is additionally our goal to maintain a much more involved participation in the data requirements and usage of Motorola's HLR consumer organizations. By keeping a finger on the pulse of their activities, we potentially can foster greater cross-divisional collaboration and cooperation. This might include formation of internal consortia to contribute to new data collection efforts of interest to several organizations, finding acquirable data which could be of use to a Motorola data consumer organization and advising them of its potential availability, or even helping connect diverse organizations so that they can better leverage each others' expertise.

2.1.3. Corporate interface to external HLR-related organizations

Another core function of the HLR center is keeping abreast of external efforts to produce HLRs, acquiring resources from external sources (purchasing, participating in consortia, contracting for production), and, representing the company in standardization efforts. The HLDRC should serve as the primary point of contact for HLR-related organizations external to Motorola

The HLDRC is responsible for membership for Motorola as a whole in HLR-related organizations (such as the LDC) and the intent is to take a more active role in making sure that the benefits of such memberships are realized throughout the corporation. Some of those organizations primarily play the role of (non-profit) publishers of information (HLRs). There are other human language data consortia which expect the participants to assume a more active role in the production of the consortial data resources (*e.g.*, SpeechDat-CAR). With the establishment of the HLDRC, Motorola now has an organization whose charter includes such activities (or at least their coordination). Likewise, Motorola may well benefit from having a single center focusing on coordination (and/or relationship building) with universities which have significant activity on the language resources front. In addition to corpora which may not make it into general distribution, many broadly useful language handling / processing tools come out of university (and consortial) projects.

And, finally, there have been and will continue to be various HLR related standards proposals under discussion by the research community at large. Motorola could (and should) play a much more active role in working with the organizations worldwide which decide on public (and *de facto*) standards so that our interests are more effectively represented. This possibility is enhanced by the existence of a central focal point, the HLDRC.

2.1.4. Standardization

And, finally, on the "basic services" front, an organization such as the HLDRC should play an active role in driving the HLR related standardization efforts within the company and serving as an advocate of best practices with respect to HLR production and utilization. It falls to us to track the various practices – *e.g.*, XML for language resource markup, annotation graphs (Bird and Liberman, 1999), and so on – and attempt to determine which represent the long term directions or emerging (*de facto*) standards so that we, as a company, can jump on the right bandwagons (and realize subsequent efficiencies in our HLR usage). We also need to track and interact with organizations which are (to greater or lesser degree) explicitly attempting to set standard practices – *e.g.*, ISLE (International Standard in Language Engineering), EAGLES, the GATE (General Architecture for Text Engineering) project at Sheffield (Gaizauskas *et al.*, 1996), and the American National Corpus effort. Again, company-wide adoption of the most useful of the results of these efforts should be advocated by the HLDRC.

In the HLDRC, we are also defining and working toward implementation of a standardized universal annotation framework which would allow flexible, cross-corpus, web-based access to annotations and towards a distributed data serving architecture which would resolve many of our data availability issues and foster much more efficient utilization of HLRs.

2.2. Human language resource (HLR) production

As noted in the introduction, commercial entities such as Motorola which are in the business of productizing human language technology (HLT) will always have a need to create proprietary HLRs specific to their products. The primary function of the HLDRC (beyond its archivist / information center role) is that of being a center of expertise for high quality design and execution of HLR production efforts. The goal is to be able to cost-effectively deliver results on large scale, multi-lingual, international data collection and annotation efforts, such that Motorola's capabilities are second to none in this area. We, also, need to be ready to meet this requirement for potentially highly variable volumes of work. Therefore, one of the requirements for the HLDRC is to build and maintain relationships with universities, individuals, and companies throughout the world who are interested in (and capable of) performing aspects of our collection and annotation projects under contract, in order to supplement the internal staff.

It should be noted that, even in the case of Motorola-proprietary resources, there is no presumption that the HDLRC would ever become the exclusive center within the company for HLR creation. Product groups and corporate research groups would likely continue to do some data collection or other forms of creation in areas of their particular interest or expertise. The role of the HDLRC *vis-à-vis* those Motorola produced resources would principally be to serve as a librarian and publisher for the collecting organization and possibly to serve in a consulting role for some aspects of their collection / annotation projects, *e.g.*, with experimental design factors or with annotation "best practices" and tools.

Motorola organizations which do not have the resources or desire to carry out projects on their own, but with specific needs, can contract with the HDLRC to produce the resources to meet their requirements. And, when Motorola would enter into participatory consortia (*e.g.*, SPEECON), then the HLDRC would be the organization to carry out the required HLR production, unless some other Motorola group related to our interest in a particular consortium desired to do so.

2.3. Extended services

These represent a collection of activities which naturally could be associated with an organization such as the HLDRC, but which under budget / manpower restrictions might be forgone. From the rich set of possible "extended services", we have singled out tool acquisition and/or creation (2.3.1) and "data research" (2.3.2) as being particularly interesting. We have found that, even in light of resource restrictions, we are reluctant to not act at all with respect to helping locate, manage, and create HLR tools, and we are, at least, looking towards doing a modest amount of "data research" (with the goal of eventually turning it into a primary activity).**Tool acquisition / creation**

The first broad category of "extended services" to be carried out by the Human Language Data Resource Center (HLDRC) is that of monitoring availability of, acquiring, and testing language data related tools from external sources. Many special purpose and general tools for acquiring, manipulating, converting, processing, and labeling human language data have been created and will continue to be created in diverse organizations throughout the world. Many of those tools are public domain or have a minimal price tag associated with them, yet when we factor in the costs for multiple, independent researchers to locate, acquire, install, and test such tools, and then verify their utility for the intended task (and possibly modify them), a substantial price tag is associated with such "free" software. The HDLRC could serve as a clearinghouse / support center for the utilization of such externally produced tools. The HDLRC could (and should) also collect, document, and redistribute generally usable language data manipulation tools which exist in the various parts of the company. In some cases, new development of data manipulation tools (*e.g.*, a tool to convert Motorola's legacy data to current encoding standards) may be called for.

2.3.2. Data research

What we are calling "data research" is, in a nutshell, research into innovative ways to leverage tools and existing resources to (semi-)automatically (or at least much more efficiently) produce new, more useful HLRs. The reality is that, to some degree, HLT researchers have been engaged in this type of activity probably since the dawn of HLT itself. Given the value of large, high quality HLRs and the substantial manual effort traditionally expended in creating them (or, in some cases, serving as a barrier to their creation), it is only natural that HLT researchers have siphoned off some percentage of their time and creativity into doing "data research" activities. Those activities though tended to be viewed as having, somehow, second class status.

In a seminal recognition of not only their legitimacy, but their criticality to the success of the field, veteran speech and natural language researcher Rich Schwartz, as part of his keynote address at the 1999 EMNLP/VLC[4] conference, threw down a gauntlet of sorts. He stated his opinion that, of the worldwide R&D budget for HLT, 10-20% ought to be spent on data resource creation *and* 50% ought to be put into technology for making data collection, preparation, and labeling more innovative, automated, and leveraged. He strongly advocated first class status and high priority to "data research" activities. It was clear that he expected his statements to be controversial, but to the contrary, the audience, which was largely made up of leading researchers in the field, seemed to be mostly in agreement with his position.

"Data research" is the most forward looking, creative aspect of the possible range of activities that we envision for the HDLRC. It is somewhat risky with respect to results (thus the appellation "research"), but has the potential to revolutionize the cost and time requirements involved in human language technology development. And, it is a natural fit for a human language resource (HLR) center to be tasked with carrying out research into how to produce better HLRs faster.

3. Industrial data resource center implementation issues

In this section we will take a look at some of the implementation issues which exist with respect to the start-up and on-going operation of a within-company HLR center such as the Motorola Human Language Data Resource Center (HLDRC).

3.1. Proper positioning of the organization

Placement of the organization may on the surface seem like a relatively inconsequential issue, but it has the potential of being a show stopper, if blatantly incorrect. The danger is that the HLR center will be wholly housed within one of several existing HLR consuming organizations such that 1) it is mandated to focus on serving the needs of local organization to the detriment of the wider needs of the company as a whole, and/or 2) existing cross-organizational distrust issues will not be able to be overcome in the attempted move to full cooperation among interested parties. Ideally, the organization should be placed high enough in the company (or funded from high enough, in the case of a contractual arrangement) to be able to maintain the focus on serving all organizations with HLR needs.

3.2. Staffing

While it may be possible to create one or two new positions specifically targeted at HLR center functions, it would be rare to have *carte blanche* to hire a bunch of people to perform the functions of an organization like the HLDRC. But neither should it be necessary; a company which has need for a HLR focussed organization probably already has many individuals distributed throughout the HLR producing / consuming organizations which spend significant amounts of their time on data resource issues. Ideally, that time can be assigned to the new organization in some form – either by matrix management type strategies (thereby making the organization a virtual entity), or by reassignment to the new organization, or by some combination of these two. Of course, issues abound with respect to making such changes – *e.g.*, (re-)location, organizational affiliation, level of interest on the part of those staff members, resistance to "cannibalization" of organizations, and so on – but, all things can be worked out when there is a will to do so.

3.3. Funding

It would be plausible to assume that funding and resource sharing issues for a strictly corporation internal organization, like the HLDRC, would be trivial relative to those facing an international consortium with industrial, government, and academic participants. The reality, however, is that in large international corporations, there is, of necessity, considerable decentralization of decision making and financial control (and some concomitant insularity). This leads to a situation where the models pioneered by (inter-)national HLR organizations – *e.g.*, formation of consortia to fund particular data collection / annotation efforts, subscription based support, etc. – are profitably retained (or even required).

As with the typical (inter-)national HLR consortium, there will probably always be a need for some degree of operational subsidy. Of course, the superior organization which values the HLR center operation enough to provide such subsidies would be the corporation rather than a governmental entity. Subsidy can considerably reduce operational friction by allowing the HLR center to provide relatively more service to its clients for less expenditure on their part. Under the logical extension of that notion, one could conclude that full subsidy of all costs would be ideal. However, full subsidy would not be desirable for at least two reasons: 1) a fixed budget caps the amount of work that the center can perform; and, 2) it would de-sensitize consuming organizations to the substantial costs involved in producing HLRs for their purposes. Ideally a corporate HLR center would be provided with sufficient funds to cover the costs of providing the basic services, and the extended services at the desired level, and additionally have some pool of money available to use in offsetting the production costs of new HLRs and acquisition costs of externally produced HLRs. The remainder of the operating costs would be derived from (subsidized) contractual production of HLRs for corporate groups which require resources and from the proceeds of serving as a (company-internal) publisher[5] of HLRs.

[4] Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, a conference / workshop held June 21-22, 1999 in College Park, MD in conjunction with the 1999 Association for Computational Linguistics (ACL) meeting. Only the abstract of his speech (entitled "Why Doesn't Natural Language Come Naturally?") was reproduced in the proceedings and it did not go into this topic.

[5] "Company-internal publication" of HLRs simply resolves to a means of equitable (usage related), cross-departmental distribution of the costs involved in acquiring or producing HLRs.

3.4. Resource sharing model

This is one of the most crucial issues in successfully achieving the objectives of a corporate HLR center (and overcoming some of the typical, historical inefficiencies with respect to the resources allocated company-wide to HLR activities). The (perhaps obvious) objectives regarding language data resources as we see them are:

1) We want to satisfy the human language data resource requirements of (internal) consumer organizations as quickly and as cost effectively as possible.
2) If the company has a resource (or can get it at a reasonable cost), we want that resource to be available to all organizations which have a need for it. And…
3) We should be cognizant of the costs involved in producing HLRs, and attempt to be as equitable as possible to all organizations (while simultaneously trying to minimize the barriers to sharing and reuse).

In designing Motorola policy with respect data resource sharing, we are trying to take some lessons from the (inter-)national consortia, such as the LDC and the ELRA, which have successfully lowered the barriers to sharing and reuse of HLRs. A couple of key aspects of their model are 1) avoidance of the need for direct (repeated) producer-consumer negotiations and 2) having a clearly defined, equitable policy regarding resource availability to consumers. When a resource (whether from internal or external sources), which might be of interest to one or more consumer organizations within the company, is first produced or first comes to our attention, the center (the HLDRC, in our case) would attempt to acquire the rights to use that resource for the company as a whole, then make it available to all interested consumer organizations within the company (under standard policies). The HLDRC thus serves as the primary point of contact for external organizations, both data consortia and individual data resource producers, wishing to "sell their (HLR) wares" to Motorola.

With respect to internally produced resources, a corporate HLR center potentially has an advantage over (inter-)national HLR consortia (involving a diverse collection of companies, universities, government agencies, and other consortia) in that it is possible for corporate leadership to mandate sharing under predetermined terms, thus further reducing the friction involved in redistribution / reuse of such resources. Here, equitability issues must be considered – producer organizations which contribute language data resources to be made available to all should be compensated in some fashion. One possibility would be to have some sort of accumulated credit toward future data production efforts targeted to their needs or toward future data resource access (in the case of existing HLRs).

A credit system, of course, presupposes that internal organizations are actually being charged for their utilization of existing HLRs and for production of new HLRs on their behalf. While it is conceivable that a company might absorb all such costs on behalf of HLR consuming organizations, such a "cost-sharing" model (all cost to corporation, no cost to consumers) would be, in my opinion, ill-advised. It ultimately would devalue the HLRs (making consumers less aware of the costs involved in their production) and unbalance weighted decisions about when resource productions should be undertaken. On the other hand, given that one of the objectives of the corporate HLR center (the HLDRC) is to diminish the barriers to sharing of resources, it would be beneficial to (less than completely) subsidize the HLR usage costs for consumer organizations, so that cost is less of a barrier in and of itself. Any collected usage charges could be fed into a fund for acquiring external HLRs (or otherwise *offset the cost of operations*).

3.5. Standardization

Standardization of data formats, tools, naming conventions, access methods, labeling, etc. is a formidable problem even when one has a completely clean slate from which to start. Unfortunately, the slate is rarely, if ever, clean – rather, typically, local conventions, habits, tools, and so on have evolved and become deeply embedded in the HLR utilization process. To the degree that HLT R&D has been carried out in geographically- and organizationally-distributed entities within the company, one would expect the engrained habits and conventions to be even more diverse, and standardization process to be correspondingly more difficult. Nonetheless, the potential efficiency gains from standardization of representations, formats, and so on are so great, that (at least, partial) standardization is a virtual imperative. The situation is not entirely bleak in that there are at least two mitigating factors:

1) Inherent commonality of tasks and requirements – Though each distinct group may have developed their own practices, the shared general goals and constraints will have led to some degree of similarity among their methods, tools, and so on; and,
2) Acquiescence to external "standards" – It is likely that the practices of each of the distinct groups will have been affected, to greater or lesser degree, by "standards" promulgated by various external organizations, such as those which *explicitly set out to produce standards* (*e.g.*, EAGLES, TEC,…) and *de facto* standards set by heavyweight players (*e.g.*, ARPA).

Once again, we have a situation where, at first blush, it appears that options are available to a within-company HLR organization, which could not be employed by public, consortial HLR organization. Namely, it would be possible, in theory, to utilize authority (however acquired) to mandate and enforce acquiescence to designated standards. However, the success of such an approach is dubious, except possibly in the most tightly controlled situations. To the degree that the confederation of HLR consumers is loose and voluntary (*i.e.*, the case of international, open HLR consortia, and probably more the reality than not for most large corporations with HLT activities), authoritarian prescription of standards is less likely to be viable. The remaining option, which admittedly is non-trivial to implement, is to provide clear and compelling advantage to people / organizations who switch to the proposed standard.

Of course, the difficulty of garnering acceptance for proposed standards will also vary with the degree to which the proposal differs from established practice. That is, a secondary, but important, aspect to this strategy is

co-option – to the degree possible making the "new" standard incorporate (at least, be compatible with) the existing, prevailing methods. This reduces acceptance friction and makes an inconvenience minimizing migration plan more possible.

4. Summary

We are in the process of establishing a responsive, comprehensive center for human language data and ancillary services to benefit Motorola business units which have need for such data in their efforts to produce tomorrow's intelligent, natural language based human interfaces for Motorola's products. This organization, the Human Language Data Resource Center (HDLRC), is similar in spirit to the Linguistic Data Consortium (LDC), but with the mission of facilitating access to and supporting utilization of HLRs by all interested parties within Motorola. The goal of this center is to significantly improve the efficiency and effectiveness of the HLR related activities of individual Motorola organizations with great benefit accruing to Motorola as a whole. The range of activities and responsibilities covered by the HDLRC have been broken down into basic services, resource production, and extended services and discussed in detail. We also took a broad look at issues we encountered (and continue encountering) in our attempt to bring the center up to speed.

We feel that the work of Motorola's Human Language Data Resource Center (while not terribly glamorous) is one of the most essential tasks facing us in the endeavor of productizing exciting, market-leading human language technologies (and one in which exceptional execution could provide the deciding advantage). Moreover, it is quite possible that industrial HLR centers such as the one that we are pioneering today may well become the standard model among commercial (and non-commercial) enterprises which have, like Motorola, a broad spectrum of human language technology activities.

5. References

Bird, S. and Liberman, M., 1999. A Formal Framework for Linguistic Analysis. Tech. Report MS-CIS-99-01, Dept. of Computer and Information Science, University of Pennsylvania.

Gaizauskas, R., Cunningham, H., Wilks, Y., Rodgers, P., and Humphreys, K, 1996. GATE – an Environment to Support Research and Development in Natural Language Engineering. In *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-96)*, Toulouse, France, October 1996.