

DELIVERABLE IDENTIFICATION

Identification number	MLAP-364-D-WP2.2
Type	Report
Title	Advanced Distribution Means for Spoken Language Corpora
Status	final
Work Package	3
Task	3.1.4
Period covered	05/95-10/95
Date	7.11.04
Version	1.2
Number of pages	10
Author	Christoph Draxler Institut für Phonetik und Sprachliche Kommunikation Ludwig-Maximilians-Universität München Schellingstr. 3/II D 80799 München
WP/TP Responsible	Institut für Phonetik und Sprachliche Kommunikation Ludwig-Maximilians-Universität München
Project contact point	Dr. Harald Höge
CEC project officer	José Soler
Status	public
Actual distribution	consortium
Supplementary notes	
Keywords	Spoken Language Corpora, distribution, CD-ROM, network, DBMS, distributed file system
Abstract	This report outlines the distribution of Spoken Language Corpora on traditional CD-ROM media and a new approach via network. High capacity CD-ROMs are being introduced, but this is only a marginal improvement in respect to the distribution of SLC. Network access however offers many opportunities: customized SLC, on-line access, and a high degree of protection. However, for network access to be feasible, the bandwidth of existing networks will have to be increased.
Status of the abstract	public

Advanced Distribution Means for Spoken Language Corpora

*Christoph Draxler
Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
draxler@cis.uni-muenchen.de*

1. Introduction

A Spoken Language Corpus (SLC) consists of signal data of recorded speech, a symbolic description of this signal data, and a set of formal procedures to access the data. The Computer Representation of Individual Languages (CRIL) guidelines agreed upon at the 1989 Kiel IPA convention require at least three distinct levels of data representation in an SLC: a digital speech signal, a narrow phonetic or broad phonemic transcription, and a phonemic citation form or orthographic transliteration [4].

SLC are mainly used in the development of Spoken Language Processing (SLP) applications, and as a reference standard in Phonetics, Phonology, and Linguistics research. Many SLC have been developed with a particular application or research goal in mind, they were relatively small, and they were not disseminated widely. The SLC required by current SLP application development and research are quite different:

- they are larger by at least an order of magnitude, and
- they are general purpose corpora.

The production of such an SLC has become a task so large and expensive that new approaches to the corpus design, the data collection, and the distribution have to be developed. Only the distribution is dealt with in this report – descriptions of design and data collection procedures can be found in the forthcoming EAGLES handbook [2].

The distribution of SLC has an administrative and a technical side. The administrative side relates to the contractual agreements between the SLC producer, distributor and the customer. Organizations such as the Linguistic Data Consortium (LDC), the European Language Resource Agency (ELRA), the Bavarian Archive for Speech Signals (BAS), and other national institutions have been established as legal bodies for the distribution of SLC. The technical side relates to the form in which SLC are made available. In this report the current means of distribution are described briefly and new approaches are outlined.

2. Classification Dimensions for Spoken Language Corpora

SLC can be described by the following dimensions:

- Size
- Content
- Access
- Stability
- Storage and format

The *size* of an SLC can be given as a physical measure such as in Megabyte, CD-ROMs or tapes.

The *content* of an SLC consists of a description of the data held in the corpus, the type of speech recorded and data on the number of speakers, sentences, recordings, the duration, etc.

Access can be divided into

- Off- vs. online access
- Partial vs. full access
- Restricted vs. unrestricted access

Access to an SLC is *off-line* if the SLC is a copy of the original SLC. This copy is usually held on a storage medium on a local machine, e.g. a CD-ROM or tape. In *online* access, the original SLC is accessed, usually via a local or wide-area network.

Full access to an SLC means that the entire SLC is available for access, *partial* access means that only selected fractions of an SLC are available. For example, in the SpeechDat corpus a partial access to the corpus can consist of all credit card number prompts only, or the read sentences.

Restricted access to an SLC means that only specific queries for data are allowed on the SLC (which may be partially or fully accessible). Access restrictions are defined in contracts, they consist of access privileges, and require an accounting of the access. For example, in the SpeechDat corpus, access to the corpus is restricted in such a way that the project partners may access all corpora for the cost of the medium immediately, and others may access the corpus if they pay a licensing fee and one year after the project has ended at the earliest.

An SLC can be *static* or *dynamic*. A static SLC, once established, it is not changed (or only changed with the release of a new version), whereas in a dynamic SLC the user accesses the most recent corpus without being notified explicitly of modifications. However, a modification record has to be kept to inform users of changes to the data.

An SLC can be stored *centrally* in one storage location, e.g. a computer centre of a research lab, or *distributed* over several geographically distant storage sites. In a distributed corpus, the user sees a virtual corpus and is not necessarily aware of the storage location of corpus fractions or subcorpora. For example, a global SpeechDat corpus could be accessed online via the ELRA with the subcorpora of each language held in the institutions that created them.

The data on the different levels of representation of an SLC can be stored in a *standard* or a *proprietary* format. Currently there exist proposals for general standards for SLC such as SAM, NIST, or CRIL, but these standards cover different representation levels and are specified to varying degrees of precision. Basically, all data should be held in machine readable form; administrative and phonemic or phonetic data should also be stored in such a way that it is readable by humans. Data can often be compressed in order to reduce the amount of storage space needed. There exist general and task-specific compression programs. The most common all-purpose compression program is gzip which is distributed as part of the GNU software library. It is available for almost any hardware platform and may be used free of charge. Task specific compression programs promise a better performance for specific kinds of data, e.g. audio data. A well-known audio file compressor is the Shorten program of Tony Robinson (<ftp://svr-ftp.cam.eng.ac.uk/pub/comp.speech/sources/shorten-2.0.tar.gz>). Task-specific compression programs perform well when applied to suitable data; for other data, they may fail miserably. Proprietary formats may be used to protect data, e.g. through encryption.

3. Current Spoken Language Corpora

A selection of well known SLC is described in *table 1* using the dimensions of the preceding section.

Corpus	Content	Size	Access			Storage	Format
			full vs. partial	on- vs. offline	restricted vs. unrestricted		
TIMIT	read sentences	1 CD	full	off-line	unrestricted	-	uncompressed
PhonDat	read sentences	7 CDs	full	off-line	restricted	-	uncompressed
Verbmobil	negotiation dialogues	9+ CDs	full	off-line	restricted	-	uncompressed
SpeechDat	prompted telephone speech	11 * 2 CDs	partial	off-line	restricted	-	compressed
TED	conference speech	7 CDs	full	off-line	unrestricted	-	compressed

Tab. 1: Spoken Language Corpora Classification

4. Current Storage Technology and Distribution Means

4.1 Storage and data transfer standards by 1996

In 1996 standard high end PCs or workstations in an SLP research or SLP application development lab are equipped with fast storage devices (usually magnetic hard disks) of 1 to 10 GB for PCs or desktop workstations, 10 GB for workgroup servers, and 100 GB for departmental servers. Jukebox CD-ROM drives that hold a large number of CDs are available at least on a departmental level, and there is at least one dedicated PC or workstation for burning CD-ROMs.

A typical desktop PC or workstation in an SLP lab has 16 to 32 MB of main memory and is equipped with a single quadruple-speed CD-ROM drive with a data transfer rate of approx. 600 KB/s and medium speed (< 10 Mb/s) local network (Ethernet) running TCP/IP under UNIX, NovellNet under Windows, or AppleTalk on Macintosh. The lab itself is connected to the outside world via primary rate ISDN connections (30 64 Kb/s data channels plus 1 16 Kb/s control channel), and high speed (> 100 Mb/s) fibre optic networks connect the major computing sites in Europe and the US.

The main higher level network protocols are *ftp* and *http* which are both based on TCP/IP.

4.2 Current Distribution Means

Currently, SLP corpora are distributed on CD-ROMs together with a symbolic description and access software. Customers have access to the entire CD-ROM and the use of the CD is governed by a licence contract.

When only a relatively small number of copies (~30) is necessary these CD-ROMs are generally produced individually at an SLP lab using a CD-ROM burner on a dedicated PC or workstation. Typically, burning a CD on a double-speed CD-ROM burner takes about 30 minutes and requires some supervision by a trained technician. When more than 50 or so copies are needed, it can be cost (and time) effective to press them.

A CD-ROM burner costs approx. \$ 2000.-, the software \$ 1000, and the high-end PC or workstation to which it is connected \$ 4000. The CD-R medium costs approx. \$ 7 each in quantities of 100 (prices are coming down at about 25% per year). With labor costs of \$ 30/hr and approx. 1000 CDs produced during the life-time of the burner, the cost per CD-ROM sums up to \$ 29.

The mass production of CDs is only possible in special CD production plants. Here, a glass master CD is created from the original data source (tape or CD-ROM). From this glass master, metal masters can repeatedly be made. These metal masters are then used to press the final CD. The cost of mass production are made up of fixed costs for mastering, setup and producing the glass master, and volume-dependent costs for media and handling. Typically, the fixed costs are approx. \$ 1500, and media costs are \$ 2.50 per CD for 100 CDs.

5. Advanced Distribution Means

Advanced distribution means for SLC will on the one hand rely on improving the traditional distribution via CD-ROM, and on the other hand on developing and deploying a network distribution infrastructure.

5.1 CD-ROM

The following improvements to CD-ROM distribution relate to different aspects of the distribution: high density CD-ROMs allow more data to be stored on a CD-ROM, encrypted CD-ROMs allow a protection of data and can be used to restrict access to SLC data, and customized CD-ROMs allow the creation of partial SL corpora.

High density CD-ROM

In 1995 the first laboratory prototypes of high density CD-ROMs with a capacity of 1.2 GB were presented which use a blue laser instead of the the traditional yellow laser. The shorter wavelength of the blue laser allows smaller pits and lands on the reflective surface of the CD-ROM medium, hereby increasing the storage density.

Another approach to augmenting the capacity of CD-ROMs is to use both sides of the disk for storing data. Technically this is not a big problem, but it will no longer be possible to use one side of the disk for a label. As an alternative, double-sided CD-ROMs could come in cartridges similar to today's CD-caddies which can be labelled; however, this would require new CD drives and is thus unlikely to be accepted.

Finally, multi-layer CD-ROMs can store data in different layers on a disk. These layers are selectively accessed either through changing the focus of the laser, or through changing the wavelength of the laser.

In September 1995 an agreement on a single high density CD format was reached. In this format, high density CDs (also called Digital Video Disk, DVD) will be single-sided CDs with up to two layers for a total capacity of 4.7 GB of user data (double-sided CDs are also covered by the standard, but as a future option only). The physical dimensions of the DVD will be the same as that of the common CD-ROM, and DVD players will also accept CD-ROMs.

Encrypted CD-ROM with keys

Many software packages, font and clip-art collections, and other large amounts of data are currently being distributed on encrypted CD-ROMs (often also called *installer CD*). Data on these CDs can only be accessed with the appropriate key, which is made available under a licence contract. The key is valid only for subsections of the CD-ROM; in general, these subsections correspond to directories or files on the disk.

Encrypted CD-ROMs allow the cheap mass production of CDs while at the same time retaining control over who may access the data. However, since the CDs are mass produced, the same key will work for all copies of the CD. Hence, great care must be taken to prevent the proliferation of keys.

Depending on the implementation of the encryption scheme it may be possible that the encrypted data on a CD may only be accessed through specific software. This requires that the original CD be used, and that access to the data is limited by the capabilities of the access software.

Customized CD-ROM

Often customers are interested only in fractions of SL corpora, e.g. all digits or chains of digits of telephone recordings, all male speakers, etc.. In such cases, customizing a CD for a particular customer offers significant advantages over mass produced CDs:

- data can be encrypted and a key can be assigned so that it is valid only for this CD-ROM,
- the customer pays only for the data he or she is interested in, and
- data from more than one source can be placed on a CD.

A key unique to a particular CD-ROM has the advantage that if copies of the CD are made it is still possible to track down the original CD from which the copy was made. Again, as with mass produced encrypted CDs, the encryption is an additional processing step which causes administrative overhead and possibly performance degradation.

Customizing CD-ROMs is only feasible with a high-level description of an SLC which allows the selection of the data of interest. Placing data from more than one SLC on a CD requires a global data model of the corpora used.

5.2 Network access

The Internet has become a major means of access to data. Making SLC available via the Internet (or other global communication networks) is substantially different from SLC distribution via CD-ROM in that it offers

- on-line access
- flexible specification of data requests
- access to data independent of storage locations
- fine tuned accounting and access control

These features have for some time been offered on the Internet, e.g. in library catalogues, but they have until now not been available to the distribution of SLC because of the large amount of data involved, the lack of a uniform data model for SLC, and because the need for such services was met with data distributed on CD-ROM.

Technology & Standards

The most widespread and best accessible global network is the Internet. The TCP/IP protocol is the basis for a whole range of protocols from distributed file systems, e.g. the *Andrew file system*, file transfer protocols, e.g. *ftp*, information retrieval protocols, e.g. *gopher* and *wais*, to hyper-text transfer protocols, e.g. *http*.

These protocols are fully sufficient for the network distribution of small and fully accessible SLC. However, the following issues have to be addressed for the network distribution of large and less open SLC.

Legal issues

Licence contracts have to be modified to cover access to data over a network, i.e. not being provided on a stable medium. The following questions have to be answered:

- Who may access the data through which network connection?
- Who is responsible for a safe transmission of the data?

The first question relates to the problem of *identity* in a network. In the Internet a machine has a unique address; users have a name and possibly a password. However, an impostor can easily forge both the machine address and the user name – a harmless version of this is the anonymous account provided by some Internet sites.

A solution to the problem of identity is to encrypt the data and to provide keys only to those users who are entitled to them. (Note: the de facto standards for encryption, RSA and PGP, are regarded as military secrets by the US government and may thus not be used outside the US. PGP however has nonetheless been available outside the US: it may legally be used if the key length is limited to a certain number of bits; furthermore, the book describing the principle and giving sample code is widely available).

The second question relates to the problem that a) the network itself may not guarantee the correct transmission of data, and b) an eavesdropper can filter network traffic and keep copies of data. Issue a) depends on the network and the transfer protocol in the network, issue b) can be solved by encryption.

Access control

The right to access data and the mode of identifying a user have been dealt with in the Legal issues section. Access in this section refers to the access to the data itself:

- Which fraction of an SLC may be accessed?

In the licence contract, a user is given a right to access specific fractions of an SLC. In order to be able to formulate access requests, the contents of the SLC have to be published in the form of a data model. This data model describes in all necessary detail the logical structure of the data. It is also possible to provide different views of an SLC to different users.

Once a user has formulated the access request it has to be checked that this request does not violate the licence contract. This check is performed by the SLC server.

Accounting

With network access to SLC two accounting schemes can be employed:

- pay per query or
- pay per item retrieved

According to the first scheme, each time a user formulates and executes a request, he or she is charged for the login time or the runtime it takes to evaluate the request. This scheme is often used in applications where a success of a request is not known beforehand, e.g. whether a book is stored in a bibliographic catalogue at all.

The second scheme is more appropriate for requests where it is known that the requested data is in the database.

Finally, two modes of payment can be distinguished:

- pay cash or

- pay per exchange of data.

Paying per cash can further be subdivided into traditional payment via banks and paying electronically. Currently there are several electronic cash schemes under development which aim at providing a safe and anonymous transfer of money electronically; however, none of the solutions developed so far has been widely accepted.

In the scientific community data is often paid for in exchange for other data. This is especially valuable in SLP where the results obtained from one SLC may be used to extend the original SLC and thus augment its value.

5.3 Distributed File System

In a distributed file system the distributed physical storage of files and directories is represented in a uniform logical addressing scheme. The most wide spread distributed file system is the Andrew File System developed at CMU [3]. It extends the UNIX file system by making visible file systems of machines anywhere in a large Wide Area Network or the Internet.

A distributed file system offers all features of traditional file systems: read/write access to files, file operations such as copy, move, rename, etc., permissions for restricting access to files, and file based accounting. Safety within a distributed file system is ensured through encryption.

Apart from handling data distributed over geographically distant machines distributed file systems do not provide any new features relevant to the distribution of SCL. Hence they can be seen as a first step in the direction of network access by allowing users from remote machines to access local files as if they were on their own machine.

5.4 DBMS access

A database management system (DBMS) provides an application independent access to the data it stores. This independence is achieved by separating the logical structure of the data, called the database schema, from all operating system or file storage details.

A database schema is represented in a data model. The relational model, introduced in the late seventies, is currently the most widespread data model and is supported by all major operating systems and programming languages. The object oriented data model is now being introduced, but it has not yet reached maturity, in particular because of the lack of a common database language.

A DBMS has total control over the access to any data stored within the DBMS. It allows the definition of access privileges on different levels of detail, and the access privileges can be based on content. For example it is possible to restrict the access privileges of a given user to particular relation tables, subtables within these tables, or even to single tuples or attributes (the same is true for object oriented DBMSs).

Current DBMS implementations are multi-user systems with network access capabilities. Hence, they can very well be used for storing and making accessible in a controlled way SLC. However, in order to fully exploit the capabilities of DBMSs, it will be necessary to define a database scheme for SLC; this scheme must be general enough to allow the extension of the SCL being modelled, and precise enough to cover all possible data requests from users.

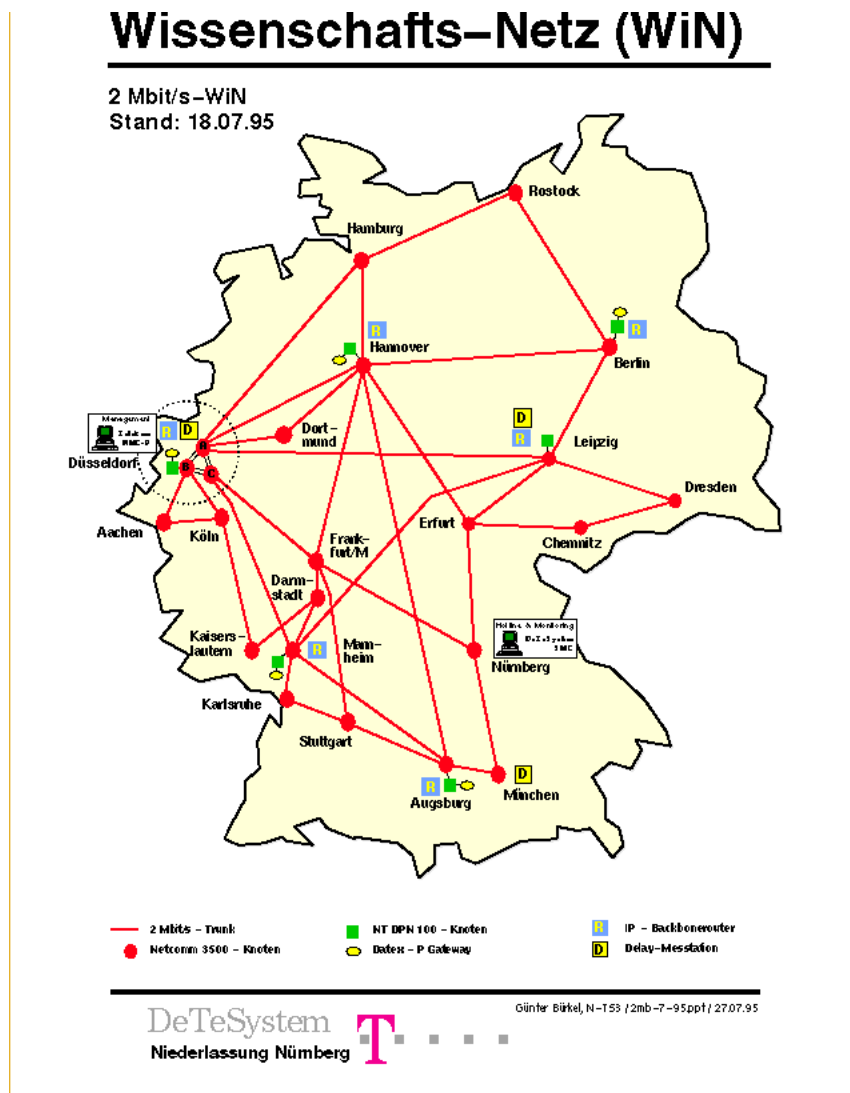
For an example of a Prolog-based DBMS for the PhonDat/Verbmobil speech data, see [1].

6. EU and world-wide infrastructure

Currently, i.e. 3rd quarter 1995, the Internet is the most accessible global network next to the telephone. Due to its recent growth both in user numbers and high-volume data transmissions (WWW, video-transfer, net-phone) it faces serious bandwidth problems.

The Internet is structured hierarchically by participating nations. There is an international gateway in each country; these gateways are connected to each other directly over fixed high-speed links, e.g. fibre optic lines. Within each country major computing centres are connected to each other and the national gateway. SL laboratories usually have a local area network as part of a university or industrial network, which again is connected to the major national computing centres.

In most EU countries, the national infrastructure consists of high- to medium-speed links based coax copper cable or fibre optic links; some centres are connected by ultra-high speed (> 2 Gb/s) links. Fig. 1 shows the current network infrastructure of the Wissenschaftsnetz („science network“) in Germany (more information can be obtained from <http://www.dfn.de/>)



The same is true for the US (currently, the national backbone is a so-called T3 line with a bandwidth of 44.7 Mb/s) and the well developed countries in Far East (Japan, Hong Kong, Korea, Taiwan) and Oceania (Australia and New Zealand).

In the US, the Internet is now run by a commercial consortium of the Telecoms MCI and Sprint, and the Internet service provider America Online. The National Science foundation is now funding the deployment of a Very High Speed Backbone Network Service (vBNS) with an initial bandwidth of 155 Mb/s and 2.5 Gb/s in 1998. This vBNS will be available to academic sites first, but it can be expected that commercial sites will follow soon after.

ISDN is available nearly everywhere in the EU by now, and intercontinental links to the US and Far East are being established. In the US itself, ISDN is not yet widely available. Currently, ISDN offers a rather low bandwidth of 2 Mb/s on a base rate interface. The main problem of ISDN is that it is not fully standardized, e.g. for coupling more multiple data channels for a greater bandwidth.

7. Summary

CD-ROMs will persist for the next three years as a distribution means for SLC because of their

- robustness,
- the ease of handling, and
- the good price/capacity ratio.

Network access is a promising approach to the distribution of large shared SLC.

- In the short term network access can be based on distributed file systems that control access on a per file basis
- In the long term database management systems will be used for the network distribution of SLC.

A DBMS offers the following advantages over traditional means of distribution and distributed file systems:

- a unified data schema in a formal data model
- independence of storage aspects from the logical view of the data
- multi-user and networking facilities are built-in
- accounting and access control on a content basis

Network distribution requires a high network bandwidth; the current transmission rates of 2 Mb/s (offered by ISDN), or shared 44.7 Mb/s lines as in the Internet, are not sufficient. The vBNS will provide sufficient bandwidth once it is installed in the US and in Europe.

References

- [1] Draxler, Chr.: Introduction to the PhonDat-Verbmobil Database of Spoken German, Practical Applications of Prolog Conf. (PAP) 95, Paris, 1995
- [2] EAGLES HANDBOOK – Standards and Resources for Spoken Language Systems, EC-DGXIII, LRE, LRE-61-100, 1995
- [3] Howard, J.H., Kazar, M.J., Menees, S.G., Nichols, D.A., Satyanarayanan, M., Sidebotham, R.N., West, M.J.: Scale and Performance in a Distributed File System; ACM Trans. on Computer Systems, vol. 6, Feb. 1988
- [4] IPA. (1989): The IPA Kiel Convention Workgroup 9 report: Computer coding of IPA symbols and computer representation of individual languages. *Journal of the International Phonetic Association* **19**, 81-82.