

DELIVERABLE IDENTIFICATION

Identification number	LRE-63314-D3.2.1.1
Type	Technical Report
Title	Tasks of a European Center for Spoken Language Resources
Status	Final
Deliverable	D3.2.1.1
Work Package	WP 3
Task	Task 3.2
Period covered	Dec 1994 - Jan 1995
Date	31 July 1995
Version	Final
Number of pages	17
Author(s)	J. Mariani (LIMSI-CNRS)
Work package (WP) / Task (T) responsible	WP3: H. Höge, J.-M. Dolmazon / T3.2: J. Mariani
Project contact point	Harald Höge, Siemens AG, ZFE T SN 5, D-81730 München Phone: +49 89 636 53374, Fax: +49 89 636 49802 E-mail: hoege@habicht.zfe.siemens.de
CEC project officer	José Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	
Key words	
Abstract	
Status of abstract	

Received on	
Recipients catalogue number	

DOCUMENT EVOLUTION

Version	Date	Status	Notes

Mlap SPEECHDAT project
Deliverable 3.2.1.1.

Tasks of a European Center for Spoken Language Resources (ECSLR)

J. Mariani
LIMSI-CNRS

I. The need for a European linguistic resources infrastructure

It is now widely acknowledged that there is a need for Linguistic Resources (LR), both spoken and written language, which are essential to develop linguistic engineering systems and to conduct research aimed at improving the performances of such systems, in order to make them usable and widely available on the market. The US LDC paved the way already some years ago. The LDC experience shows the need for linguistic resources, but it is hardly conceivable that LDC acts as the only worldwide entity gathering and distributing linguistic resources for all languages, including the European ones. Multilingualism is actually a European communication problem but also the strength of Europe, and appears as a barrier against the intrusion of linguistic engineering products from abroad. There are many commonalities between spoken and written language resources, including terminology. This pleads for a single distribution agency, together with networks of production and validation centers for the 3 kinds of resources: spoken language, written language and terminology. We propose a non-profit entity, having permanent staff, to play this role, in coordination with the 3 on-going Mlap projects and with existing bodies (CEC projects, national efforts and international committees).

II. The US Linguistic Data Consortium

The US was the first to understand this need and launched the Linguistic Data Consortium (LDC) in 1991. LDC received a start-up funding of 5 M\$ from ARPA, and has a 20 M\$ budget over 4 years. LDC has a strong relationship with the ARPA Human Language Technology (HLT) program, and provides data to be used in this program to develop and test Linguistic Engineering systems. LDC has members that pay an annual fee (2 K\$ for public research, 20 K\$ for industry, 200 K\$ for industry wishing to seat in the Management Board), and in return get linguistic resources for their own use. In 1994, LDC has 47 databases (DBs) (33 speech and 14 text) and 73 members (51 public research and 22 industry) up from 39 members in 1993. The LDC is hosted by the University of Pennsylvania and its Management Board comprises two representatives of the Academic world, two representatives of government agencies (ARPA and NIST), and two industrials (Nynex and Texas Instruments) which pay 200 K\$ yearly to seat there.

It should be highlighted that LDC distributes 5 databases (2 speech and 3 text) coming from European centers, and that the number of European members went from 9 in 1993 (6 public, 3 industrial) to 22 members in 1994 (19 public, but still only (and fortunately) 3 industrial).

III. The need for a European Linguistic Resource infrastructure

The above numbers show the interest of European research laboratories and industry in such an entity, for the acquisition and distribution of linguistic resources. But more generally for a general European infrastructure for Linguistic Resources production, validation and distribution. It appears that the European industry needs resources in order to build systems that will be used by customers from the various European countries. This effort should be properly funded but should also be coordinated in order to be efficient and to meet the needs of European industry and public research.

It also raises the concern that the management of linguistic resources would be done worldwide by the US, which seems to be unacceptable for various reasons:

- *political:*

- Part of the mandate of European Union governments is to preserve the national culture of each member country, and language is a large part of this culture. The Commission of the EU has a special responsibility to provide a linguistic framework for all countries which are members of that community. Giving this responsibility to a third party outside the community would appear as a "constat d'impuissance". Also, there is a danger that only the most profit-making language (ie those with a large speaking population) may be covered. The minority languages should also be supported for political, not only economical, reasons within the EU.

- *economical:*

- If the linguistic resources are managed by a US entity, they will be provided first to industries which are members of that entity. This puts the US industries in a very strong position to address not only their own market in the languages which are spoken in the US (mainly American English, but also American Spanish), but also to attack the European market.

- The present situation is that exchanges of corpora are made on a language to language basis. In the recent Polyphone international speech database action, the idea was that each country would bring its own language and get the databases for the other 11 languages from the 11 other participants. This obviously puts the LDC, and the US, in a very strong position as they can get 11 languages and distribute those languages to their members, which can develop recognition systems for those languages while none of the other participating countries has any LR distribution infrastructure, or any company in a strong position, compared with the US ones, on this technology. Having such a European infrastructure would mean the mastering of all European languages, which are numerous, and a possible exchange with the US on a much stronger, united basis. Why should the US receive 10 times the amount of data they produce ?

The conclusion is that there is a need for a European distribution infrastructure. And that this European infrastructure should be supported by the CEU and by the various European governments. But this infrastructure, which will compete to some extent with the LDC (which already exists, has strong experience in the field, many members and LRs and has an efficient, government supported, single country, american-style management), should also attract interest from the public research laboratories and from the industrial companies (European and non-European). The official statement of LDC is that they want to keep the LRs US market, but that they would be pleased to have a European counterpart to discuss with.

IV. Multilingualism: the strength of Europe

The main interest of such a European entity will actually be Multilingualism. Several examples in the past show that Multilingualism is a very important factor for the development of linguistic engineering products (the localization of the DecTalk speech synthesizer never succeeded for other languages than American English, the Polyphone action, launched by LDC within Cocosda (International Coordinating Committee on Speech DBs and Speech I/O Systems Assessment), has some difficulty finding european volunteers for doing the recordings, and even more difficulties for sharing the resulting DBs). Having a European agency for LR distribution relying on a network of sites in the different European countries for DBs production would put Europe in a very strong position, easing and

coordinating interaction with the already existing LDC. Also, the position of the EU towards the European Central and Eastern countries is a major advantage, and would allow for enlarging the number of languages covered, and the corresponding systems to a larger set of countries, ie a larger economic market.

In this regard, it is obvious that this ability to address more languages within a European infrastructure will attract not only the European laboratories, but also the American industries. And the policy will have to be carefully defined, taking into account the funding obtained from the CEU/European governments vs the funding obtained directly from the members, including the non-European ones. Having resources of different status (public vs restricted distribution) will probably solve this problem. But even for publicly available data, it may be wise to sell resources on an individual basis in order to avoid the possibility to get multilingual data at the same price as one would pay for monolingual data.

V. Commonalities between Spoken and Written Language Resources: a single Distribution Agency

Our point of view is that there is a close relationship between spoken and written language databases, and also a large difference. Speech databases are specific as they include the signal corresponding to the pronunciation of utterances by various speakers in various conditions. These databases are used for example in order to build acoustic models for speech recognition systems. But there is also a need for written data, corresponding to the kind of utterances which are pronounced in the target application. These data can be for example newspaper texts, if text dictation is aimed at, or transcribed dialogs, if vocal dialog is aimed at. In the first case, the texts are already available in large quantity, in the second case, the transcription is to be made either manually or by semi-automatic methods. The third kind of data is spoken language lexica, comprising the various pronunciation of each word, depending on the speaker, the style, the accent or the dialect. Those two last types of data are very similar to those that could be used and distributed for written language processing. While the technical aspect of the pure speech signal databases is different from text, the legal and business aspects are closely related. Finally, databases aiming at Optical Character Recognition include an image part, and a textual one, and have many commonalities with speech. In the US, the LDC is conscious of this close relationship and has addressed from the very beginning both textual and vocal databases. The US National Institute for Science and Technology (NIST, formerly National Bureau of Standards (NBS)) also produces and distributes both speech, text and image (characters) databases, through NTIS (National Technical Information Service). However, there may also be different written language resources, such as dictionaries, that may be permanently updated, and that may need direct agreement between the users and the producers to be made fully or partly available.

VI. The European Linguistic Resource Infrastructure: a Distributed Production Network, a Network of Validation Units and a Centralized Distribution Agency

The proposal would thus be to create a central distribution agency for distributing spoken and written language databases, and eventually terminological ones, and a distributed production network, built from the on-going 3 Mlap projects on linguistic resources (Speechdat, Parole and Pointer). Each database production infrastructure (speech, text, terminology) should address the same topics with a similar structure: short-term database, long-term databases, use of common standards and tools and quality evaluation with a distributed network topology including the centers that would have the expertise to produce databases in various countries for different languages (in Europe, but also elsewhere). Validation of databases should be conducted in "clearing houses" situated in the various countries, as it appears that mastering a given language is necessary in order to validate a database in that language. This validation centers network can be started with centralized validation units, mostly addressing language-independent validation, or hiring

a technical staff mastering, preferably as native speakers, the languages of the resources to be validated. The distribution should be organized as a centralized entity, in order to facilitate the access to information and the survey of existing resources, at a single point. It should also ensure the quality of resources, in close connection with clearing houses for various languages, but also on its own, or by subcontracting approved validation centers, for language independent quality assessment. It will take care of both spoken and written language resources. This agency should comprise a permanent staff including an executive director, administrative, legal commercial and technical staff members. This structure should be independent from any specific laboratory or industry or government or commission directorate, but should report to a Council comprising the executive director, representatives of the speech, text and terminology production entities and representatives of the European public research laboratories and of the industry. Like the LDC, the distribution agency will distribute data to paying members and/or subscribers, both from the industry and from public research. It should get support from the CEU, and may also get support from the various EU governments in order to ensure a proper coverage of their language.

VII. Legal status: founding a European Linguistic Resources Association (ELRA)

The distribution agency should be independent in order to have the freedom to provide the laboratories, both public and private, with the linguistic resources they need, but should report to a European Linguistic Resources Advisory Committee in order to respond to the European needs as a first priority. Several possibilities have already been checked by I. de Lamberterie (Institut de Droit/CNRS), as a subcontractor within the Relator project. A study has also been independently conducted by O. Caisou-Rousseau at DGXIII. The best would be to have a European status, in order not to belong to a single country. A possibility would be a European Economic Interest Group (EEIG). Unfortunately, those entities should have unlimited liability, and many possible participants, such as universities, are very reluctant to participate in an unlimited liability structure. Another possible status would be to start a European non-profit Association, the status of which is presently discussed at the European level. Such associations exist in Belgium, but need to include a Belgian partner, and this brings a constraint. The proposal is thus to start a non-profit association in one EU country (and France was proposed, as I. de Lamberterie is from a French laboratory and has an excellent knowledge of the French law) for an initial period of 5 years. This association would have institutional members from various European countries, and would be modified to an European Association, as soon as this legal status exists. The council of the Association will be made up of founding, elected and supporting members, both from industry and public research. The availability of data will be open to non-members and to non-European laboratories, through a yearly subscription scheme.

VIII. European Linguistic Resources Advisory Committee

The complete scheme covering both production and distribution, should be headed by a LR Advisory Committee, including the members of the Council of the Association, representatives from the CEU and from each European Government involved in this LR structuring effort. This committee should be independent from the association, but, as it provides funding, may decide to adapt the funding to the quality and success of the action.

IX. Permanent staff in the Agency

It seems that a staff of 8-10 permanent employees would be sufficient for the distribution agency in a European framework (LDC has about 6 employees. In The Netherlands, Spex has 4 permanent employees (plus 11 part-time people employed in 1993)). This would comprise the Executive Director, an administrative assistant, a secretary, a research coordinator, an industry coordinator, a CEU/EU Government coordinator, 2 technical

assistants and a half-time lawyer. Initially, however, a set of 3.5 people would be sufficient: an Executive Director, a secretary, a technical assistant (for database edition on CDROMs or on other media) and a half-time lawyer.

X. Getting funding and material to distribute

The agency may get support coming from existing projects, from the CEU, from European governments and from its own members. It may be difficult to get financial support from EU governments, and this possibility should be considered as optional. The internal funding mechanisms will include both a relatively modest yearly membership fee (that may be similar for public research and industry), a yearly resources subscription fee (different for public research and industry, and open to non-members) and the sale of resources at a subscriber price, or at a higher non-subscriber price. The sales income will partly go to the Agency, and partly to the resource producer (non-linearly, in order to first reimburse the distribution investment). All profit made by the agency should be reinvested in the production of data, according to the priorities defined by the Council of the Association. The initial databases will be obtained from past or on-going CEU or EU national projects, and from donations of public research or industrial corpora (they could get free membership in return). There should be a strong policy from the CEU that all data developed within CEU programs should be given for distribution to the Association, according to specific distribution agreements, and that all databases should be made in agreement with the standards defined by the corresponding committee. The Agency will also have the role of broker for data that owners do not wish to make publicly available, but are ready to trade on specific conditions. This will also bring revenues to the Agency. The interest of LR providers is to have an agency promoting and advertising their data, validating and formatting them, and ensuring the proper study and defense of the IPR allowing for a return on the investment made to create the data. The interest of LR users is to have access to a large set of multilingual LRs through a single entry point, to get validated and formatted data, with cleared-up IPR, and to influence the choice of the kind of data to be produced, in case of a positive budget balance, through the Council of the Association.

XI. Starting-up the Association and the Agency

The Agency should be established quickly, but needs personnel and facilities (offices, computers...). One possibility would be that the Association first subcontracts, as soon as funding is available, already existing public institutes and/or private publishers or companies working in this area, in order to ensure data gathering, validation and distribution. However, a new, independent Agency is preferred from the very beginning, or at least aimed at in the medium term, that would cover all acquisition and distribution aspects, in close relationship with production and validation centers.

Annex 1: European Linguistic Resources Infrastructure

Annex 2: Answers to a questionnaire on the European Linguistic Resources Infrastructure

Annex 3: Possible databases for distribution

Annex 2: Questionnaire on The European Linguistic Resource Infrastructure

This questionnaire has been sent to the 17 Speechdat partners (8 industrial, 8 public research and 1 EFTA partner) and to the 8 Relator partners, associated partners and subcontractors, after the presentation of the Relator draft proposal for a European Linguistic Resources Infrastructure made at the Speechdat "kick-off" meeting in Garmisch-Partenkirchen on October 10-11, 1994.

1. There should be a European alternative for LDC ?

- *DEFINITELY YES !!*

- *Yes*

- *yes*

- *For the time being yes, later on a merge might be possible, although recent developments in the US may put LDC in a bad position.*

- *Yes*

- *Yes, although LDC might start to cover European languages, we need a mechanism for commissioning new databases for European use. (I suppose it might be possible to separate the commissioning from the distribution, in which case LDC might be given the job of distribution)*

- *yes*

- *Yes provided it gives clear benefits:*

i) *The collection of databases for European languages are placed at higher priority than LDC would give them.*

ii) *It attracts more European funding and therefore the Databases for European Languages are generated more quickly.*

iii) *The European alternative can be more responsive to the needs of European Industry and Academic Institutions.*

Note that we should not alienate the LDC - the aim is not competition or protectionism - it is to progress European needs more effectively.

2. This entity should cover both spoken and written resources (and possibly also Terminology) ?

- *it seems to be the best solution to cover both aspects (so many initiatives to try to get them funded and clearly the two basic components of the Linguistic Engineering field)*

- *Spoken + written + terminology*

- *in due course*

- *Not necessarily. It might be much more practical to have three separate centres that work together under one joint structure (Association or otherwise)*

- *if possible, SLR and WLR*

- *I suggest that written resources, etc, should only be part of the centre if they can be dealt with at no extra cost - they need to bring their own financial resources with them. The centre may well be very poorly financed and it should not stretch itself beyond its capacity.*

- *yes*

- *Only if it is efficient to do so. Spoken resources are the priority. There may be cost savings by combining language with speech but:*

i) *the two do have some different needs and obtaining agreement from both communities simultaneously may slow things down*

ii) *there is less industrial interest in language technology at present, depending on the funding arrangements to begin with, speech may end up subsidising the language side.*

I think that the needs for Terminology are quite different from those of speech and language and that terminology could be handled quite easily by other routes.

3. What about character recognition ?

- *NOT mandatory ... but written processing ... so that it could be valuable*

- *no strong feelings*

- *useful*

- *No need yet, let them solve their own problems first.*

- *isn't this only a technology of WLR (as is speech recognition in SLR)?*

- *See 2.*

- *maybe, why not*

- Not yet.... too much else to think about.If it can be picked up at a later date once the centre is running then that is fine but it is not a priority and might complicate things.

4. The distribution infrastructure should be centralized ?

- Probably YES : for querying info on the whole, to have identified points of contact ...; etc
- centralized
- or at least with well defined nodes
- Yes, although the owner of a certain database should always keep the right to distribute his material himself.
- yes, for visibility and responsibility reasons; it should have powers to enforce rulings on associated providers
- I'm not sure. Technical support might be better from national distribution locations. The centre might still process orders and invoices even if the discs themselves came locally.
- yes
- Yes. It is the only way which will really work.

5. The production / validation infrastructure should be distributed in European countries, with separate sites for speech, NL and terminology

- Distributed YES but separate for Speech NL or Term. will be decided by circumstances ... who is willing to accept .. and so on. No 'theoretical position'
- common centre if possible for the 3 categories
- collaborating sites for speech, NL and terminology
- The production can take place anywhere. Final validation should be centralized. Separate sites might be more efficient and realistic for the time being.
- yes, but central national institutions for WLR and SLR and/or terminology should be possible (I guess you had this in mind too)
- The production and validation should be performed on a per-product basis. There is no need for a fixed number of sites/capabilities per country.
- yes
- Yes. Local, national, specialist knowledge is needed for validation. The general rules can be agreed and centralised but the actual work can be distributed.

6. The distribution should be organized by a (possibly european) association

- We saw in Garmisch that it is THE solution (the other propositions have showed their limits ...)
- yes
- or at the very least supported by it
- Yes
- I don't feel competent to answer this question; is an association some legal status according to EU legislation?
- A non-profit making company with an appropriate board of governors.
- ok
- Yes

7. Would 3.5 people in the executive management be enough/too much for the first year ?

- No comments
- enough
- too many to start
- Looks OK to me
- sufficient
- Too much. A better design would be one full-time central executive with part-time workers in each country. I + 8*0.1 say.
- no, it may be insufficient for all areas including character recognition
- It depends how much work has to be done to set everything up. I dont expect that there will be lot of real material to distribute/validate but the actual work will be in setting up the ground rules, administrative framework and infrastructure. My gut feeling is that it depends on the inputs they receive from Speechdat and Relator. If most of the preparatory work actually gets done then this effort will be too much, if not then it is about right.

8. Would 8/9 people be enough/too much for the following years ?

- If the executive manager is well identified .. no problem

- don't know. Will there be an advisory and an executive board ?
- too big
- This might be a bit too much.
- keep it small but powerful: 6 people seems more appropriate
- Too many - effort should be concentrated on contracts to make corpora.
- no
- Too much I think if the validation is performed by a supporting distributed infrastructure of local specialists. Much of the validation can be sub-contracted.

9. Should the executive management be new or based on an existing institute ?

- NO necessity to be NEW, but necessity to be efficient !!
- no strong feelings, although an existing structure would allow a faster set-up.
- existing in order to minimise delay
- The management structure should be set up for this institute specifically, however, it could largely overlap with an existing institute.
- ?
- new
- Based in an existing institution, parasitic for administrative support.
- Some of these people need to be full time while others can act as consultants. I think that it would be better if it were new but I can see practical difficulties! The most important issue is that it be clearly independent.

10. If it should be new, is it acceptable to have an existing institute as a start-up ?

- Probably required ...
- yes
- yes
- Definitely yes, SPEX might be a good candidate.
- yes, to get it going
- Yes
- yes
- yes

11. Is it realistic to consider 13 KEcus a reasonable amount for distributing an existing Database ?

- It seems to be difficult to decide such an amount without correlation with the database itself (volume, real state of data ...)
- don't understand what is "distributing". It sounds too expensive for just dissemination, but I haven't time to evaluate this now.
- no, but it is a useful budget estimate
- I don't understand this figure. It is too low if you mean collection costs. Pearce mentioned 130 kECU per language for 5000 speaker Polyphone-like DB, this roughly coincides with the basic costs of the Dutch Polyphone, although one can easily argue that the actual costs are much higher. 13 kECU is too high if mean reproduction costs on CDROM, or selling costs. The Italians ask for their Eurom-1 1600 ECU (800 for academic partners).
- I think the figure is ok.
- The pricing should be flexible and on a per-disc basis. 13kecu may be alright for the entire EUROM1 distribution, but not for one language.
- it really depends on the size! (you have my estimates)
- Is this the purchase price? If so, then this is about the upper limit for a database of the similar practical utility as the 5000 speaker polyphone type database per language.

12. Should the data be distributed on a yearly membership fee basis only, on a sale basis only, or hybrid ?

- I would prefer some hybrid arrangement
- hybrid
- by low cost sale, without restriction for European research
- Hybrid seems to be most realistic. There should of course be some benefit for being a member.
- hybrid: membership entitles to very low prices
- The LDC model has members with certain numbers of discs free per year plus special rates on other discs (Note from the editor: wrong. LDC went back to a yearly membership fee only mechanism, as this hybrid system proved to be too complicated). Non-members pay more per disc.

- both
- I prefer a sales only basis or a hybrid.

13. In the first case, is a yearly fee of 10KEcu (ie half of LDC), and 15 KEcu starting on 2nd year realistic for industry ?

- no Comments
- yes
- perhaps a bit too low
- These (LDC) figures seem to be realistic, since several European industries were willing to pay that to LDC. However, the European alternative may not have as much to offer as LDC in the start-up phase.
- ...
- Yes.
- why not the same as LDC. we need to get off the ground with a lot less public financing
- It depends what we would get for the investment. In our organisation, it is easier for us to justify a purchase for a specific purpose rather than a general ongoing membership. I can foresee it being difficult for us to join on a membership basis.

14. and 2 KEcu realistic for public research labs ?

- It could be too expensive
- yes
- yes
- This is a realistic figure, although for small labs like mine a yearly payment of 2 kECU may not be possible.
- yes
- Yes.
- yes
- Please note: the issue is not so much whether they are industrial or public research labs but whether the database is being used purely for research or sold on to be exploited in revenue generating business. Thus an industrial research labs may use a database for research only and never exploit the technology or the databases to create models - similarly a public research labs may receive funding by a company or PTT to develop technology for them which is then exploited in the market. We thus need to be careful about how the definitions (public and industrial) are used.

15. Is it realistic to think that at least 2 industrials would take a "supporting" 5 year membership ?

- ...
- yes
- more likely
- I hope so.
- ...
- Yes.
- I would hope so - if not industry is clearly not interested
- Not us!

16. And that 5 public research may do the same ?

- ...
- not sure
- yes
- There are indeed at least 5 big public resarch labs that most probably will subscribe. Think fi. of LIMSI, ICP, CNET, IPO, KTH, IDIAP, UCL, DRA, etc.
- yes
- Probably more: 10-20?
- ok
- ...

17. Is it realistic to think that 8 EU countries may give 50 to 100 KEcu (yearly ?) to support the effort ?

- I don't think !! (sorry !)
- no. At least I don't believe Portugal will be able to contribute this much :-)
- OK if they feel that they can influence policy

- National contribution are NOT realistic, however, contributions from national research programs or projects might be possible (such as Francil, Verbmobil, Dutch Priority program).

-...

- No. Countries will support national efforts, expecting the Commission to provide finances for European wide ventures. National finances might be obtained for national databases, which could then be distributed by the European centre. The fact that there is a mechanism for European exploitation might make obtaining national funds easier.

- probably not initially, but hopefully after several years

- No. I cant see the UK government contributing - there is no infrastructure for them to do so. They will probably say that the money has already gone into Europe to the 4th framework programme and that it should come from that.

18. Is it acceptable that all "profit" be reinvested in LRs ?

- It is required !!!

- sure, what else should it be reinvested in ?

- yes

- Yes, if this also includes improving the infrastructure of the centre itself.

- highly desirable

- Yes, after royalties have been paid to information providers.

- yes

- Probably yes but there does need to be an incentive to make a success of the venture.

19. What is your opinion on the possible roles of publishing companies ?

- With what we saw in garmisch I have strong concerns on the real benefit of such a solution (but the publicity will be very well done !)

- there may be some conflicts of interest if the center is supposed to distribute written language resources as well. This question should be answered by NL people preferaly.

- useful to consider

- I would love to see a joint venture between SPEX and Elseviers. I think several publishing companies are interested and could be of great help. However, they should not get exclusive property right.

- since they are profit oriented, they will have different interests than the SLR community; also, their know-how on SLR-specific topics is too limited.

- Need an enthusiastic company who could provide marketing expertise as well as a home for the centre. Maybe a better alternative would be to use a CD-ROM publishing company as a data management/mastering and distribution organisation; separating the commissioning into a central body.

- it could work for distribution, but no say in creation or needs analysis. not a top choice

- i) They could run the whole ECSLR, drawing on existing skills and building on them by recruiting the appropriate people (similar to the executive management above). To make it a suitable legal entity there could be a separate "association" (I'm not sure that this is the right legal term) which has a representative management board and sub-contracts the services of the publishing company. Alternatively with a lower degree of involvement: ii) They might be able to provide a low cost distribution and order handling mechanism with the other issues being handled by the association's executive management. The choice depends on the number of sales which will need handling (see the final point at the end of this response), the degree of responsibility placed on the publishing company and the logistics/practicality of starting it up. An important issue will be the individual people who are involved, their skills and commitment to making a success of the venture. This is probably more important than whether they are employed directly by the centre or indirectly by the publishing company.

20. Who should seat in the Association Management Board: elected members, + founding members (?) (who are they ?) + "supporting" members (?) + Executive manager ?

- We need elected members knowned by their scientific competences ... and we must accept representatives from founding members ... and of course the Executive manager.

- elected + executive manager would be my first choice but who does the election ????

- all can be considered BUT must have active people as priority and source of funding as political necessity

- If SpeechDat/Relator/CEU sets it up, then these partners should be represented (not each individual partner of course) with a proper balance between academia and industry. By including a representative from ESCA, and perhaps a few other professional organizations, one give the impresion that the whole user and research

community is represented. Also the CEU should have a chair, in order to force them to contribute and to take formal responsibility for their initiatives.

- elected members, executive manager

-

- o Commission representative
- o European Speech Community representative
- o Supporting Industrial members representatives.
- o Industrial member representative
- o Public research member representative
- o Chief Executive

-...

- ...

- Should be industrial.

21. Should there be SIGs (telephone, office systems, synthesis...)

- ??????

- may be not at the beginning

- not at first - to fragmenting

- This might be a solution for stimulating DB activities in these fields, although there may be other and better ways to do that, such as good LE or national projects.

- Time will tell

- There clearly needs to be good communications between members: via the Internet (WWW, mailinglist, FTP), and also meetings on particular topics. These interest groups will arise naturally I suspect if there is a need for them.

-...

- Yes - they could provide a good communication channel to establish needs and test ideas on

.

22. Do you have databases that you would like to distribute ?

- BDSONS, EUROM1, BDBRUIT from the french GDR-PRC

- soon (Eurom.1)

- yes

- SPEX does have several.

- we could (if legal issues are solved B-) contribute TED and PhonDat

- There are a number of data sets that we could make available: I think if the commissioning procedure and royalties are sorted out then many labs will see the advantages of offering data.

- yes

- There are some existing telephone databases and in-vehicle databases. They are only worth our while distributing them if the cost of preparing them for distribution (quality assurance and distribution standards) is less than the expected revenue from sales.

23. Are there DBs that you are aware of and that you would like to get ?

-...

- no. But that's because for the time being I'm only interested in my own language and these databases have not been created yet.

- yes

- I was the external reviewer of Polyglot and I know that they collected interesting speech material. The same holds for various other Esprit and LRE projects, such as Sundial, Spell, Onomastica, etc. These DBs should be made available.

-...

- Yes, even just English ones that I know are in the UK. At present however there is no one site I can approach for details of availability!

- yes

- Not many that already exist, although there may be some which I am not aware of. Perhaps some of the databases which were collected on previous European projects. The Onomastica dictionaries. Certainly the Polyphone type databases which will be collected. Some already exist e.g. Dutch, Italian

24. Other comments

- A final thought:

Perhaps the most important question which may shape our decisions concerning finance and distribution is that of the number of transactions which will actually take place e.g. how many organisations (split into exploiters and non exploiters of the data) will want to purchase which databases. Assuming a particular product portfolio, how many purchases will there be? (I suspect that the figures I guessed at and circulated at the Garmish meeting were somewhat high). This is really a sales forecast. It would probably be useful to get some real feedback on this from those who will be involved - assume a certain set of databases were available for purchase from the centre - who would buy what, at what price and what timescale. LDC may be able to provide some insight too.

Annex 3: Possible databases for distribution

The following databases can be identified as possible targets for edition and distribution in the first year:

Speech:

- Groningen corpus (from the Dutch SPEX project)
- BREF (Text extracted from the journal Le Monde read aloud)
- ESPRIT SAM Eurom-0 (digits + Phonetically rich sentences)
- ESPRIT SAM Eurom-1 (digits + Phonetically rich sentences)
- Esprit Polyglot (Multilingual (7 languages) speech IWR + CSR + synthesis)
- ESPRIT BRA ACCOR DBs (multi-channel recordings)
- MAPTASK (HCRC-Edinburgh)
- SCRIBE (British English in UK), PHONDAT and VERBMOBIL corpora (Germany), FRANCIL corpora (Francophonie)
- WSJCAM0 (British English Wall Street Journal)
- TED (Transnational English Databases) (English pronounced by speakers with different accents + related texts)
- Polyphone Databases (agreement to be reached with LDC & producers) (telephone recordings IWR + CSR)
- MLAP Speechdat Polyphone-Like Public Research Databases (Telephone recordings IWR + CSR)

Text:

- European Corpus Initiative (ECI)
- MLCC (Lexica)
- Multilingual Parallel Texts from the European Parliament
- Multilingual Parallel Texts from the CEU translation services
- LRE MULTEXT: Multilingual Encoded Text corpora

Other data can be thought of (from ESPRIT projects (ARS, Spell, Sundial), and from LRE and MLAP projects). There should be a strict policy within CEC projects, that LRs data which are project deliverables should be deposited at the Distribution Center, with distribution (public, partly public, private, exchangeable, distribution delayed...).