H. Zimmermann

Department of Information Science
University of the Saar
Saarbrücken, Germany

Towards Standardisation in European Language Technology

1. Introduction

There is no industrial development without standards. Therefore, the definition and the application of industrial standards in Language Industry are an urgent and indispensable need.

In this context, Language Industry is defined as the "*industry which produces and distributes specific language based software or applications with linguistic features*." (DANSIN p. 34, orig. French).

In the report of the preliminary consultations with industry and user organisations about Language and Technology (February 1992, further referred to as LT), Language Industry is defined as "*The set of enterprises or professionals that produce final or intermediate goods and processes intended to improve the efficiency of verbal communication especially in a multilanguage community and to decrease the cost of producing efficient texts*" (where "verbal" refers to both written and spoken language) (LT I, p. VII,2).

One main aim of this study is to go through the description and arguments of the reports, to order the different items, to give a first impression of the problems to be solved and the possible steps of realisation of this big task of qualified and systematic standardisation in this special area of software development. Especially the role is handled the CEC and the European Community can or has to play in this regard.

2. Aims and types of standardisation

The LT report gives a lot of arguments for standardisation which are **common to other software technology standardisation**:

Technical reasons:

- "to improve portability, usability, robustness and openness" (LT I; p. 11).

General reasons for improving quality:

- "to help developers to create competitive advantageous facilities and features" (LT I, p. 11).

Competition as advantage for users:

- "to allow the use of different tools" (LT I, p. 25).

Cost reduction:

- "incompatibility leads to the loss of investment especially for SME" (LT I, p. 1.15).

Better understanding of functions:

- "help end users to compare services and products" (LT I, p. 11).

Quality Assessment:

- "creation of non-subjective quality assessment methods and criteria" (LT I, p.11).

Some of the arguments are oriented at **general problems in natural language processing**:

- "to enable natural language processing module development and integration including the provision of *common criteria on text structuring and writing*" (LT I, p. 16);
  "to support the *design of systems with multilingual capabilities*" (LT I, p.38);
- "definition and implementation of *interchange formats for documents and lexicographical or terminology databases*" (LT I, p. 25);

Others are directed to **special advantages**:

- "development and shared use of linguistic resources (written and spoken)" (LT I, p. 11);
- "*convergence of speech and text processing*" (LT I, p. 11);
- "guidance on the correct use of words (...) to ensure that the desired information imparted in different languages, especially when translating between languages" (LT I, p. 17);
- "development of standardised access methods to *linguistic profiles*" (LT I, p. 17);
- "address *document quality healths* and safety implications via legal authorities performing liable studies" (LT I, p. 25);
- "establish *improved cooperation between speech and natural language processing researchers and practitioners* by setting suitable standards" (LT I, p. 41; - this is also a general argument);
- "to *save time for translators* who often spend hours discussing terminology with subject field experts" (LT I, p. I.11).

By looking at these examples, two main directions can be identified:

(1) *Standardisation by formal criteria* (exchange, adaptability of language data and software tools);

(2) *Standardisation under content aspects* (quality, applicability).

3. Domains of standardisation in LT

The following main domains or areas (fields) of standardisation in LT can be differentiated:

3.1 Terminology:

Terminology is understood as standardisation of monolingual terms and of bi- or multilingual thesaurus development. Main arguments are:

- "adoption of standard terminologies in products, services and for their data interchange" (LT I, p. 11);
- terminology control (in general);
- "definition and implementation of interchange formats for (...) terminology databases" (LT I, p. 25).

3.2 **Character set, keyboards, form**ats:

There has been done a lot in standardisation of character sets. But the user, especially in multilingual environment, feels very unhappy with the existing results. Main arguments or suggestions are:

- "definition and implementation of an international character set, including Russian, Japanese and other languages for computer platforms" (LT I, p. 25);
- "development of a European standard for keyboards and character changes as well as keyboards with softkeys or key displays using flat screen technology. This would allow keys to be changed automatically according to the application program" (LT I, p.28);
- the problems of incompatible file and character formats should be solved primarily through integration;
- in relation to existing systems, there is a need for format conversion.

In any case, this will be a long haul, but it seems that a common language based European effort has to be done.

3.3 **Document production and exchange**:

The market standard (if one can speak about de facto standards) seems not to be "good enough" for the development and integration of linguistic features. Some of the arguments are to be mentioned here:

- "standards for integrated document production documents" (LT I, p. 25)
- "definition and implementation of interchange formats for documents" (LT I, p. 25)
- "definition of standard interfaces for integrated document creation, production, communications and library systems: define the required architecture" (LT I, p. 25)

The software industry seems to understand that the development of open systems will be an important precondition for additive functions. But LT still has to prove that the relevant functionality is available.

3.4 **Lexical data exchange, linguistic resources**:

In this field, LT is not starting from scratch. In the past, a lot of work has been done, especially by publishers. And there exists a lot of machine readable material. Therefore, there are proposals for standardisation as follows:

- "definition and implementation of interchange formats for documents and lexicographical or terminology databases" (LT I, p. 25);
- "there is a strong need for standards in the field of linguistic resources" (LT I, p. 1.12).

At first sight, this seems to be an easy task. In my opinion, the *pure technical solution* (exchange format) is not a real problem. The main difficulties are in contentual differentiation and structurisation. And it seems that practical work with existing material will reveal the gaps in these data.

3.5 **Textual (reference, representative) corp**ora:

Because of the lack of basic linguistic material (see above) one argues about standards on (described or tagged) language sources:

- "users require standards for tagged text and speech corpora, multipurpose dictionaries, terminology data banks and multilingual thesauri" (LT I, p.29);
- "text and speech corpora, language resources and archives require standards for efficient collection, exchange, sharing of diffusion and re-use" (LT I, p.29);
- "users need standards for tagged text and speech corpora (...). Integrated document processing is not possible without such standards (...), nobody will be interested in investing in large corpora without some kind of standards" (LT I, p. 1.12).

On the one side, this need is quite clear, but it also shows that Language Processing is far from being a well established Language Industry.

3.6 **Machine Translation**:

This area is of great importance. There is a potential market, and there are, in reality, available and partly usable products (SYSTRAN, LOGOS, METAL, GLOBALINK, to mention some of them). But the users are not happy with the existing systems, so that the following arguments in direction of standardisation occur:

- "assembly and updating of translation systems" (LT I, p. 22)
- "standardisation should address the subjectors of rough to fully automatic translation" (LT I, p. 31)
- "agreement on specific domain language structures to make them more amenable to machine translation" (LT I, p.34).

There is some feeling that MT will not reach, within a short time, a quality which is sufficient in the "linguistic sense". So testing instruments for comparing existing systems and rules for "good-enough" translations are imagined.

3.7 Simplified languages / natural sublanguages:

There is the opinion that one should concentrate, at least at the beginning, on standards for restricted languages:

- "definition of simplified languages (LT I, p. 1.9).
- "agreement on specific domain language structures to make them more amenable to machine translation" (LT I, p.34)

3.8 Others:

Some special applications are further cited:

- "formulation of standards on *language and culture independent icons*" (LT I, p.28)
- "formulate standards for multilingual teleservices, including quality assessment and metrics" (LT I, p.44).
- "message understanding" (LT I, p. 31)

4 Quality assessment:

Quality assessment can be seen as an application variant of LT standards. Without standards, no real comparison between suppliers can be done, no real assessment is possible. There are some suggestions for standardisation of quality assessment:

- "standard procedures to perform the evaluation process, including tools and techniques" (LT I, p.29)
- "standard results of benchmark tests" (LT I, p.29)
- "standards for databases associated with multilingual processing and knowledge transfer" (LT I, p. 35)
- "conversion routines" (LT I, p. 35) - which will allow the comparison of LT routines.
- "definition and description of reduced languages for specific industrial and administration domains" (LT I, p.37).
- "(standards, interfaces) so that multilinguality is included in the software architecture" (LT I, p.38).

5. Feasibility of standardisation in LT: General Problems

LT is no exception of the rule: Technical solutions need standards. But the question is: is standardisation in this application field feasible? With regard to the two main dimensions of standardisation, one has to give different answers:

- The feasibility of handling of natural language data by standardisation of exchange formats seems to be evident. If further solutions are a real need in terms of market interest, they will be done (e.g. in the field of character sets, keyboards, document structuring, terminology or dictionary data exchange formats). Within the development of computer equipment, especially with open system architectures and client-server concepts, there will be - more or less automatically - a push in this direction. But one has to be very careful with theoretical concepts in this direction: The example of the LCD keyboard (now available at Hohe Electronics) shows that there is a big difference between wishes (and technical possibilities) and the real market.

- On the other side, *with regard to content and quality*, there is nearly no approved or accepted standard in sight. There are two reasons for that: One is the *vagueness* of "linguistic data" and the *problem of formalisation* itself. The other is the *copyright problem*: Companies who spent a lot of time and money in creating and maintaining linguistic functions normally don't want to reveal their concepts and to give away the internal knowledge. Even if they know that they would need a lot of further investment of capital and manpower, there is a great resistance in opening their functions and processes.

6. State-of-the-Art

If one looks at existing standards, it seems that a lot of standards are available, but that they don't fulfil the aims of the LT. Some examples:

- *character sets* (several ISO and national standards)
- *lexicography* (exchange formats like MATER)
- *text encoding* (SGML)

In some domains, standardisation is under way, e.g. the Text Encoding Initiative (TEI), the standardisation in Speech Processing (SAM Esprit project). Further proposals will come from projects like GENELEX, AQUILEX etc.

On the other side, there is a gap between the *word processing systems frequently used worldwide* (like WordPerfect, MS-WORDS) and the integration of linguistic functions with higher quality or functionality. The reasons are very complex:

- Existing systems have (or at least: had) to follow the basic technical possibilities: availability of character set standards, the technical equipment of the "standard" user (sometimes still working with an inflexible primitive terminal) etc.

- The final user - up to now - is not very interested in "deeper" or more complex language processing.

- On the other hand, when oriented at English, functions like spell checking and automatic hyphenation (which nowadays are common in word processing and DTP) can be handled, to a sufficient degree, by simple tools (word lists, exception lists). The 'big" (US-) companies developing WP systems normally got specialised partners delivering such functions via a simple interface and efficient in speed and storage (we are speaking of the generation of XT personal computers and comparable solutions).

That is why the existing interfaces are not able to handle more complex (but still "simple") problems, like hyphenation in German or decomposition of nouns or upper/lower case.

- In the field of data base software (on mainframe, not to speak about CD-ROM solutions), there are the same problems. Most of the systems are "English" based, that means that problems of word inflection / deflection are handled by simple word form truncation and adjacency functions which are - to some degree - "good enough" for English, but not for French or German etc.
- On the other hand, the application of "higher quality systems" like GOLEM / PASSAT for German makes clear that development in this domain is very expensive, needs user activities and even will not be "perfect".

- In machine translation, systems like SYSTRAN, LOGOS or METAL show that one can "live", at least to some sufficient degree, with the existing interfaces, even if it is hard to update the interfaces when new versions appear. The SYSTRAN interface CONVERT (commercial SYSTRAN), for example, allows the direct application of MS-WORD, WordPerfect and other standard text processing systems: The systems METAL and LOGOS have similar interfaces to INTERLEAF resp. FRAMEMASTER or HIT. It is quite sure that with the development of more "neutral" interfaces like RTF (MS-Windows), the introduction of SGML and the client-server concept, such "text based processing" will get broader application, and the demand will grow for higher quality solutions.

- In special domains (like libraries), there are well structured data bases following standardised and agreed structures. Millions of linguistic data are available in machine-readable form, so at least in these domains; this is also true for material in bibliographic databases like CAS, CELEX (multilingual) and others.

7. Need and procedures for further development

Within the promotional programme on Language and Technology under development at CEC, the *creation and harmonisation of standards* will play a major part, one could even speak of a key role.

To avoid pure theoretic actions, some general insights have to be taken into consideration:

- *The basic push for extended standardisation must come from the industry itself, both from the developers of application and environmental tools* (like DTP), text processing, data base developers and distributors, the office system creators, the Telecom services, *and* from the *specialised deliverers of the linguistic software products* (ancillary industry). If possible, these activities shouldn't be limited by European partners, but be oriented worldwide.

- As one of the first activities, different *levels of interfaces* have to be standardised (word based, sentence based and text based) and to be combined with the basic functions (spell check, lemmatisation / indexing, style check, translation). A *platform-independent client-server model* (as a general interface tool) will be a "must", *standardised interfaces* will provide the framework developer with alternatives.

- The main effort must be spent on content and quality description and measuring of linguistic data and functions. It must be quite clear, for the customer / (final) user as well as for the competitor, what has to be understood by "hyphenation", "spell check", "style", "text-to-voice", "speech recognition", etc., where are the possibilities and where are the restrictions of a special function, an electronic dictionary etc.

- The final user has to be provided with standardised functions which allow the reuse of material created and modified by himself during previous applications.

Because it is quite clear that the existing linguistic features don't fulfil all the wishes - even basic needs - of the user, there must be a fundamental change of the general strategy in promotion in this field.

8. Role of the Commission, Infrastructure

The Commission is already involved both in the standardisation of exchange formats and the development (see above).

But the CEC is also "obliged to propose activities for coordination and harmonisation via standards" (DANZIN p. 23, orig. French).

There are a lot of suggestions for the role of the CEC in the promotion of standards:

- "Promote the definition and use of standard interfaces for the software modules to enable easy assembly and updating of translation systems" (LT I, p. 22);
- "The CEC should actively support the development of standards" in integrated document management (LT I, p. 25);
- "CEC should take the lead in an corporative endeavour to formulate standards supporting Community actions on speech and natural language technologies, services and products and providing mechanisms for the dissemination of standards and related information" (LT I, p. 31)
- "CEC should arrange agreement on specific domain language structures to make them more amenable to machine translation (LT I, p.34);
- "promote consensus on standards formulation" (LT I, p. 35)
- "CEC should participate in international standardisation bodies with the objective to add a linguistic dimension to ISO 9000 series" (LT I, p.37);
- "A pan-European panel should be set up (i.e. by the CEC) to identify needs, set common standards and present ideas as how to proceed" (LT I, p. 45);
- "Arrange, with standardisation organisations and terminology data bank developers, the creation and management of terminology banks for the Community (LT I, p.45);
- "This makes it even more important to provide a solid basis for the elaboration of standards for integrated document processing and for the Commission to back these standards all the way from proposals via industry standards formally adopted ISO or CEN/CENELEC documents" (LT I, p. 1.8).
- "Exploration of ongoing activities and distribution of results from AQUILEX, GENELEX, MULTILEX and Text Encoding Initiative" (LT I, p.1.13).
- "CEC support for propagation of standards" (LT I, p.1.13).

So it seems that there is a great confidence in the role of the CEC, within the planned programme, to take an active leadership in coordination of ongoing standardisation activities.

In my opinion, the CEC can take over the following important roles:

(i) The role as *coordinator and clearing house* by establishing, together with the relevant government agencies of the member states and with national language research institutes and national standards committees, an infrastructure.

This can be done by setting-up a network, combined with a European based clearing and information processing center specialised on Language Technology.

(ii) The CEC could be, as big international and multilingual institution, one of the "born" leaders in setting and applying "de facto standards" for *their own multilingual text and speech processing*. But, to compare this with former activities (like the development of the CEC-SYSTRAN machine translation system and the software for EURODICAUTOM), there should not be a new additional CEC internal development of tools: On the contrary, the strategy should be directed on defining and applying, later on, *quality levels and interfaces* in such a way that the (external, industrial) LT developers or suppliers will have guidelines and encouragement for their strategic developments.

(iii) The CEC should give *sufficient (including financial) support* on development where specialised suppliers of LT follow the guidelines by creating functions based on (preliminary) standards and interfaces.

But even if such a concept will be realised, the success will depend predominantly on the *insight of the LT industry partners* that *they* have to agree on their own standards and not to wait for external suggestions and "abstract" recommendations.

9. Outlook

Language Industry still suffers from undercapitalisation and an underdeveloped market. If LT functions are available, they are more or less ad hoc or "island solutions".

The only ways to overcome this situation, in my opinion, are:

- finding consensus and agreements in standards (even if this might, sometimes, lead to restrictions),

- more cooperative work, at least on the base of *strategic alliances of companies* (in different countries or with different functionalities).

The latter concepts will also help in solving the problem that solutions have to be developed for all the (European, including Easter European) languages, "ideally" with similar quality and functionality, even if the market will be dominated by the so-called "bigger" languages (English, French, Spanish, German, Russian).

Existing instruments (EUREKA, ESPRIT, IMPACT, to mention some of them) can be utilized, but they cannot solve the basic problems in Language Industry.

There is no doubt that the promotional programme for the development of Language Industry and Technology planned by the CEC has an important and even crucial mission and is worth to be supported by the government institutions of the Member States.

The main instruments of this programme:

- *creation of an effective and efficient infrastructure*
- *development; improvement and tentative / demonstrative application of LT standards*
- *financial support* as a subsidiary and substitutional solution up to a functioning market

will - together with an engaged Language Industry - lead to a *balanced model* that can fulfil the multilingual "needs" of the European user.

There is also no doubt that Language Technology has an important political function in a situation where the interest in an expanding European Community will depend increasingly on the solution of the problem of language barriers.

(T27EGST2) date: 15-6-92 Version: 1.2