Chapter 12

Language Resources

12.1 Overview

John J. Godfrey[a] **& Antonio Zampolli**[b]

[a] Texas Instruments Speech Research, Dallas, Texas, USA
[b] Istituto di Linguistica Computazionale, CNR, Pisa, Italy

The term *linguistic resources* refers to (usually large) sets of language dat and descriptions in machine readable form, to be used in building, improving or evaluating natural language (NL) and speech algorithms or systems. Exam ples of linguistic resources are written and spoken corpora, lexical databases grammars, and terminologies, although the term may be extended to includ basic software tools for the preparation, collection, management, or use of othe resources. This chapter deals mainly with corpora, lexicons, and terminologies

An increasing awareness of the potential economic and social impact of nat ural language and speech systems has attracted attention, and some suppor from national and international funding authorities. Their interest, naturally, in technology and systems that work, that make economic sense, and that dea with real language uses (whether scientifically *interesting* or not).

This interest has been reinforced by the success promised in meeting suc goals, by systems based on statistical modeling techniques such as hidden Marko models (HMM) and neural networks (NN), which learn by example, typicall from very large data sets organized in terms of many variables with many po sible values. A key technical factor in the demand for lexicons and corpora, i fact, is the enormous appetites of these techniques for structured data. Both i speech and in natural language, the relatively common occurrence of relativel uncommon events (triphones, vocabulary items), and the disruptive effect o even minor unmodeled events (channel or microphone differences, new vocabu lary items, etc.) means that, to provide enough examples for statistical method to work, the corpora must be numerous (at the very least one per domain o application), often massive, and consequently expensive.

The fact that we still lack adequate linguistic resources for the majority of our languages can be attributed to:

- The tendency, predominant in the '70s and the first half of the '80s, to test linguistic hypotheses with small amounts of (allegedly) critical data, rather than to study extensively the variety of linguistic phenomena occurring in communicative contexts;
- The high cost of creating linguistic resources.

These high costs require broadly-based cooperative efforts of companies, research institutions and sponsors, so as to avoid duplications and to widely share the burden involved. This obviously requires that linguistic resources not be restricted to one specific system, but that they be *reused*—by many users (*shareable* or *public* resources), or for more than one purpose (*multifunctional* resources). There are many examples of the former, such as the TIMIT corpus, TI-DIGITS, Treebank, the Celex Lexical Database, the Italian machine dictionary, and a few of the latter, such as SWITCHBOARD (used for speaker identification, topic detection, speech recognition, acoustic phonetic studies), the GENELEX dictionaries and the MLCC corpus.

A controversial problem, especially with natural language materials, is whether, in order to be reusable and multifunctional, linguistic resources must also be *theory-neutral*: the requirements for linguistic information of a given natural language or speech system may depend not only on the intended applications, but also on the specific linguistic theories on which the system's linguistic components are explicitly or implicitly based.

At the scientific and technical level, the solution is to attempt a consensus among different theoretical perspectives and systems design approaches. Where successful, this permits the adoption of common specifications and *de facto* standards in creating linguistic resources and ensures their harmonization at the international and multilingual level. The Text Encoding Initiative, jointly sponsored by ACH (Association for Computing in the Humanities), ALLC (Association of Literary and Linguistic Computing), and ACL (Association for Computational Linguistics), has produced a set of guidelines for encoding texts. The project LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards), recently launched by the CEC DGXIII, is pooling together the European efforts of both academic and industrial actors towards the creation of *de facto* consensual standards for corpora, lexicons, speech data, and for evaluation and formalisms.

At the organizational level, we can recognize, with regard to the present state of the art, the need for three major action lines:

(a) to promote the reuse of existing (partial) linguistic resources. This can imply various tasks, from reformatting or converting existing linguistic resources to common standards, to augmenting them to comply with common minimal specifications, to establishing appropriate

(b) to promote the development of new linguistic resources for those languages and domains where they do not exist yet, or only exist in a prototype stage, or exist but cannot be made available to the interested users; and

(c) to create cooperative infrastructure to collect, maintain, and disseminate linguistic resources on behalf of the research and development community.

The most appropriate way to organize these activities is still under discussion in various countries.

In Europe, the CEC DG-XIII LRE-RELATOR project, begun in 1995, aims at creating an experimental organization for the (c) tasks. The LE-MLAP (Language Engineering Multilingual Action Plan) has launched projects for activities of type (a) and (b) in the field of written and spoken corpora, lexicons, and terminology.

In Japan, plans for a central organization for speech and text databases have been under discussion. The EDR (Electronic Dictionary Research) Institute is, a the time of the writing of this volume, about to conclude the creation of large monolingual Japanese and English lexicons, together with bilingual links, a large *concept* dictionary and associated text corpora.

The approach taken in the U.S. was to create the Linguistic Data Consortium (LDC); although started with a government grant, it depends on membership dues and data collection contracts for its continued operations. LDC's principal mission is exactly (c) above, but in fulfilling the needs of its worldwide membership it addresses (a) and (b) as well. In its first three years it has released over 275 CD-ROMs of data for public use. Examples of its activities include:

- Publication of existing corpora previously available only to government contractors;
- Collection of speech and text data in languages of interest to members (English, Mandarin, Japanese, Spanish, French, and others);
- Creation of Common Lexical Databases for American English and other languages, with free commercial licenses for members;
- Acting as a clearinghouse for intellectual property rights to existing linguistic resources;
- Campaigning for the release of government-owned resources to researchers.

The need for ensuring international cooperation in the creation and dissemination of linguistic resources seems to us a direct consequence of their infrastructural role, precompetitive nature, and multilingual dimension. The CEC is taking a leading role for the coordination, among the EU countries and EU languages. COCOSDA (for speech) and LIRIC (for NL) are spontaneous initiatives of the R&D international community which aim at ensuring world-wide

coordination. Inside the framework of EAGLES and RELATOR, the possibility of defining a common policy for cooperation between the major sponsoring agencies (CEC, NSF, ARPA, MITI) is being explored.