Forschungsberichte des
Instituts für Phonetik und
Sprachliche Kommunikation
der Universität München
(FIPKM) 35 (1997) 139-143

EIGHT MAIN DIFFERENCES BETWEEN COLLECTIONS OF WRITTEN AND SPOKEN LANGUAGE DATA

Hans G.Tillmann

Institut für Phonetik und Sprachliche Kommunikation
Ludwig-Maximilians-Universität München
Schellingstr. 3
D-80799 Munich, Germany

Author's note

The following paper was written some time ago as a contribution to the EAGLES activity of the EU and has now been published as Section 2 of Chapter 3 in the *Handbook of Standards and Resources for Spoken Language Systems* (D. Gibbon, R. Moore and R. Winski, eds.). At the start of the project not many representatives of the commission seemed to understand the distinction between NLP = Natural Language Processing and SLP = Spoken Language Processing. Indeed, when we started - at a meeting with members of the EU in Luxembourg - to discuss the necessity of introducing a new fifth working group on spoken language into the already existing EAGLES consortium, Adrian Fourcin had to point out that the acronym *NLP* could as well be read as *Nonspoken Language Processing*. (We all know that exactly this fact caused Hirose Fujisaki to coin the term SLP !)

In its original version my contribution contained (and was entitled:) "*Seven main differences....*". The editors of the new EAGLES handbook decided to add a further difference into my list and introduced a new Section 7, concerning (aptly) "*The different legal status of written text and sampled speech signals*". I am not going to discuss the legal status of written and/or published pieces of texts with respect to PPRs, but simply feel free to introduce here as a postscript my own Section 7 (see PS7, below) which is to point out the multimedia nature of spoken language[1]. I do so because it seems to be quite clear to me that this new aspect is going to play a very important role in future collections of spoken language data.

In order to keep my list of differences in agreement with the official ("standard") numbering, I have changed my own original Section 7 into Section 8, and have inserted the new section entitled

"*PS7. The multimedia nature of spoken language*"

as the seventh in my own list of eight differences.

Introduction

Traditionally, linguists and natural language processing (NLP) researchers understood language corpora to consist of written material collected from text sources which already exist and often are available in published form (novels, stage and screen plays, newspapers, manuals, etc.). In this context the term "spoken language text corpora" was used to indicate

[1]The published Section 7 is cited here as an appendix. In addition, two or three minor changes to my original manuscript (probably introduced by Els den Os or Christoph Draxler during the editorial process) are left unflagged.

that the data are not taken from existing texts but that speech had to be written down in some orthographic or non-orthographic form in order to become part of a data collection. However, the differences (and relations) between text and speech data are far more complex. There are at least eight important differences, which must not be ignored because they determine relevant properties of the resulting data collections. For future (technological) developments of Spoken Language Processing (SLP) they should be taken into account very seriously.

These eight differences have to do with:

1. the durability of text as opposed to the volatility of speech,
2. the different time it takes to produce text and speech,
3. the different roles errors play in written and spoken language,
4. the differences in written and spoken words,
5. the different data structures of ASCII strings and sampled speech signals,
6. the two reasons that cause the great difference in the size of NL and SL data collections,

PS7. the multimedia nature of spoken language

8. the most fundamental distinction (as well as relation) between symbolically specified categories and physically measured time functions.

A closer look at these eight differences between written and spoken data will reveal why the traditional term "natural language processing", NLP, also could well be read as standing for "Non-spoken Language Processing". As it is our goal to call special attention to the relevant differences we will refer to the written language data as NL data meaning *non-spoken language data*, and set it in opposition to the term SL data, the acronym for *spoken language data*.

1. Durability of text, volatility of speech

The first distinction may seem rather trivial but it must nonetheless be mentioned, because it affects specific properties of the collected NL and SL data. While text generally stays on the paper when it is written down, speech is transient. It is the nature of the phonetic facts which speakers create during speech acts that they disappear at the moment they come into existence.

The first difference (which in the step from speaking to writing has helped our cultural development) explains why to collect SL data is less trivial than to produce NL data. The former must necessarily be recorded, for example on a tape or a disk, to make it accessible for future use.

2. Different production times for text and speech

Another difference between NL and SL corpora is due to the fact that speech data are time functions in a sense in which text data are not. Whilst a writer may consume any time he wants (or needs) to invest in producing a text, a speaker must code and transmit the phonetic information through syllabically and rhythmically organised sound transitions. Speech must run in its own natural time with a typical syllable rate of a value between 120/min and 180/min. The time for writing new text is normally much longer than it takes to read it aloud (which does not mean that silent reading, as well as short-hand-writing, cannot be much faster than speaking the text).

3. Correcting errors in the production of text and speech

In spontaneously spoken language the editing behaviour of the speaker is audible and remains a part of the recorded data. Interruptions, hesitations, repetitions of words (and parts of words), and especially self-repairs are a characteristic feature of naturally spoken language and must be represented in SL data collections of spontaneous speech. On the other hand, the writer who has even more correcting and editing options in producing a text document, will normally intend to produce a "clean" version. In the final version of the text all corrections which may have been carried out have disappeared; this is especially true for text intended to go into print. In the recent past SL data were often recorded as clean speech collections. A typical example is so-called laboratory speech which is produced when a speaker who is sitting in a monitored recording room reads a list of prepared text material, and then only the proper reproductions of the individual text items are accepted to enter the data base. Examples of speech corpora collected in this way are EUROM-0 and EUROM-1, as well as the early PHONDAT corpora of German. More recently, however, interest has shifted towards corpora comprising "real-world" speech, including hesitations, corrections, background noise, etc.. This is especially true of the German data collections for VERBMOBIL distributed by the BAS (cf. Schiel, this volume).

4. Orthographic identity and phonetic variability of lexicalised units

In correctly written texts any morphologically inflected lexical item generally has just one distinct orthographic form. Thus the words of European languages are easily identified and also well distinguished from each other, and there is usually only one version of each possible orthographic contextual form of any given word. The spoken versions of

orthographically identical word forms show a great phonetic variation in their segmental and prosodic realisation. In most European languages the phonetic form of a given word is in fact extremely variable depending on the context and other well defined intervening variables such as speaking style and context of situation, strong and weak Lombard effects (the influence of the physical environment on speech production via acoustic feedback), etc. A given word can totally disappear phonetically, or can be reduced to - and only signalled by - some reflection of segmental features in the prosody of the utterance. Most of these inconspicuous variations appear only in a narrow phonetic transcription of a given pronunciation.

It makes a great difference whether a word has been uttered in isolation or in continuous speech . Only if a word is consciously and very carefully produced in isolation can we observe the explicit version of its segmental structure . These phonetically explicit forms produced in a careful speaking style are called citation forms or canonical forms. The segmental structure of so-called citation forms is modified as soon as it is integrated into connected speech (probably systematically, although relatively little of the system is currently understood). For the design of spoken language corpora this is very relevant. It has also been taken into account in the conventions of the IPA proposed for Computer Representation of Individual Languages (CRIL, see Appendix A in the *Handbook of Standards and Resources for Spoken Language Systems*).

In dealing with SL data one must be able to know which words the speaker intended to express in a given utterance. This is reflected in the CRIL convention of the IPA. Here it should be mentioned that an SL data collection should ideally have at least two and possibly three different symbolically specified levels which are related to the acoustic speech signal:

1. On the first level the words of the given utterance are identified as lexical units in their orthographic form.
2. On the second level a broad phonetic transcription of the citation form should be given (which may be the result of automatic grapheme-to-phoneme conversion, as for very large SL corpora it would cost too much time and too much money to make broad phonetic transcriptions manually). If a reliable pronunciation dictionary is available the canonical representation of orthographically given words (cf. first level above) can easily be looked up.
3. How the given words have been actually pronounced in a given speech signal must be specified in terms of a narrower phonetic transcription of each individual utterance on a third, optional CRIL-level. This third level can then be directly aligned to the segments or acoustic features of the digital speech signal in the data base, which can be done automatically or manually. This information is especially relevant if also multi-sensor data are to be incorporated in SL databases.

Detailed phonetic transcriptions are subject to intra and inter-transcriber variability. Furthermore, they are extremely expensive, to the extent that they are likely to be prohibitive for large corpora. However, recent attempts using large vocabulary speech recognisers for the acoustic decoding of speech show some promise that the process can be automated, at least to the extent that pronunciation variation can be predicted by means of general phonological and phonetic rules. The Munich MAUS system has been especially helpful in processing the spontaneous speech material of the VERBMOBIL project (Schiel, this volume).

In addition to phonetic detail on the segmental level, several uses of spoken language corpora may also require prosodic annotation. In this area much work remains to be done to develop commonly agreed annotation systems. Once such systems exist, one may attempt to support annotation by means of automatic recognition procedures.

5. Printable ASCII-strings and continuously sampled speech

Taken as pure data, written texts and spoken utterances are completely different. In all European languages written NL data consist of strings of printable alphanumerical and other elements coded in 7- or 8-bit ASCII-Bytes. The resulting NL strings possess already a characteristic information structure which is not available in the case of primary SL data. Separated by blanks, punctuation marks or control codes, ASCII-strings are grouped into lexical substrings; also, the explicit punctuation of phrases and sentences is an important property of NL data. None of this type of information can be found in the recordings of primary SL data, since in natural speech there are no ASCII elements representing word boundaries, full stops, commas, colons, quotation, question, exclamation marks. Recorded SL data are primarily nothing but digitised time functions, oscillations of values in a sequence of numbers.

6. Size differences between NL and SL data

Comparing the pure size of stored NL and SL data reveals a great quantitative difference. There are two reasons why SL data require orders of magnitude more storage space than written language corpora. The first one is simply the difference in coding between text and speech. Whereas the ASCII string of a word like *and* needs only three bytes, many

more bytes are required as soon as the phonemes of this word are transformed into an acoustic output for storing the AD-converted data. If in the given example we assume that in clear speech the utterance of a three-phoneme-syllable takes about half a second and if we apply an amplitude quantisation of 16 bits and a non-stereo hi-fi sampling rate of 48 kHz, the NL/SL ratio amounts to approximately 1:16000.

The second reason follows from the great variability in the phonetic forms of spoken words. As pointed out above, any written text must be reproduced by many speakers in more than one speaking style (at least at slow, normal and fast speeds with low, normal, high voice, etc.), if the corpus is intended to reflect some common sources of variability.

PS7. Multimedia dimensions of future SLP

There is a third reason which will cause a further very dramatic expansion in the sizes of SL data collections as opposed to NLP data collections, as well as in the resources required for processing these data. Whereas the multi-media aspects of written data can be reduced to the ascii-string of a given text and to the form and appearance of its graphical representation (possibly specified in HTML), spoken language is always of a totally multi-media nature. Not only can the acoustic time function of any natural speech signal be directly related to the articulatory movements of speech production (which introduce large amounts of additional multi-sensor data such as glottograms, EMG-data, recordings of electromagnetic articulography, etc.), but, equally, the visible speech movements observable in the face of the speaker as well as the "prosodic movements" of the whole body of a person (acting in the situational context of a recorded speech act) lead to very large data collections.

This new type of multi-media speech data can now not only be properly collected, but can also be effectively dealt with since such large amounts of data can be stored and processed by means of newly available modern database management systems. Multi-media speech data can thus be effectivly used to further study human speech in all relevant details as well as to develop new and better applications of SLP-technologies, especially for man-machine-communication.

8. The different nature of categories and time functions

The last difference, and the most important one, must be looked at from two different angles. The first thing to understand is that the relevant category of the data (that determines its collection) is already inherently given in the case of NL, but totally unknown in the case of physically recorded speech. The ASCII symbols of a given text are elementary categories by themselves, and are directly used to form syntactically analysable expressions for the representation of all the different linguistically relevant categories. Thus relevant categorical information can be directly inferred from categorically given data and their ASCII representations. In contrast to this NL situation, the data of a digital speech signal do not signal any such categories, because they only represent a measured time function without any inherent categorical interpretation. At the present stage in the development of SLP it is not yet even possible to decide automatically whether a given digital signal is a speech signal or not. Therefore the necessary categorical annotations for SL data must still be produced by human workers (with the increasing support of semi-automatic procedures).

The second matter that must be considered in judging the different roles of categories and time functions in speech technology is that speech signals contain relevant prosodic and paralinguistic information that is not represented by the pure text of what was pronounced within a given utterance. As long as NLP can be restricted to non-spoken language processing the restriction to NL data does not pose severe problems. But as soon as real speech utterances are to be processed in an information technology application, the other, non-linguistic, but communicatively extremely relevant categories cannot be ignored. They must be represented in future SL data collections, and much effort has still to be invested by the international scientific community to deal with all these information-bearing aspects of any given speech utterance.

APPENDIX

(Citation of the published Section 7)

"7. The different legal status of written texts and spoken words

With few exceptions, the texts in NL corpora have previously been published. From a legal point of view, this implies that any use of electronic copies should adhere to copyright rules and regulations. In most countries copyright laws were passed long before the era of electronic publishing. However, laws designed to protect printed materials may not be optimal for the protection of machine readable text. Neither is it obvious how abuse of electronic texts can be detected and prevented. These problems have impeded the distribution of NL corpora quite considerably and it would be optimistic to suggest that all problems are close to a solution.

For SL corpora the legal issues are even less well understood. Has a speaker who is recorded while reading sentences presented by an experimenter any legal rights with respect to the sounds produced? Recordings of spontaneous speech are even more complex in this respect, since a speaker might claim rights as to the contexts and details of the formulations used. If speakers are recruited to contribute to a

SL corpus, legal problems can be avoided by requesting them to sign a consent form. Building corpora from existing recordings (e.g. from radio and television broadcasts) is more difficult in this respect, because it may not always be feasible to contact all relevant speakers. Under the law of EU countries unauthorised re-broadcast of recordings made from radio or television is illegal. It is less clear what the legal status is of limited redistribution of recordings for research and development in speech science and technology. For more information on this topic, we refer to Section 4.3.4." (op.cit., p.85)