Published in: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000). Athen, Griechenland. - European Language Resources Association (ELRA), 2000. 5 pp.

Enhancing speech corpus resources with multiple lexical tag layers

Andreas Witt, Harald Lüngen, Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
{gibbon|luengen}@spectrum.uni-bielefeld.de
witt@lili.uni-bielefeld.de

Postfach 10 01 31
33501 Bielefeld, Germany

Abstract

We describe a general two-stage procedure for re-using a custom corpus for spoken language system development involving a transformation from character-based markup to XML, and DSSSL stylesheet-driven XML markup enhancement with multiple lexical tag trees. The procedure was used to generate a fully tagged corpus; alternatively with greater economy of computing resources, it can be employed as a parametrised 'tagging on demand' filter. The implementation will shortly be released as a public resource together with the corpus (German spoken dialogue, about 500k word form tokens) and lexicon (about 75k word form types).

1. Introduction

The present project is the result of a co-operation between the VERBMOBIL lexicon data subproject and the text technology working group at the University of Bielefeld. VERBMOBIL is a large joint project dealing with the machine translation of spoken language, where 1128 mono- and multilingual dialogues in the domains of appointment scheduling, travel planning and hotel booking have been collected since 1993 (Jekat et al., 1997). The corpus consists of digitised audio recordings and their transcriptions (marked as 'TRLs' in Figure 1), and is used for the training of acoustic and language models for speech recognition, for grammar and domain modelling, and for training and evaluating a statistical translation system. The languages covered are English, German, and Japanese. At the time of writing, the German subcorpus used for the work reported here contained 489,722 word form tokens, and 7,926 different word form types, a type-token ratio of 0.016 with a high level of saturation for the domain.

From the transcriptions, a lexicon for speech related lexical information below the word level is derived in various stages of automatic and semi-automatic processing (Lüngen et al., 1998; Lüngen and Sporleder, 1999). For the present project, we have additionally converted the transcriptions to an XML representation, and the lexicon to a DSSSL list representation. We have then employed the DSSSL engine OPENJADE to combine these into an enhanced, reusable transcription in XML format. The data flow is depicted in Figure 1.

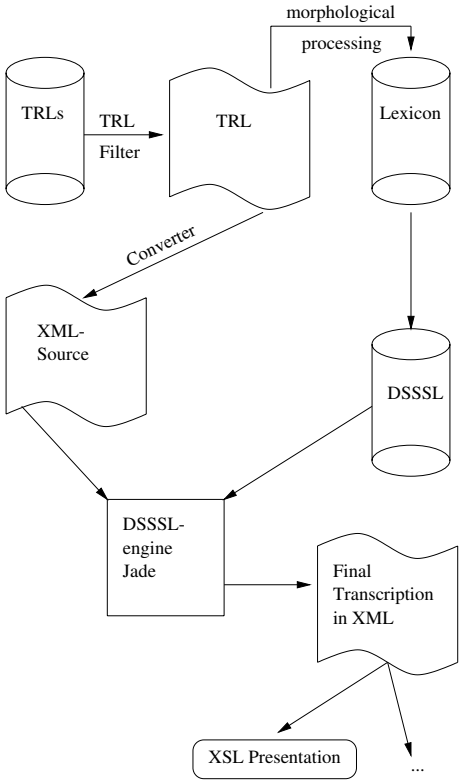

Figure 1: Architecture

2. Data

2.1. Transcription source format

The dialogues are transcribed orthographically with project-specific character-based markup conventions (TRL = 'transliteration', (Kohler et al., 1994; Burger, 1997)), devised for both manual and machine processing. Based on orthographic transcription on the word form level, the markup also encompasses spontaneous speech phenomena such as interruptions, corrections, repetitions, reductions, hesitations, speaker overlaps, and human noise as well as background and technical noise types. The example in Figure 2 illustrates the markup type:

On the one hand, the TRL markup, like SGML/XML, provides marking elements in angle brackets such as <*P*>; on the other hand, some markers are diacritic characters added to elements or word form orthography strings (such as #, or =). Figure 2 is taken from a multilingual dialogue with German and Japanese speakers. The same extract is shown in our XML-markup in Figure 5 and in the enhanced XML-markup in Figure 6.

```
m852arr1_032_HAB_000002:  <*tGER>
mhm .  <#Rustle> <uhm> ich habe
hier Hotelpl"ane .  <#Rustle> <uhm>
<#Rustle> <P> ja , <#Rustle> <P>
<#Rustle> <uhm> die wie w"ar' 's
Ihnen denn lieber , sollen wir in
der Stadt <uhm> ein Hotel nehmen
oder , <P> <uhm> ja , vielleicht
eher etwas au"serhalb .  oh , ich
seh' , hier ist gar kein Hotel
au"serhalb .  die liegen alle
ziemlich in der Innenstadt , also
da"s sie auch verkehrsg"unstig
gelegen alles , so da"s wir nicht
so weit zur Firma m"ussen .  <uhm>
welche Preiskategorie <uhm> ste=
<uh> stellen Sie sich denn so vor
?  <P> <#Rustle> eher etwas im
mittleren Bereich , h"oherer Bereich
oder vielleicht eher etwas billiger
?   m852arr1_033_HBD_000002:  <*tJAP>
<eto> watashi wa temoto ni hoteru no
puraN ...
```

Figure 2: TRL markup

2.2. Lexicon source format

For lexicon acquisition, these transliterations are automatically error-checked, filtered, and the orthographic word forms are morphologically processed (Gibbon and Steinbrecher, 1995; Lüngen et al., 1998). In this way, a morphologically structured background lexicon is semi-automatically acquired, and subsequently, the projected vocabulary of all morphological lemmata contained in the background lexicon is automatically generated and stored in an ASCII database (Bleiching et al., 1996). It is distributed project-internally both as a file and via an HTML query form and CGI access on the Internet. Currently, the projected lexical database contains 73578 records, and contains amongst other things the following types of information about each word form: Internal morphological boundaries, sequences of morpheme types, canonical phonological transcription (in SAMPA notation), syllable boundaries, lexical stress marks, morphological lemma, orthographic stem, phonological stem, possible morphosyntactic categories, and name class. The intensional coverage for morphological and phonological information is illustrated in the following sample query output.

Orth (ASCII-string) is the orthography according to the Verbmobil-II transcription conventions, it is the key to the word form token in the transliterations.

Phon (SAMPA-string) is the canonical phoneme transcription after the conventions in Gibbon, 1995).

OrthSeg is the regular Verbmobil-II orthography as under the **Orth**-attribute extended by morphological boundary information:

```
Entry 10546 matches String key
verkehrsg"unstig:
 Orth:         verkehrsg"unstig
 Phon:         f6ke:6sgYnstIC
 OrthSeg:      ver+kehr#+s#g"unst+ig
 PhonSeg:      f6.+k'e:6#+s#g''Yns.t+IC
 OrthStem:     ver+kehr#+s#g"unst+ig
 PhonStem:     f6.+k'e:6#+s#g''Yns.t+Ig
 MorLemma:     verkehrsg"unstig
 UnkTag:       *nil*
 CorpusFreq:   5
 PhonSep:      f-6-k-e:6-s-g-Y-n-s-t-I-C
 MorphCats:    root#+lm#root+suf
```

Figure 3: Lexicon database Query output

- # compound boundary
- + derivational or enclitic boundary
- #+ inflectional boundary
- - compound boundary (instead of # when also used in ORTH).

PhonSeg (SAMPA-string with morphoprosodic markers) is the pure phoneme transcription as under the **Phon**-attribute extended by:

- # compound boundary (# implies a syllable boundary except in a sequence `.C#` where `C` is a consonantal phoneme (example: *einander*: `?aI.n#'an.d6`)
- + derivational or enclitic boundary
- #+ inflectional boundary
- ' (preceding a vowel) primary stress
- " (preceding a vowel) secondary or tertiary stress
- . syllable boundary (when not collapsing with #).

MorLemma morphological lemma (orthographic citation form)

UnkTag Unknown word class (Schaaf and Dorna, 1998).

MorphCats is a sequence of morpheme categories that corresponds to the sequence of morphemes in the **OrthSeg** and **PhonSeg** representation. The inventory of morpheme categories is:

root_n	native root
root_nn	non-native root
pre1	prefix class I (native)
pre2	prefix class II (native)
pre_nn	non-native prefix
part	verbal particle
infin	infinitive marker *zu*
suf_n	native suffix
suf_nn	nonnative suffix
interfix	stem-extending interfix
lm	linking morpheme (*Fugenelement*)
infl_a	adjectival inflectional suffix
infl_n	nominal inflectional suffix
infl_v	verbal inflectional suffix
infl_pro	pronominal inflectional suffix

The lexicon format is a relational database format where much of the morphological tagging is represented using di-

acritics (boundary symbols, stress symbols) in strings that refer to the word form level. When linguistically interpreting these strings and diacritics, we can see that at least three different word form constituency levels are treated: the phoneme level, the syllable level, and the morpheme level. For each of these constituency levels in the objects described, there are different tagging layers, e.g. lexical morphemes have tags for orthographic representation, phonological representation, morpheme category, and nativeness.

```
(verkehrsg&uuml;nstig
  (("f" "6" "k" "e:6" "s" "g" "Y" "n" "s"
     "t" "I" "C"))
  (("ver+kehr#+s#g&uuml;nst+ig"))
  (("f6.+k'e:6#+s#g''Yns.t+IC"))
  (("ver+kehr#+s#g&uuml;nst+ig"))
  (("f6.+k'e:6#+s#g''Yns.t+Ig"))
  (("ver" "kehr" "s" "g&uuml;nst" "ig"))
  (("f6" "ke:6" "s" "gYnst" "IC"))
  (("pre1" "root" "lm" "root" "suf"))
  (("native" "native" #f "native"
    "native"))
  ("A")
  (#f)
  ("verkehrsg&uuml;nstig")
)
```

Figure 4: LEXICONTREE (pre-processed DSSSL compatible bracketed tree format)

For the present project, the entries of this lexicon data base have been converted into DSSSL association lists, cf. Figure 4.

3. TEI-based XML-version of the transcriptions

The first stage of the enhancement procedure is the automatic transformation of the corpus transcriptions into an intermediate corpus in XML format. XML offers a lot of advantages: Firstly, whereas the TRL conventions were newly developed for the VERBMOBIL project and a parser for format checking had to be implemented accordingly, an XML document type definition maybe either chosen or newly defined, and a steadily increasing number of standard SGML or XML tools for format checking is already availabe. Secondly, SGML/XML can be *edited* using standard software like e.g. Adept Editor, XMetaL, or the free PSGML mode for the Emacs editor. Thirdly, XML is supported by current WWW browsers, enabling straightforward dissemination procedures, as in our presentations in Figures 7 and 8. Moreover, the application of an already existant transcription scheme (whether XML-based or not) avoids retooling of development environments different local notations.

For our XML-markup, we have observed the TEI conventions for the transcription of speech (Sperberg-McQueen and Burnard, 1994), but use a simpler, more domain specific XML-DTD, which means that the annotated documents are both compliant to the TEI-DTD, which is not XML-based, and to our own, XML-compliant DTD cf. (Witt, 1998). Following the TEI-guidelines, we extended the TEI-markup scheme in those instances where the Verbmobil transcription convention is more detailed, i.e. we defined new tags and/or attributes to preserve all the information that is present in the original (Witt et al., 1997).

```
...
<u who="HAB" n="032" lang="GER">
mhm .  ich habe hier Hotelpl&auml;ne
.  ja , die wie <reg orig =
"w&auml;r'"> w&auml;re </reg>
<reg orig = "'s"> es </reg>
Ihnen denn lieber , sollen wir
in der Stadt ein Hotel nehmen
oder , ja , vielleicht eher etwas
au&szlig;erhalb .  oh , ich <reg
orig = "seh'"> sehe </reg> , hier
ist gar kein Hotel au&szlig;erhalb
.  die liegen alle ziemlich in der
Innenstadt , also da&szlig; sie
auch verkehrsg&uuml;nstig gelegen
alles , so da&szlig; wir nicht
so weit zur Firma m&uuml;ssen
.  welche Preiskategorie <del
type="truncation"> ste </del>
stellen Sie sich denn so vor ?
eher etwas im mittleren Bereich ,
h&ouml;herer Bereich oder vielleicht
eher etwas billiger ?
</u>
<u who="HBD" n="033" lang="JAP">
watashi wa temoto ni hoteru no puraN
...
```

Figure 5: CORPUSTREE_1 (XML markup)

In Table 1, some of the correspondences we established between the VERBMOBIL transcription markup (Burger, 1997) and the TEI tag set for the transcription of speech are shown. A new element we had to introduce is '<neol>' (neologism). Presently, we first filter out the non-lexical information tier in the source transcription that deals with overlaps of noise and speakers. Our `vm2xml` conversion tool is a perl script encoding a cascade of finite-state transducers. The original dialogue extract in Figure 2 is presented in XML markup in Figure 5.

4. Transforming XML to XML using DSSSL

The second stage of enhancement is multi-layer tree tagging of the intermediate corpus, expressed as a tree transformation function

$$T : (CORPUSTREE_1, LEXICONTREE) \rightarrow CORPUSTREE_2$$

The implementation of this function is a novel exploitation of a further feature of SGML/XML-annotated transcriptions. In most current applications, a highly structured document is *filtered* to produce a new document containing less information. In this project, however, we used a highly structured document, and a further knowledge base (our

Tag Category	Example TRL	TEI-XML Encoding
Lexical	<*ENG>Miss	<foreign lang="ENG"> Miss </foreign>
Lexical	Hannover	<name> Hannover </name>
Lexical	*Treffi	<neol> Treffi </neol>
Lexical	#zehn	<number> zehn </number>
Technical gap	<T_>tz	<gap reason="empty_signal" pos="word-initial"> tz </gap>
Unclear	wir%	<unclear> wir </unclear>
Human	<B>	<vocal desc="breath">
Human	<Laugh>	<vocal desc="laugh">
Non-human	<#Rustle>	<event desc="rustle">
Variant	'ner	<reg orig="'ner"> einer </reg>
Morph. coordination	be--	<m mcat="nonheadcoord"> be </m>
Truncation	heu=	<del type="truncation"> heu </del>

Table 1: Example of corresponding Verbmobil and TEI speech annotations

lexicon) to generate an even more highly structured document, containing more information than the original document. For doing this, we employed the ISO-Standard Document Style Semantic and Specification Language (DSSSL, ISO 10179:1996). Amongst other things, DSSSL defines a style language for the visual appearance of documents, and a tree transformation component for the transformation of SGML-documents. Up to now, these two functions have mainly been exploited for the purpose of *presenting* SGML-documents (e.g. as a presentation language for printout, or as a language for transformation to HTML), but DSSSL allows for more sophisticated kinds of processing. It is for example possible to incorporate a parser for natural language into text processing using DSSSL, providing automatic syntactic annotation of (maybe previously untagged) text (Witt, 1999).

Thus, the arguments of the transformation function are firstly the XML transcriptions (Figure 5) according to TEI conventions and secondly the lexicon transformed into DSSSL-conform notation (Figure 4).

```
...
</word>
<word sampa="f6.+k'e:6#+s#g''Yns.t+IC"
      cat="A"
      morph-readings="1">
<m sampa="f6" morph-cat="">ver</m>
<m sampa="ke:6"
   morph-cat="">kehr</m>
<m sampa="s" morph-cat="">s</m>
<m sampa="gYnst"
   morph-cat="">günst</m>
<m sampa="IC" morph-cat="">ig</m>
</word>
<word sampa="g@.+l'e:.g+@n
      cat="A"
...
```

Figure 6: CORPUSTREE$_2$ (enhanced XML markup)

The transformation program runs with the DSSSL engine OPENJADE, and converts the source XML file into a new XML file augmented by our morpho-lexical information, producing the output in Figure 6.

5. Presentation

The resulting XML document may be used in different applications. The most obvious one is to provide interested researchers with all the information available about a corpus in one document i.e. the transcriptions and the lexical information whenever it is needed. For this reason, we have developed a presentation using standard software (e.g. Internet Explorer 5), which allows the user to specify which components of the structured document should be displayed. Figure 7 shows this standard presentation mode.

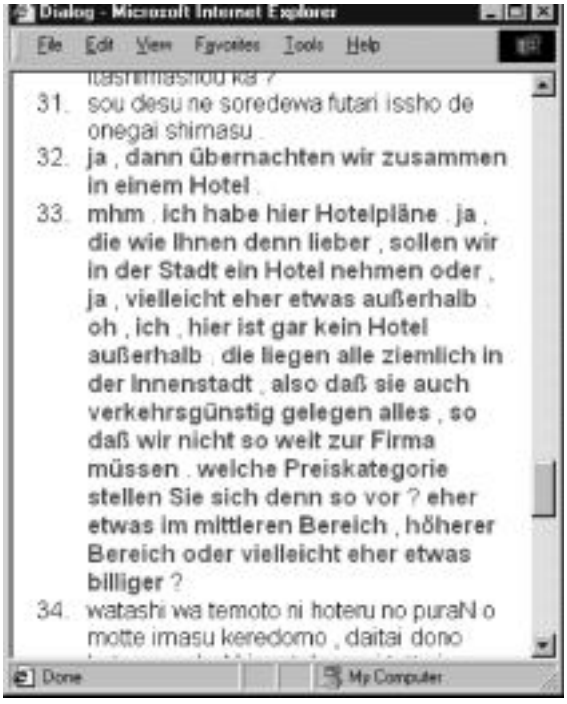

Figure 7: Standard presentation in Internet Explorer 5

If the user clicks on a word in the dialogue, a new window pops up showing the respective morphological decombination information (Figure 8).

Figure 8: Pop-up window showing morpholexical information

The presentation has been rendered using XSL (Extensible Stylesheet Language). XSL style sheets are presentable using standard WWW browsers such as Netscape or Internet Explorer. As an alternative, the DSSSL transformation component can be used to convert the XML source to HTML, but then the advantages of XML we mentioned above are lost. Moreover, no standard software for the presentation of style sheets designed with the DSSSL presentation component exists yet.

6. Results and prospects

We have described a procedure for automatically generating a multi-layer XML treebank from a conventionally annotated corpus, a relational lexical database, and a corpus which was produced using this procedure. It will shortly be made publicly available. The procedure is designed to enable re-use of existing corpora and lexica with new corpora and lexica in multimodal system development.

7. References

Bleiching, Doris, Guido Drexel, and Dafydd Gibbon, 1996. Ein Synkretismusmodell für die deutsche Morphologie. In Dafydd Gibbon (ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996*. Berlin: Mouton de Gruyter.

Burger, Susanne, 1997. Transliterationen spontansprachlicher Daten - Lexikon der Transliterationskonventionen - Verbmobil II. Verbmobil Technisches Dokument 56. LMU München.

Gibbon, Dafydd, 1995. Verbmobil Lexicon: Conventions for spelling and pronunciation. Verbmobil Technisches Dokument 31. Universität Bielefeld.

Gibbon, Dafydd and Daniela Steinbrecher, 1995. Verbmobil-Standardfilter für Transliterationen Version 2.2. Verbmobil Technisches Dokument 38. Universität Bielefeld.

Jekat, Susanne, Christian Scheer, and Tanja Schultz, 1997. VM II Szenario: Instruktionen für alle Sprachstellungen. Verbmobil Technisches Dokument 62. Universität Hamburg, LMU München, Universität Karlsruhe.

Kohler, Klaus, Gloria Lex, Matthias Pätzold, Michael Scheffers, Adrian Simpson, and Werner Thon, 1994. Handbuch zur Datenerhebung und Transliteration in TP14 von VERBMOBIL - 3-0. Verbmobil Technisches Dokument 11. Universität Kiel.

Lüngen, Harald, Karsten Ehlebracht, Dafydd Gibbon, and Ana Paula Quirino Simoes, 1998. Bielefelder Lexikon und Morphologie in Verbmobil Phase II. Verbmobil Report 233. Universität Bielefeld.

Lüngen, Harald and Caroline Sporleder, 1999. Automatic induction of lexical inheritance hierarchies. In Jost Gippert (ed.), *Multilinguale Corpora. Codierung, Strukturierung, Analyse*. Prague: Enigma Corporation.

Schaaf, Thomas and Michael Dorna, 1998. Behandlung unbekannter Wörter im Verbmobil System. Verbmobil Memo 132. Universtität Karlsruhe/Universität Stuttgart.

Sperberg-McQueen, C.M. and Lou Burnard, 1994. *Guidelines for electronic text encoding and interchange (TEI P3*. Chicago: Text Encoding Initiative. Volumes I and II.

Witt, Andreas, 1998. TEI-based XML-Applications: Transcriptions. In *Joint Conference of the ALLC and ACH (ALLCACH98)*. Debrecen. Http://www.arts.klte.hu/allcach98/abst/jegyzek.htm.

Witt, Andreas, 1999. DSSSL zur Verarbeitung linguistischer Korpora. In Jost Gippert (ed.), *Multilinguale Corpora. Codierung, Strukturierung, Analyse*. Prague: Enigma Corporation.

Witt, Andreas, Dafydd Gibbon, and Harald Lüngen, 1997. Standardisierung orthographischer Transkriptionen: Ein Vorschlag für Verbmobil. Verbmobil Memo 117. Universität Bielefeld.