Consistent Vocabularies for Spoken Language Machine Translation Systems

Dafydd GIBBON Harald LÜNGEN

1. Introduction

In this article, we pursue the task of defining consistent lexical coverage criteria for lexica to be constructed in a Spoken Language Machine Translation environment. We call the data structure defining the coverage of a module lexicon a *wordlist* and with the help of this explain the criteria of corpus consistency and translation equivalence for monolingual wordlists in an SL MT context. We then describe a practical application of the criteria, developed in the Verbmobil project, for which additional requirements stemming from the system architecture of the transfer based MT approach had to be taken into account.

2. Crucial Parameters in SL Applications

It is important to differentiate spoken language and written language corpora and lexica for these. In DEN OS (1997) eight main differences between collections of written and spoken language data are enumerated, some of which have immediate impact on lexical coverage criteria. Whereas many written corpora are simply to be found 'on the shelf', spoken language corpora are always designed for specific purposes and thus depend heavily on scenario specifications. In the case of Verbmobil, this is the appointment scheduling and travel planning task. For building lexica over a spoken language corpus, it is also necessary that the spoken data be transcribed according to transcription conventions that have to be carefully designed. Since a sufficiently narrow transcription will contain more than only lexical items, tools for filtering the transcribed corpora for lexical acquisition must be provided.

Differences lie also within the vocabularies as such. A certain category of words called discourse particles, for example, will be found ex-

clusively in spoken language; it is nevertheless appropriate to store them as lexical units, cf. FISCHER (1998).

3. Words and Wordlists

We define lexical coverage in terms of wordlists which contain words to be used as keys for lexical entries. The term word is frequently used in different senses when talking about the various module lexica (i.e. a speech recogniser dictionary vs. a syntax-semantics lexicon). The most common usages are:

1. *Word = Word form*: a phonological or orthographic representation of an inflected word; member of an inflectional paradigm. The items `sagen, sage, sagst, sagt, gesagt` count as different single words with this definition.
2. *Word = Morphological lemma*: a common stem representation of all elements in one inflectional paradigm; label of an inflectional paradigm.

The term *lemma*, in turn, may appear in other contexts, too:

1. *Semantic lemma*: language specific lexical meaning of at least one morphological lemma.
2. *Conceptual lemma*: language independent unit of meaning.
   a. Bilingual conceptual lemma: minimal conceptual lemma shared by one language pair.
   b. Mulitilingual conceptual lemma: minimal conceptual lemma shared by a language tuple of arbitrary size.

Figure 1 gives an example of how the different meanings of word and lemma are related. Bilingual conceptual lemmata correspond roughly to bilingual transfer rules in the translation task, whereas multilingual conceptual lemmata correspond to the language independent concept in the tip of a modified lexicon-oriented Vauquois-triangle (translation triangle), cf. Figure 2. In feature-based hierarchical lexica, such as ILEX (GIBBON

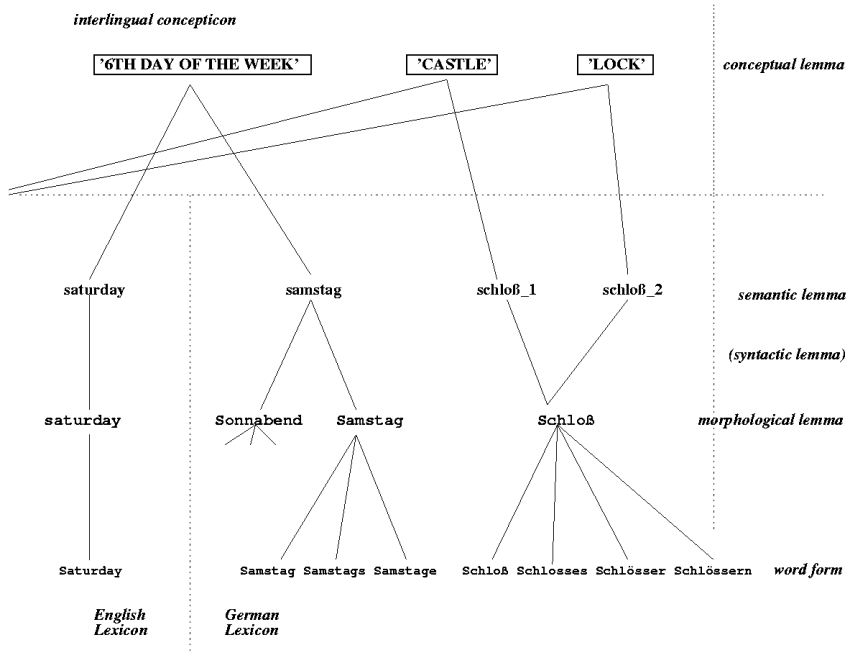

Figure 1: The relation between different usages of *word* and *lemma*

1991), or the HPSG lexicon (FLICKINGER 1987), the entries are modelled as default or typed feature structures encoded as attribute-value matrices (AVMs). The different levels of usage of the terms *word* and *lemma* evidently correspond to the use of different attribute-value specifications as dimensions to distinguish between types of signs. In other words, a certain level of subtypes in the type hierarchy is chosen to constitute a set of words, and another to constitute a set of *lemmata*. When a morphological mapping or relation between *word form* and *lemma* is explicitly defined, the term *word* can be used in many cases without needing to specify which of the two senses is meant. Within the Verbmobil project, wordlists are defined on the basis of word forms as they occur in the corpora, as these are the linguistic units employed in current word recognition systems. They also appear at the interface to the language modules, which is a word hypothesis graph (WHG). Thus, a wordlist also defines what is permissible as an arc label in a WHG.

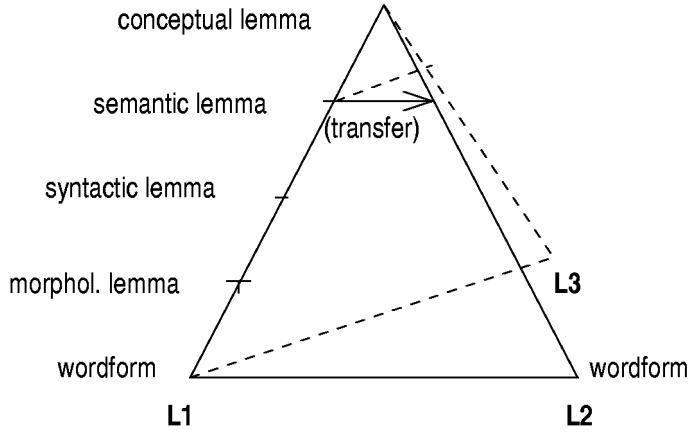

Figure 2: Extended translation triangle (Vauquois triangle)

4. Extraction of Wordlists from Corpora

4.1 Extensional Coverage Criteria

Extensional Coverage Criteria define the selection of lexical entries (forms, lemmata, or records). If a lexicon is to be built for a specific application, in most cases a desired vocabulary size M will be specified. In the case of Verbmobil Phase 2, this is 10000 word forms for German. We assume that a procedure such as the one described in GIBBON and STEINBRECHER (1995) is provided to obtain a monolingual word form list automatically from a transcribed corpus and thus operationally defines what a word form is. But for several reasons it is not desirable simply to define all the words to be found in the corpus as lexical since words with a very low token frequency may turn out as a.) transcription errors (such as `micht`), b.) ad-hoc word formations (such as `Diaabend-Weintrink-Revisionstreffen`), c.) words totally deviating from the given scenario and domain (`Safaribüchse`), all of which should not be stored in the module lexica, because it is quite certain that these will never occur again. Thus, it is desirable to include in the vocabulary only words of the corpus that appear more than *N* times, and to choose the largest *N* which permits

M to be reached. In the example in Table 1, at least all word forms occurring more than $N = 3$ times in the corpus of 708426 word form tokens have to be included in order to reach a vocabulary of 5000 word form types.

Frequency	Relative type/token ratio
N ≥ 1	9523 / 708426
N ≥ 2	5385 / 708426
N ≥ 3	4116 / 708426
N ≥ 4	3485 / 708426

Table 1: Frequencies of word forms and resulting vocabulary sizes in a corpus with 708426 word form tokens

4.2 Intensional Coverage Criteria

Intensional coverage criteria define the lexical properties of information types associated with lexical entries (i.e. features, values of attributes, or fields). If applied, they are independent of the corpus but, like the corpus, depend on the given scenario. They may be applied if it cannot be guaranteed that all the words that are known to be needed in the application actually occur in the corpus. If the corpus consists of spontaneously spoken language, as is the case in Verbmobil, this can certainly never be 100 % guaranteed.

Intensional coverage criteria are used to define addtional criteria for extensional coverage in the Verbmobil domain:

1. Include:
   - Control Commands: `lauter, leiser, wiederholen, ...`
   - Forms for inflectional paradigm extension, e.g. for nouns always include accusative singular form
   - Full set of function words: `für, mit, angesichts, ...`
   - Restricted set of cardinal numbers (for prices)

- Restricted set of ordinal numbers (for dates)
- Discourse particles: `äh, ähm, hm, puh, ...`
- Full set of time expressions: `Stunde, Montag, Januar, ...`
- Scenario-relevant adverbs: `heute, tagsüber, ...`
- Forms of address: `Herr, Frau, Doktor, ...`
- Spelling Vocabulary: `A, B, Berta, doppel, ...`

2. Exclude:

(a) Words that receive a class-based treatment

- Names:

`Czerczinsky`	→	UNK_Surname
`Parkhotel`	→	UNK_Hotel
`Mönckebergstraße`	→	UNK_Street

- Non-lexicalised Spelling Combinations:

`H-O-L-G-E-R, ...`

(b) Words that are unlikely to occur again

- Ad-hoc foreign language words: `cinema`[1]
- Scenario-external words: `Safaribüchse, Begehr`
- Neologisms: `vereinzubaren, Treffi`

5. Multilinguality and Translation Equivalence

Multilinguality in the Verbmobil context requires that the monolingual word list WL_A for language L_A must contain all the words that are needed to translate the words in the monolingual wordlist for language L_B . This means that in addition to the words obtained from the L_B -Corpora, the list WL_B must also contain the *translation equivalents* of WL_A . Of course, translation equivalents for a word cannot simply be given out of the blue. Each possible translation is context-dependent. The task may be feasible for a restricted domain, however. Note that these conditions are met in an SL MT project such as Verbmobil. We have a restricted domain of application and since the wordlists are corpus-based, the contexts in which each word of WL_A occurred are known. Thus, exact translations can

[1] These are sometimes accidently used by intepreters when they have difficulties with code-switching.

be given. There are in fact two sources, where translations can be obtained from: 1. the Transfer Rules 2. Aligned Translations of the data, produced by human translators or through automatic translation. We can thus define the 'translationally equivalent wordlist of a wordlist':

> The translation equivalent of a given Wordlist WL extracted from a dialogue corpus *C* is the list of words of the target language that are needed for the translation of *C*.

Note that the translation equivalent is defined for a corpus-derived word list, and deliberately not for a single word. In this way, translation equivalence can be defined context-dependently for words. But at the same time this makes the above definition not exactly operationalisable for Verbmobil, since the extraction of *WL* from *C* will not be performed exactly dialoguewise or dialogue-turn-wise (but rather frequency-based, and to some degree intensionally defined, as we saw above). Therefore, we describe the following transfer-rule-based and operationalisable approach, which is an approximation of the above given definition:

> The translation equivalent of a wordlist *WL*, which was extracted from a dialogue corpus *C*, is the list of lemmata that occur on the right hand side of a transfer rule *T*, whose left hand side contains a semantic lemma with a corresponding entry in *WL*.

5.1 Practical Application

The basis of the Verbmobil German word recognition dictionaries are (until 1999) 1131 recorded dialogues in the scenarios of appointment scheduling, travel planning and hotel booking collected since 1993 (cf. JEKAT et al., 1997). They are a prerequisite for the training of acoustic and language models for speech recognition. Moreover, they are the empirical basis for grammar and domain modelling, and training material for other statistic-based processing such as statistic translation. Until the beginning of 1998, the number of all word form types in all German Verbmobil dialogues was 7349. Of these, only 3661 occurred more than once, i.e. 3688 were *hapax legomena*. The aim is to be able to process 10000 German words in the travel planning and hotel booking scenario by the year 2000.

	1998 I	1998 II	1999	2000
# dialogues	850	948	1128	
# word form tokens	342380	408112	489722	
# word form types	6249	7399	7926	
# hapax legomena	2489	2991	2878	
# morphological lemmata	5087	5983	6391	
# vm-whg-wordlist	5836	7126	8321	10000

Table 2: Lexicon-related corpus statistics in Verbmobil

The wordlist vmII-whg.wl.2.0 was generated from the currently available Verbmobil transliterations of spoken dialogues by means of a largeUNIX-Script. The wordlist obtained directly from the corpora was filteredand extended according to the criteria defined above. It contains 5836 wordform types, 4701 of which actually occurred in the dialogue corpora. Thegeneration procedure `wlgen` is displayed schematically in Figure 3.

6 Conclusion

We have defined lexical coverage criteria for consistent, multilingual lexica constructed for a Spoken Language Machine Translation environment. We have differentiated between extensional and intensional coverage criteria for wordlist selection, and we have discussed requirements of multilinguality and system architecture. Finally we described the automatic wordlist generation procedure used to define the lexicon of the Verbmobil system.

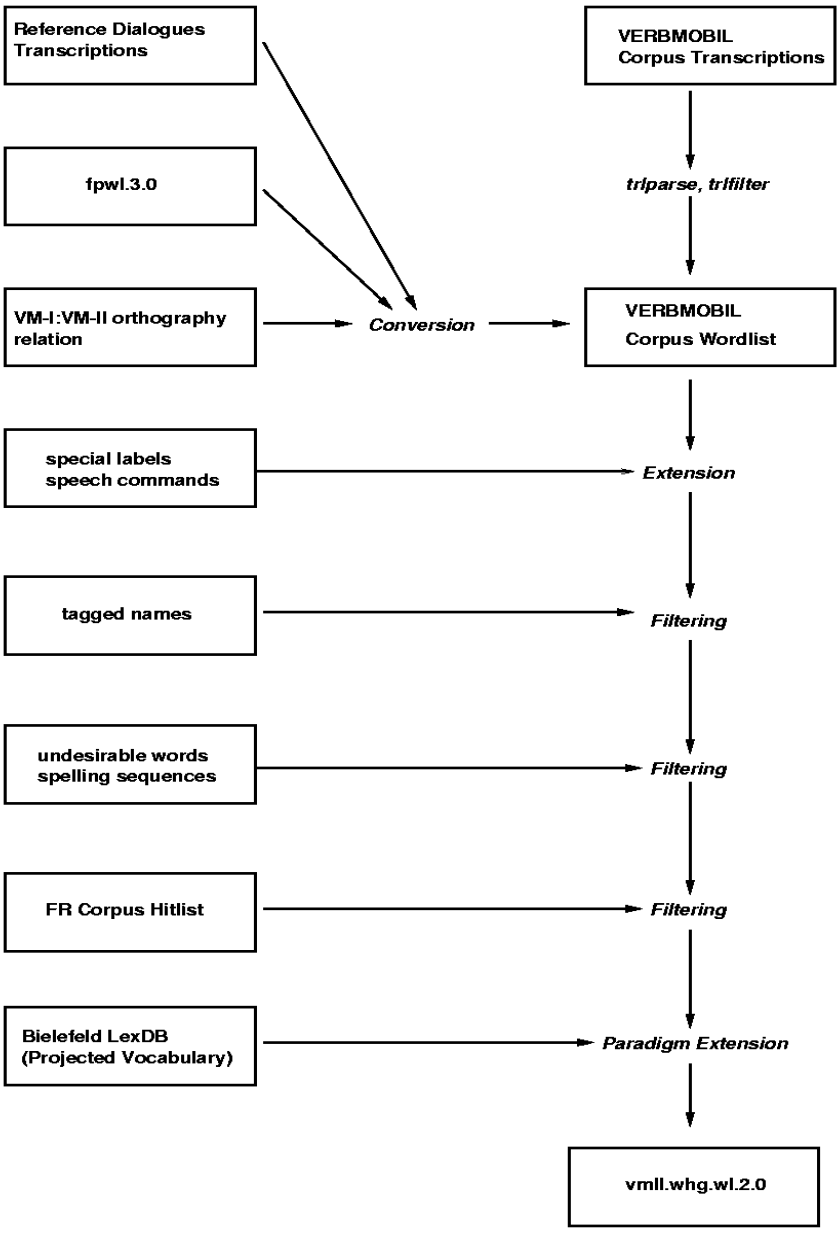

Figure 3: Architecture of the script `wlgen`

References

DEN OS, E. (1997): Spoken language corpus design. In: D. GIBBON, R. MOORE and R. WINSKI (eds.), EAGLES Handbook of Standards and Resources for Spoken Language Systems, pp. 79-118. Berlin: Mouton.

FISCHER, K. (1998): A cognitive lexical pragmatic approach to the polysemy of discourse particles. Ph.D. thesis, Universität Bielefeld.

FLICKINGER, D. (1987): Lexical Rules in the Hierarchical Lexicon. Ph.D. thesis, Stanford University.

GIBBON, D. (1991): ILEX: A linguistic approach to computational lexica. In: *Computatio Linguae*. Zeitschrift für Dialektologie und Linguistik, Beiheft 73.

GIBBON, D. and STEINBRECHER, D. (1995): VERBMOBIL-Standardfilter für Transliterationen Version 2.2. VERBMOBIL Technisches Dokument 38. Universität Bielefeld.

JEKAT et al. (1997): S. J., C. SCHEER and T. SCHULTZ, VMII Szenario: Instruktionen für alle Sprachstellungen. Universität Hamburg, LMU München, Universität Karlsruhe. Verbmobil Technisches Dokument 62.