Accessing Foreign Languages with COMPASS*

Elisabeth Breidt **Helmut Feldweg**
Seminar für Sprachwissenschaft,
Universität Tübingen,
Wilhelmstr. 113,
D-72074 Tübingen, Germany.
E-mail: {breidt,feldweg}@sfs.nphil.uni-tuebingen.de

February 28, 1996

Abstract

This paper presents an approach to avoid translations by utilizing a context–sensitive dictionary look–up system to assist in the comprehension of on–line texts in foreign languages. The system employs state–of–the–art finite state technology for analysis of the context and uses converted bilingual print dictionaries as its primary resource of lexical information. The print dictionaries are opened up for the system by analysing the dictionary structure and parsing the type–setting tapes. The lexical information is validated and augmented by corpus based lexicographic revision, where special attention is paid to the treatment of multi–word lexemes, for which formalized local grammars are added to the augmented dictionaries. Although the system is currently implemented as a device for comprehension assistance, the same technology and methods can be employed to select information from the dictionaries that is relevant for tranlation processes.

1 Comprehension Assistance instead of Translation

The growing availability of information in electronic form creates a new situation with respect to tools that permit the access to material in different languages. As long as the main type of document in a foreign language was written material that was meant to be of persistent value, translation was the best response to that need. Now however, often this material is not of a type that warrants translation: because the interest of the material itself is short–term, or because it is directed at only one user, or a small group of people, interested in exactly this material.

*Submitted to 'The Machine Translation Journal', special issue on new tools for human translators, edited by K.W. Church and P. Isabelle. This paper reports collaborative work done in the COMPASS project, funded under LRE 62-080 by the Commission of the European Communities. We would like to thank everybody involved in the project for their contributions.

This evolution represents a dramatic change in the type of tools that are needed, a change that has not yet been very widely recognised. It is obvious that in many cases a single user cannot be expected to achieve sufficient knowledge of the foreign language(s) that might be involved to reach a full understanding of the provided information.

In this situation we need tools other than standard translation: tools that give quick access to changing material where there is no need to store a 'text' in the foreign language. To give a couple of examples of such situations: the reading of foreign newspapers or other documents whose information value is short–term; the understanding of material that is not meant for a mass market, e.g. technical material, or which needs frequent updating.

Besides, a study undertaken by Bossard Consulting and OVUM within the EU's Multi–Lingual Action Plan (Engelien and Johnson 1991) uncovered the fact that the vast majority of business uses of documents in foreign languages by Europeans does not involve formal translation at all. In practice, in most countries there is a large set of people who have had some instruction in one or several other languages but who cannot use what they know efficiently. One major category here are the so-called 'faux débutants', people whose knowledge of the syntax of a foreign language is, in most cases, sufficient for them to figure out the basic relations between the parts of a sentence but who do not have a sufficient vocabulary to understand the sentence. Along with translation tools there is a need for tools that help people understand foreign language text on the fly.

The LOCOLEX system developed at Rank Xerox Research Centre, Grenoble, and the COMPASS prototype aim at filling this gap between full translation and genuine foreign language understanding[1].

2 What is COMPASS?

The COMPrehension ASSistant provides intelligent access to specially adapted, electronic dictionaries and in this way supports the understanding of on–line foreign language texts. Once a text is loaded into the COMPASS reader — the system supports HTML format — clicking on any word in the text starts a context analysis and dictionary look–up, and a selection of possible translations is presented in an unobtrusive way. The whole process being very fast, the user is distracted as least as possible from her primary task of reading the text.

Rather than producing a full translation of the foreign–language text, which often is condemned to be of poor quality — at least with current MT systems applied to unrestricted text —, a pre–selection of translations for the queried word is made. COMPASS filters out all information in the dictionary entry irrelevant to understand

[1]The only other attempt in this direction we know of is WordTranslator (Martin 1993), a prototype based on a passage of Van Dale's 'Unabridged Translation Dictionary English–Dutch'.

the foreign language expression. In case the information selection done by COMPASS might not be satisfactory, the user always has the possibility to see the full dictionary entry on demand.

COMPASS is a novel type of 'translation aid' based on electronic dictionaries. It makes use of existing translation information in bilingual dictionaries and takes over part of the intellectual task of dictionary look–up by using state–of–the–art NLP techniques. In the current system, this comprises:

- morphological analysis of the inflected text word and determination of the base form (or lemma) for look–up,
- determination of the part of speech (POS) of the queried word,
- accessing the relevant entries of the dictionary and storage of the complete information in a datastructure,
- check if the queried word is part of a multi–word lexeme (MWL) or other complex expression requiring a particular translation,
- marking of subsections of the entry fitting the context of the queried word for presentation to the user.

Resolution of ambiguities that are beyond currently known techniques, such as the difficult task of selecting the correct one among various (sub)senses — or the best translation, if this is desired — is left to the user. However, as dictionary information is presented very efficiently, this part is made as little time–consuming as possible.

Using the same dictionaries and techniques, but modified strategies for information selection and filtering, the system can also be used for machine–aided translation. Over and above that, COMPASS supports language learning 'on the fly', building on the users' existing knowledge of the foreign language. To this end, it can be integrated into a computer–assisted language learning (CALL) environment.

2.1 Project Details

The COMPASS project is funded by the Commission of the European Communities within the *Linguistic Research and Engineering* Program under LRE 62-080 from April 1994 to April 1996. The consortium consists of:

- Rank Xerox Research Centre (RXRC) (A. Zaenen), Grenoble, France,
- Seminar für Sprachwissenschaft (E. Hinrichs), Universität Tübingen, Germany,
- Fraunhofer Institut für Arbeitswirtschaft und Organisation (R.P. Wetzel), Stuttgart, Germany,
- Linguacubun LTD (N. Ostler), London, UK,
- Word Trade Centre (S. Atkins), Lewes, UK,
- Langues étrangères appliquées (H. Béjoint), Université Lumière, Lyon, France,
- Dept. of Marketing, Advertising and Public Relations (H. Lewis), Bournemouth University, UK.

Compass currently provides English–to–French and German–to–English language assistance, based on the Oxford–Hachette French Dictionary (OUP 1994; in the following OHEF) and the Collins German–English Dictionary (HarperCollins 1991; in the following CGE), for which the publishers kindly provided research licenses for the purposes of this project.

2.2 The System Components

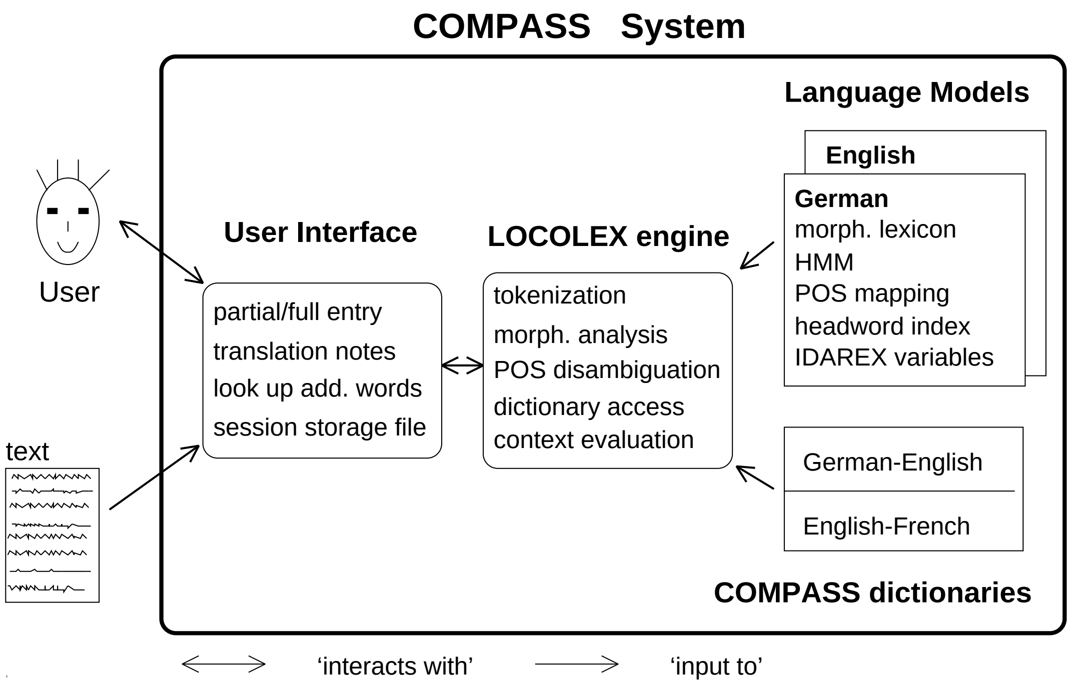

Figure 1: Compass Scenario

As illustrated in figure 1, the system has four main components. The Locolex engine links the various system components and controls the context evaluation and look–up process. For each source language (SL) there is a language model containing all language–dependent modules except the dictionaries. The bilingual dictionaries in Compass format are based on electronic versions of bilingual dictionaries which have been parsed into *dictionary databases (DDB)*. The DDBs have been lexicographically adapted and formalized to allow automatic evaluation of the dictionary entries. A graphical interface for use of Compass completes the system architecture. It presents a selection of translations, and more information on demand, up to the whole dictionary entry. Further, it allows the user to make translation notes for words looked up, and produces a 'Session Storage' file. In the following sections, we describe these modules in more detail.

3 The LOCOLEX Engine

LOCOLEX, the core look–up system utilised by COMPASS has been developed independently of the project and is patented by RXRC. The platform–independent architecture of LOCOLEX controls the access to the language–specific lingware components which are all realized as finite state networks (FSNs), with the notable exception of the SGML–tagged DDBs.

For each SL, language– and dictionary–specific knowledge is represented in a separate *language model*. The language models are interfaced to the LOCOLEX system, but are developed independently of the system using the RXRC finite state tools (Karttunen and Yampol 1993, Tapanainen 1994). A language model mainly consists of a morphological lexicon, information for POS disambiguation, an index to the dictionary entries, a mapping of morphological POS to POS information in the dictionary, and variable definitions used for the description of contextual patterns. In addition, the SGML–tagged DDB is supplied as a separate entity, together with regular expressions as a formal representation of the relevant contextual patterns in the dictionary.

In the following paragraphs, we describe the interaction of the LOCOLEX engine and a language model.

Tokenization A look–up process is triggered by a user's request for information on a word in the on–line text, normally by a mouse click. This activates the LOCOLEX engine, which, as its first action, isolates the sentence containing the activated word and breaks it down into individual words. This tokenization step turns the sentence and its words internally into a FSN, which are disclosed using a number of source language specific modules contained in the external language model. The realization of these modules as FSNs facilitates effective processing and convenient comparison with the sentence and word automata.

Morphological Analysis The following, language–specific processing step is the morphological analysis of each word in the sentence. For this purpose, a compiled version of a two–level based morphological analyser forms an integral part of each language model (Karttunen and Beesley 1992). For German, the morphological analyser used is a licensed derivate of Lingsoft's GERTWOL (Hausser 1996). For English, a morphological analyser of RXRC is used. Each of these analysers reduces inflected word forms to their base form and assigns a number of morpho–syntactic labels to every word in the sentence. Morpho–syntactic ambiguity is not resolved in this step, multiple analyses are assigned to ambiguous words.

Normalization Words not recognised by the morphological analyser receive special treatment by a SL–specific normalizing transducer: another FSN which maps frequent orthographic variations such as upper case at sentence beginnings, accents, umlauts etc. to the standard used by the morphological analyser. After normalization, these words are then morphologically re–analyseed.

Guesser If no morphological analysis can be assigned even after application of the normalizing transducer, a guesser is employed to assign morpho–syntactic information to unknown words. The guesser uses regular expressions to determine which labels should be assigned to the unknown word's character sequence. The regular expressions may also encode heuristic information: given that the morphological analyser has a fairly wide coverage of the language — as is the case for the analysers used for COMPASS — chances are high that unknown words belong to the group of proper names, neologisms, or acronyms.

Part-of-speech Disambiguation Morpho–syntactic ambiguities are resolved by a POS disambiguation module based on a first order Hidden–Markov–Model. The language–dependent parameters for this consisting of lexical and transition probabilities for ambiguity classes are stored within each language model. POS disambiguation is mainly done according to syntactic criteria (49 tags for German, 68 tags for English). Both models have reported success rates of over 96 % (Cutting et al. 1992, Feldweg 1995).

Dictionary Access The result of the preceding processing steps is a fully morphologically analysed representation of the sentence's words with disambiguated POS. The base form of the word selected by the user is looked up in a pre–compiled index of the dictionary to locate the dictionary entry. Spelling variants of dictionary lemmas are dealt with in the index by allowing multiple lemma forms to point to the same entry. To improve the speed of looking up and evaluating an entry, the dictionaries are pre–compiled into the LOCOLEX dictionary data structure which is common for all dictionaries used with COMPASS. Similarly, all regular expressions are pre–compiled into FSNs and stored with the dictionary.

POS mapping POS information is used as one of the major means to find relevant parts of the word's dictionary entry. Differences in the POS category systems used by the POS diambiguator and by the dictionary are resolved with a POS mapping transducer that relates POS tags from the disambiguator to the POS labels used in the dictionary. The POS category chosen by the disambiguator is mapped to the dictionary's category and the sub–entry matching this POS is marked — including any regular expressions defined within the sub–entry.

Context Evaluation Besides POS information, LOCOLEX takes into account the lexical context of the queried word. In the dictionary, lexical contexts requiring a particular translation can be stated in the form of regular expressions (REs) in the finite state formalism IDAREX (Segond and Tapanainen 1995). Details of the formal description of multi word lexemes (MWLs) and other multi–word expressions as REs are given in section 5.3.4 (see also Segond and Breidt (1995) and Segond et al. (1996)). All REs are pre–compiled into finite state networks. During look–up, those FSNs associated with the marked sub–entry are compared with the FSN of the input sentence. If a FSN from the dictionary can be mapped onto the sentence FSN, the corresponding contextual pattern in the dictionary entry with its translation is marked for selection, e.g. an idiom or a collocation. The entry is then passed to the interface for presentation, all relevant sections to be displayed, e.g. a matching idiomatic use, are marked in a particular way.

IDAREX uses the two–level morphological lexicon contained in the langugage model and is part of the Xerox Finite State Compiler (Karttunen and Yampol 1993, Tapanainen 1994). REs in IDAREX are sequences of morphological surface forms, base forms and IDAREX variables, combined by various operators. *Surface forms* are preceded by a colon, e.g. '`:went`' for the past form of 'to go'. If the RE should match various inflections of a word, the *base form* is used, together with a *morphological variable* that stands for particular morphological features. The colon at the end of the expression indicates the lexical side. For example, '`go V:`' allows any form of the verb, whereas '`go V3sg:`' is restricted to match the third person singular of the verb in any tense. The variable definitions are based on the two–level morphology used for that language and thus are part of the language models. Another type of variable, *word–class variables*, can be defined for complete classes of words sharing certain morphological features, e.g. `ADV` to match any adverb, or `POSS` for possessive pronouns. For complex syntactic categories, further variables can be defined as sequences of such basic word–class variables, e.g. `NPgen` for genitive noun phrases (in German).

Variability on the syntactic level, e.g. German word order variation[2], can be systematically represented with *macros*. A macro takes multiple parameters as arguments which can be surface or base forms, variables, operators or other macros. In each use of the macro, these parameters are then instantiated with the patterns for the particular multi–word expression described. An example is given in section 5.3.4 on p.23.

An advantage of the Xerox approach is that the FSNs are written as REs using morphological surface and base forms and allowing generalizations to be expressed on either side. The compactness, flexibility and modularity of the formalism considerably simplifies the description of the different patterns of variation occuring in MWLs compared to other finite state approaches that describe contextual patterns as FSNs, e.g. Maurel (1989) and Silberztein (1993).

[2]It is known that not all dependencies found in natural language can be represented with finite state techniques. They are, however, capable of representing a great number of phenomena in an efficient way.

4 The COMPASS Interface

The COMPASS interface is responsible for an easy information flow between the user and the system. It is designed as a text browser displaying HTML formatted input text and providing access to the look–up system by mousing a word. For testing and debugging purposes, it also provides a simple edit mode to change the input text and a read–only mode that allows reading of the text with no access to the look–up system. On the technical side, the interface is the only component of the COMPASS system that is platform–dependent: it has been developed for computer systems running Apple MAC OS 7.0 or higher.

The most prominent design principle of the COMPASS interface is to destract the user's reading process as little as possible. The user's attention should be kept on the on–line text being read rather than on the dictionary look-up process or the dictionary text. This is facilitated not only by the COMPASS system taking over the look–up process and (parts of) the information selection process, but also by optimal display of the selected information (cf. figure 2).

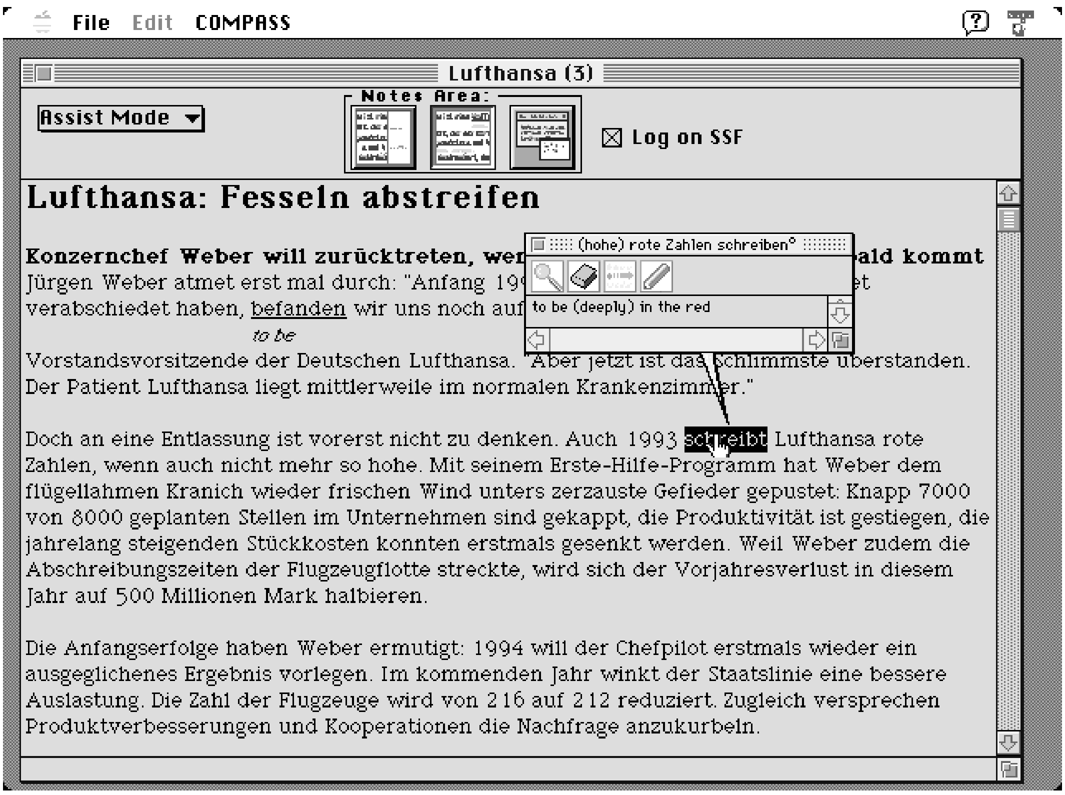

Figure 2: COMPASS interface screen dump

Assist Window If a user requests assistance for a word in the text, the interface passes the text to the look–up system and collects the information returned from it. This information normally consists of a full dictionary entry represented in the system–internal dictionary data structure. The most important task of the interface is to display the information marked as relevant by the system to the user in a non–disturbing way. To achieve this, the system determines a convenient location on the user's screen for an *assist pop–up window*. This is the place on the screen where a minimal amount of the word's context is hidden by the assist window. The window is just large enough to hold those parts of the dictionary entry which are most relevant in the given context. The title bar of the assist window displays the SL base form of the queried word, or, in the case of a matching MWL, the 'canonical' form of the MWL.

The assist window disappears after a user–configurable display time if the user does not perform any action on it. Alternatively, the user can explicitly close the window.

To enable COMPASS to select only translations relevant to language comprehension, each translation should ideally be marked as production–specific or as relevant for comprehension. However, current bilingual dictionaries do not formally distinguish whether several subsequent translations are given as production variants or as alternatives for different subsenses. As this is a very costly lexicographic task where each single entry is affected, COMPASS uses an approximation instead. Only the first translation within each sense group is presented in the assist pop–up window, to give the user an idea of the concepts related with the particular word. To reduce redundancy even more, if the same translation is possible for several senses of the queried word, it is displayed only once. Additional translation equivalents, usage information as well as examples and the like remain hidden.

Full Dictionary Entry Further information on the queried word, e.g. additional translations for various subsenses or usage information, is displayed to the user on request in the *partial entry view*. This view is activated by selecting a translation and clicking on the *magnifying class icon*. The partial entry view presents all the information contained in the dictionary for that particular sense.

The user is also given the option to click on the *book icon* to see the full dictionary entry, giving her the opportunity to back–up to intellectual analysis in case the system's disambiguation methods have failed, or to allow her to learn more about the uses of the word in question. In this mode, the system acts like a conventional electronic dictionary and no preselection of senses is active any more.

Translation Notes An additional feature of the interface is the possibility to make translation notes within the text using the *pencil icon*. Translation equivalents selected by the user during display of the assist window are copied to the text, either as side notes, interlinear between the lines, or into a separate note window, as defined in the user's settings. This way, the interface also serves as a memorizing device.

Session Storage File The life span of the information retrieved by the COMPASS system can be extended by activating the session storage file: a protocol text data file to which a variety of (configurable) information from the COMPASS system can be logged for later use. Its primary use is to record unknown vocabulary for later recapitulation, if desired with the related dictionary entry and sentence context. This functionality is an important characteristics of COMPASS to use it as a language learning tool. The session storage file has also been used within the project to supply data for the statistical evaluation of look–up times and number and types of looked–up words and expressions that had been protocolled during the two rounds of user testing organized within the project.

5 The COMPASS Dictionaries

Used with COMPASS, machine–readable, standard bilingual dictionaries can be turned into speedy reading or translation aids. To this end, the dictionaries must be turned into DDBs satisfying various requirements regarding format and contents. In this section, we will briefly go through these requirements. Part of these arise from the fact that in a certain sense, a COMPASS dictionary is a dictionary for human use and a computational lexicon for machine use at the same time, because the contained information is processed and evaluated by the system *and* is presented to the human user.

A COMPASS dictionary is a *decompacted DDB* (called 'database view of dictionary data' in Ide and Véronis (1995)) in SGML format: the detailed structure of the dictionary entries is indicated with SGML tags, quite similar to the recommendations for dictionaries in the TEI guidelines[3] (Sperberg–McQueen, Burnard 1994), and various compactions have been resolved. The whole conversion process is illustrated in figure 3.

Other database formats are also conceivable, but SGML is sufficient to make any single lexicographic item accessible. To ensure comfortable and flexible access to the SGML tagged dictionary data, we have developed an SGML–sensitive variant of the Unix *grep* utility, called SGREP[4]. SGREP is no longer line oriented and allows the search for (POSIX) regular expressions within SGML elements. Various restrictions on the SGML environment can be specified, such as the higher–order SGML element within which the search pattern should occur, or the SGML element which should

[3]We use a more detailed tag set that better captures the semantics of the tagged elements. All in all it should be possible to map our DTD (document type definition) onto the TEI DTD for dictionaries.

[4]Implementation by Michael Hipp.

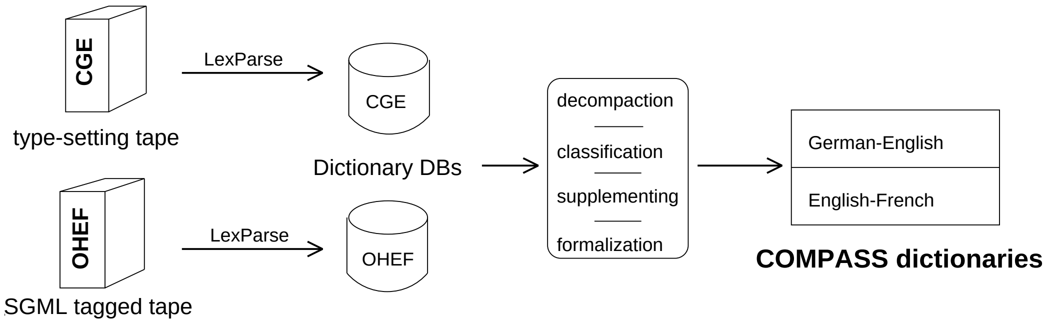

Figure 3: COMPASS dictionary conversion

be produced as output if the search pattern matches (e.g. output the lemma form if a particular collocator is used in the entry).

5.1 Comprehension vs. Translation Dictionary

Most bilingual dictionaries translating between two major languages are multi–functional (Bogaards 1990). Each language direction from SL to TL (target language) serves three purposes: translation from native SL into foreign TL (also called 'encoding'), understanding of the foreign SL, and translation from foreign SL into native TL (also called 'decoding')[5]. Hence, the dictionaries in parts contain information only relevant for one or the other of these functions. For an automatic evaluation of dictionary entries it is essential to identify the related lexicographic items[6] carrying such information. This is possible once the requirements described in the following subsections are met, allowing in addition the use of one and the same decompacted DDB for comprehension assistance and for translation assistance (into and from a foreign language).

5.2 From Raw Dictionary to Dictionary Database

Although in the meantime some electronic dictionaries are SGML tagged, the tags often express a mixture of typographic and structural information rather than the semantics (or lexicographic function) of the tagged elements. Also, the tagging usually does not describe the entry structure in enough detail such that the various types of information present in the entry can be processed by LOCOLEX. Besides, many dictionaries are still only available as type–setting tapes. Therefore, to get a completely, structurally SGML–tagged dictionary, namely a *parsed DDB* (called 'editorial view of dictionary data' in Ide and Véronis (1995)), the dictionaries are

[5] This is less true for bilingual dictionaries including a language of a small linguistic community, e.g. Dutch, where one language direction is usually ment only for encoding or for decoding.

[6] The computational lexicography terminology used in this article is based on Breidt (1994).

(re–)parsed with a dictionary entry parser[7].

Structural Analysis Before dictionary parsing, a profound structural analysis of the dictionary as a whole (macrostructure) and of its entries (microstructure) is necessary. This is very important to ensure that the resulting DDB can be correctly mapped onto the LOCOLEX dictionary datastructure. To describe the analysis method, we need to introduce the following terms: the text that constitutes a dictionary entry contains complex and elementary *lexicographic items* (e.g. sense group, translation equivalent resp.) and *non–typographic structure indicators* (e.g. comma, semicolon, brackets). In addition, there are *typographic structure indicators* (e.g. bold face, italics) that are properties of the text rather than part of it.

Following Storrer (1996), the procedure to get from the raw dictionary text to a parsed DDB can be split into four steps:

1. Development of a microstructure grammar

   Various entries of different types from the printed dictionary are split into minimal, lexicographically meaningful units, resulting in a preliminary list of terminal categories of the microstructure grammar. The combinations of these lexicographic items into bigger functional units are described with further, complex lexicographic items, yielding the non–terminal categories of the grammar and a number of context–free rewriting rules (see also Wiegand 1991, 22ff). Analysing further entries according to this method the resulting microstructure grammar is gradually improved[8].

2. Extension of the microstructure grammar into an entry structure grammar

   In the next step, the inventory of the various structure indicators and their functions are determined. Their addition to the microstructure grammar yields an entry structure grammar. Typographic structure indicators should not form part of the grammatical categories, though. Their functions can be considered indirectly through attributes of the categories or by creating subtypes of the lexicographic items in question.

3. Adapting the entry structure grammar to a parser–specific grammar

   Only in this step, specific characteristics and structure–indicating information of the electronic dictionary source are considered and related to the structure indicators described in the entry structure grammar. On this basis, the entry structure grammar is translated into a grammar to be used for parsing.

[7]The analysis and parsing of the two dictionaries used in the project is described in Breidt and Thielen (1996).

[8]The microstructure grammar for the CGE, for example, contains 59 terminal and 32 non–terminal categories, and 30 rewriting rules.

4. Modification of the parser–specific grammar according to parse results

   Those entries that could not be parsed in a first run of the parser are analysed, leading to corrections of the grammar or of the entries themselves if they contradict one's lexicographic intuitions. This step is repeated until all well–formed entries are successfully parsed. Wrong or inconsistent use of structure indicators and varying lexicographic items for the same type of information are two major sources of error[9].

Parsing We have used the general–purpose dictionary entry parser LexParse developed at the University of Tübingen (Hauser 1993) to parse the CGE and to re–parse the OHEF. Rather than discuss the special properties of the parser (cf. Hauser and Storrer 1994), in figures 4 to 7 we give a sample entry, first as presented in the printed dictionary and as on the rudimentary SGML–tagged tape, and then after parsing (SGML and tree notation).

> **anfangen** *sep irreg* **1** *vt* **(a)** (*beginnen*) *Arbeit, Brief, Gespräch, (inf: anbrauchen) neue Tube etc* to start, to begin; *Streit, Verhältnis, Fabrik* to start.
> **(b)** (*anstellen, machen*) to do. **das mußt du anders** ~ you'll have to go about it differently; **was soll ich damit ~?** what am I supposed to do with that?; (*was nützt mir das?*) what's the use of that?; **damit kann ich nichts** ~ (*nützt mir nichts*) that's no good to me; (*verstehe ich nicht*) it doesn't mean a thing to me; [...] .
> **2** *vi* to begin, to start. [...] **es fing zu regnen an** *or* **an zu regnen** it started raining *or* to rain; [...] **fang nicht wieder davon** *or* **damit an!** don't start all that again!, don't bring all that up again!; **mit etw** ~ to start sth; **klein/unten** ~ to start small/at the bottom; **er hat als kleiner Handwerker angefangen** he started out as a small–time tradesman; **bei einer Firma** ~ to start with a firm or working for a firm.

Figure 4: Printed dictionary entry from the CGE (HarperCollins 1991)

To obtain a proper SGML document, LexParse can generate a DTD on the basis of the dictionary grammar. If desired, the SGML–tagged dictionary entries can then be loaded into an SGML editor for post–editing.

The next step is to convert the *parsed* DDB into a *decompacted* DDB to make the information contained in the DDBs fully accessible for Locolex. Depending on the dictionary parser used, part of this can already be done during pre–/post–processing as with LexParse. In the next section, we give a summary of the necessary (lexicographic) adaptations.

[9]In the rudimentary SGML–tagged version of the CGE kindly put at our disposal, about 600 entries had to be corrected, which is not too bad (approx. 0.7 %) for about 84.500 entries processed altogether.

```
*****
<HWME> a>.<nfangen
<HWGR> sep irreg
<CT1P> 1
<POSP> vt
<CT2M> a
<LBIN> beginnen
<COLL> Arbeit, Brief, Gespräch
<LBRR> inf
<LBRR> YYY
<LBIN> anbrauchen
<COLL> neue Tube etc
<TRAN> to start
<TRAN> to begin
<COLL> Streit, Verhältnis, Fabrik
<TRAN> to start
<CT2M> b
<LBIN> anstellen, machen
<TRAN> to do
<PHRS> das mußt du anders a>.<nfangen
<TRAN> you'll have to go about it differently
<PHRS> was soll ich damit a>.<nfangen?
<TRAN> what am I supposed to do with that?
<LBIN> was nützt mir das?
<TRAN> what's the use of that?
...
<PHRS> das fängt ja schön {or} heiter an!
<LBRR> iro
<TRAN> that's a good start!
...
```

Figure 5: Unparsed dictionary entry from the CGE tape

5.3 From Dictionary Database to COMPASS Dictionary

The main fields of adaptation, described in the following subsections, are (i) decompactions, (ii) retagging of lexicographic items, mainly concerning the various types of phrases, (iii) corpus–based supplementing of data, and (iv) formalization of contextual information. Some of these tasks can be fully automated, for others heuristics can be applied, but many of these adaptations require individual decisions on a case by case basis and can thus not be fully automated[10]. Sceptical readers who might think that such an amount of semi–automatic or even manual dictionary adaptation is undesirable should keep in mind how much more manual work is necessary to build a comparable full–scale, broad–coverage MT lexicon, even on the basis of existing dictionary data.

[10] It is envisaged to support this manual editing with interactive tools making suggestions for decompactions.

```
<ENTRY ID=a2801><DE><SE>
<FG><LF>anfangen</LF><FGGRU>sep irreg</FGGRU></FG>
<HNO>1</HNO>
<HG><GRG><POS>vt</POS></GRG>
    <SG>
       <SNO>a</SNO>
       <SENSE><USGS><SUS><SIND>beginnen</SIND></SUS>
                <COLL>Arbeit, Brief, Gespräch</COLL></USGS>
           <USGS><USGCOMBS><LDIAS><LSTYL>inf</LSTYL></LDIAS>
                     <SUS><SIND>anbrauchen</SIND></SUS></USGCOMBS>
                <COLL>neue Tube etc</COLL></USGS>
           <SUB><TR><EQ>to start</EQ></TR>
                <TR><EQ>to begin</EQ></TR></SUB>
           <USGS><COLL>Streit, Verhältnis, Fabrik</COLL></USGS>
           <SUB><TR><EQ>to start</EQ></TR></SUB></SENSE>
       <SNO>b</SNO>
       <SENSE><USGS><SUS><SIND>anstellen, machen</SIND></SUS></USGS>
           <SUB><TR><EQ>to do</EQ></TR></SUB>
           <MWG><MWE><MWEFORM>das mußt du anders anfangen</MWEFORM></MWE>
                <SUB><TR><EQ>you ’ll have to go about it differently</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>was soll ich damit anfangen?</MWEFORM></MWE>
                <SUB><TR><EQ>what am I supposed to do with that?</EQ></TR></SUB>
                <USGS><SUS><SIND>was nützt mir das?</SIND></SUS></USGS>
                <SUB><TR><EQ>what ’s the use of that?</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>damit kann ich nichts anfangen</MWEFORM></MWE>
                <USGS><SUS><SIND>nützt mir nichts</SIND></SUS></USGS>
                <SUB><TR><EQ>that ’s no good to me</EQ></TR></SUB>
                <USGS><SUS><SIND>verstehe ich nicht</SIND></SUS></USGS>
                <SUB><TR><EQ>it doesn ’t mean a thing to me</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>nichts mit sich/jdm anzufangen wissen</MWEFORM></MWE>
                <SUB><TR><EQ>not to know what to do with oneself/sb</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>mit dir ist heute (aber) gar nichts anzufangen!</MWEFORM></MWE>
                <SUB><TR><EQ>you ’re no fun at all today!</EQ></TR></SUB></MWG>
   ... </SENSE></SG></HG>
<HNO>2</HNO>
<HG><GRG><POS>vi</POS></GRG>
    <SG><SENSE>
           <SUB><TR><EQ>to begin</EQ></TR>
                <TR><EQ>to start</EQ></TR></SUB> ...
           <MWG><MWE><MWEFORM>mit etw anfangen</MWEFORM></MWE>
                <SUB><TR><EQ>to start sth</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>klein/unten anfangen</MWEFORM></MWE>
                <SUB><TR><EQ>to start small/at the bottom</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>er hat als kleiner Handwerker angefangen</MWEFORM></MWE>
                <SUB><TR><EQ>he started out as a small-time tradesman</EQ></TR></SUB></MWG>
           <MWG><MWE><MWEFORM>bei einer Firma anfangen</MWEFORM></MWE>
                <SUB><TR><EQ>to start with a firm {or} working for a firm</EQ></TR>
                </SUB></MWG></SENSE></SG>
</HG></SE></DE></ENTRY>
```

Figure 6: LEXPARSE SGML output (manual indentation)

```
Collins (German/English)
  +--> DE (dictionary entry)
     +--> SE (standard entry)
        +--> FG (form group)
        |  +--> LF (lemma form): "anfangen"
        |  +--> FGGrU (form group grammatical usage item): "sep irreg"
        +--> HNo (homograph counter): "1"
        +--> HG (homograph group)
        |  +--> GrG (grammatical group)
        |  |  +--> Pos (part of speech): "vt"
        |  +--> SG (sense group)
        |     +--> SNo (sense counter): "a"
        |     +--> Sense (sense)
        |     |  +--> UsgS (usage items (source language))
        |     |  |  +--> SUS (semantic usage items (source language))
        |     |  |  |  +--> SInd (sense indicator): "beginnen"
        |     |  |  +--> Coll (collocator): "Arbeit, Brief, Gespräch"
        |     |  +--> UsgS (usage items (source language))
        |     |  |  +--> UsgCombS (combined usage items)
        |     |  |  |  +--> LdiaS (diasystematic label (source language))
        |     |  |  |  |  +--> Lstyl (style label): "inf"
        |     |  |  |  +--> SUS (semantic usage items (source language))
        |     |  |  |     +--> SInd (sense indicator): "anbrauchen"
        |     |  |  +--> Coll (collocator): "neue Tube etc"
        |     |  +--> Sub (subsense)
        |     |  |  +--> Tr (translation)
        |     |  |  |  +--> Eq (translation equivalent): "to start"
        |     |  |  +--> Tr (translation)
        |     |  |     +--> Eq (translation equivalent): "to begin"
        |     |  +--> UsgS (usage items (source language))
        |     |  |  +--> Coll (collocator): "Streit, Verhältnis, Fabrik"
        |     |  +--> Sub (subsense)
        |     |     +--> Tr (translation)
        |     |        +--> Eq (translation equivalent): "to start"
        |     +--> SNo (sense counter): "b"
        |     +--> Sense (sense)
        |        ...
        |        +--> MWG (multi-word expression group)
        |        |  +--> Mwe (multi-word expression)
        |        |  |  +--> MweForm (multi-word expression form): "was soll ich damit anfangen?"
        |        |  +--> Sub (subsense)
        |        |  |  +--> Tr (translation)
        |        |  |     +--> Eq (translation equivalent): "what am I supposed to do with that?"
        |        |  +--> UsgS (usage items (source language))
        |        |  |  +--> SUS (semantic usage items (source language))
        |        |  |     +--> SInd (sense indicator): "was nützt mir das?"
        |        |  +--> Sub (subsense)
        |        |     +--> Tr (translation)
        |        |        +--> Eq (translation equivalent): "what 's the use of that?"
        ...       ...
Entry #1: Accepted!
```

Figure 7: LEXPARSE Tree output

The editing of the SGML DDB should preferably be done with an SGML editor taking advantage of the DTD[11]. Within the project, all automatic conversions were done on the complete dictionaries, while manual conversion tasks were restricted to the vocabulary of various selected test texts, between 1000 and 2000 entries per source language.

5.3.1 Decompactions

Traditional bilingual dictionaries have been produced for use by humans, so far mainly in printed form. This purpose has lead to a number of compaction schemes (cf. Wolski 1991) that are fairly easily interpreted by the human user and are used in most dictionaries. However, for processing of the dictionary information by LOCOLEX, such condensed information has to be resolved. This includes:

- Splitting of complex information into separate lexicographic items:
  – lemma variant forms (*Neugier(de), Bade(hand)tuch*)
  – translation variants (*klein anfangen – to start small/at the bottom*)
  – MWL variants (*es fing zu regnen an* {*or*} *an zu regnen – to hold to oneself*).
- Making implicit information explicit (e.g. POS for nouns in the CGE, given implicitly through the gender information restricted to nouns; indirect cross references of the type '*see vt*' in entries of derived nominals).
- Expansion of niche/nest entries (e.g. compound blocks) into separate full entries.
- Expansion of partial lemma forms, e.g. morphological variant forms (German *Lehrer, –in* → *Lehrer, Lehrerin*)
- Addition of explicit translation equivalents where only usage examples are given instead to save space, e.g. in the CGE:

  > **aufhalsen** *vt sep (inf)* **jdm/sich etw** ~ to saddle *or* land sb/oneself with sth (*inf*), to land sth on sb/oneself (*inf*); ....

- Substitution of place holder symbols, e.g. tilde replacing the lemma form.

Whereas most of these tasks can be (and have been) automated, the decompaction of translation and MWL variants is more difficult. The scope of the *or* and '/' used to indicate disjunctions does not follow predictable rules and is sometimes even difficult to determine for non–native speakers. Similarly, the addition of explicit translations and the substitution of indirect cross references require manual editing.

[11] In the project, AuthorEditor (by SoftQuad) proved to be a good choice under Unix (less so on the MAC), at least for the dictionary extracts edited in the project.

5.3.2 Improvements of the Tagging

Corrections of the tagging results of the parser are not limited to tags that are structurally correct but semantically wrong because the information should be presented as a different lexicographic item (e.g. object collocators tagged as sense indicators/meaning paraphrases). As discussed above (5.1), bilingual dictionaries contain a lot of information mainly relevant for language production, to which among other things belong many of the so–called 'phrases'. In the CGE, the various types of phrases are not distinguished at all, so a complete classification was necessary. In the SGML–tagged tape of the OHEF, idioms, compounds, lexical collocations, phrasal verbs, grammatical collocations and examples are tagged differently. Unfortunately, the tagging is quite inconsistent, so many phrases had to be re–tagged.

In COMPASS, we distinguish between usage examples and semantically transparent collocations on the one hand, and MWLs, fixed phrases and grammatical collocations on the other hand. The former are only important for (stylistically adequate) language production in the TL, the latter are also relevant to properly understand the SL expression. For example, in the CGE in addition to the grammatical collocation *sich in etw or zu etw verwandeln* (to change or turn into sth), another phrase is given with an additional possible translation variant for this use of *verwandeln*, information which is only interesting for language production: *Zeus hat sich in einen Stier verwandelt* (Zeus turned or transformed himself into a bull).

Multi–word lexemes comprise idioms, idiomatic expressions, support verb constructions and other complex predicates, multi–word compounds, (lexical) collocations, and phrasal verbs. *Fixed phrases* include proverbs, adverbial MWLs and complex function words, and any other semantically non–transparent constructions that are completely invariable. *Grammatical collocations* illustrate syntactic environments and subcategorization patterns of the lemma and thus specify syntactic or grammatical contexts of the lemma[12]. On the basis of such a classification of the various phrases in the dictionary, MWLs, fixed phrases and grammatical collocations can then be formalized (see 5.3.4) so that LOCOLEX is able to check them against the sentence context being read.

5.3.3 Corpus–based Supplementing

To improve dictionary coverage[13], various techniques for data extraction from corpora have been successfully applied in the project. For German, CQP and the interface XKWIC (Christ 1993, 1994; Schulze and Christ 1994) have been used as concordance tools, together with corpora POS–tagged at the University of Tübingen, to find corpus evidence for particular word usages and word combinations. In addition, NGRAM (Feldweg and Hipp 1993) could be used to calculate ngrams, their

[12] A more detailed definition of the various phrase types is given in Breidt (1995).

[13] Though the OHEF has already been produced using text corpora, to our knowledge the CGE is not based on electronic corpora.

frequencies and mutual information values based on the POS–tagged corpora. On the basis of bi– and trigrams for particular syntactic patterns, their frequencies and MI values, candidate lists for MWLs were automatically produced, following Church and Hanks (1989) and Breidt (1993).

For English, KWIC lines extracted from the Oxford Corpus (OUP, more than 60 million word tokens) could be used which were subsequently POS–tagged with the Xerox DDS tagger (Kupiec and Wilkens 1994), together with t–scores for all bigrams. In addition, various syntactic relations (subject, object, modifier) for the bigrams were calculated with SEXTANT (Grefenstette 1994).

These tools were used to detect missing MWLs, grammatical collocations and MWL variants. For example, the following bigrams for *anfangen/angefangen/-anzufangen*[14] were selected with a simple *grep* command from the NGRAM file containing all adjacent bigrams of the POS–tagged 'Frankfurter Rundschau' corpus (40 million word tokens). As long as no lemmatized corpora for German are available, the bigrams can only be calculated for word forms. For example, the variant forms *(von) vorne anfangen* and *(von) vorn anfangen* are considered as two distinct bigrams. The bigrams are sorted according to descending MI value, with a minimal cooccurrence frequency of 3.

```
MI      f(x,y)  f(x)   f(y)       x           y
16.750  3       3       357     zeitiger_ADJD angefangen_VPP
13.942  3       21      357     Kampfsport_NN angefangen_VPP
10.921  23      1051    444     vorne_ADV anfangen_VINF
10.493  6       369     444     Null_NN anfangen_VINF
10.164  4       309     444     null_CARD anfangen_VINF
9.996   3       1051    110     vorne_ADV anzufangen_VIZU
9.807   3       369     357     Null_NN angefangen_VPP
9.266   5       895     357     klein_ADJD angefangen_VPP
9.195   5       756     444     vorn_ADV anfangen_VINF
8.998   7       4896    110     neu_ADJD anzufangen_VIZU
8.845   6       1156    444     unten_ADV anfangen_VINF
8.552   14      13341   110     nichts_PROS anzufangen_VIZU
8.510   4       1209    357     Dornbusch_NN angefangen_VPP
8.346   49      13341   444     nichts_PROS anfangen_VINF
8.305   4       1121    444     Viel_NN anfangen_VINF
8.214   3       895     444     klein_ADJD anfangen_VINF
8.031   4       5471    110     wenig_PVIEL anzufangen_VIZU
8.016   7       9670    110     viel_PVIEL anzufangen_VIZU
7.924   15      5471    444     wenig_PVIEL anfangen_VINF
7.847   3       1436    357     spät_ADJD angefangen_VPP
7.763   12      4896    444     neu_ADJD anfangen_VINF
7.692   3       1599    357     Leuten_NN angefangen_VPP
7.638   9       16156   110     damit_PROAV anzufangen_VIZU
7.517   3       1451    444     Verkehrsberuhigung_NN anfangen_VINF
7.460   3       19714   347     da_KOUS anfangen_VFIN
7.283   17      9670    444     viel_PVIEL anfangen_VINF
...
```

[14]See Breidt (1993) on the use of the infinitive forms only for automatic MWL extractions.

Going through the bigrams it is very easy for the lexicographer editing the dictionary to add the remaining MWL components in addition to the parts of the bigrams. In dubious cases, and also to check for the variability of a MWL, a query with the corpus tools is possible to verify the lexicographer's hypothesis. Especially if one adds up the various inflection forms, the following MWLs are clearly evidenced:

> *von vorn(e) anfangen* (to start from the beginning/to start from scratch)
> *bei Null/null anfangen* (to start from scratch)
> *klein/unten anfangen* (to start small)
> *neu anfangen* (to make a new start)
> *mit etw nichts/nicht viel/wenig anzufangen wissen* (not to know what to do with sth)
> *mit jdm ist nichts/nicht viel/wenig anzufangen* (to be no fun (at all))
> *mit jdm nichts/nicht viel/wenig anfangen können* (not to be of (any) use for sb/not to mean a thing to sb)

Most of these are also mentioned in the CGE entry for *anfangen* or in other entries, but not all, e.g. *bei Null anfangen, neu anfangen* are missing.

The KWIC lines were mainly used as evidence for the modifiability of MWLs. The KWIC lines in figure 8 have been produced from the 'Frankfurter Rundschau' corpus by searching for occurrences of 'die Fäden' (99 times) with 'Hand' as collocate (18 times). The concordance gives evidence that the idiom *die Fäden in der Hand halten* (to hold the reins) allows *haben* (lines 3, 9, 15–19) and *behalten* (l. 14) as variants of the verb. Besides, *die Fäden* can be modified by a genitive NP (l. 4, 7).

```
1   Ob Bein in Istanbul wieder <die Fäden> in der Hand hält , entscheidet sich erst
2   r im Schatten eines Großen <die Fäden> in der Hand hält , ohne eigene Ambition
3   mand mehr genau weiß , wer <die Fäden> in der Hand hat und wer wirklich die Pup
4   euchtkäfer in den Kulissen <die Fäden> des Spektakels in der Hand hält , entwir
5   abseits der tobenden Menge <die Fäden> in der Hand hält . Und dann geht's los ,
6   erliner Schalousie-Theater <die Fäden> in der Hand . Für Erwachsene wird im Par
7   hmen . Deng hält seit 1978 <die Fäden> der chinesischen Politik in der Hand und
8   n hält auch in der Politik <die Fäden> in der Hand . Er rühmt sich : " Kluge Le
9   ogenannte Dalang hat dabei <die Fäden> in der Hand . Der Priesterähnliche ist
10  er des Chronik-Ausschusses <die Fäden> in der Hand hielt . Es sei ihm eine Freu
11  te Drogenboß Pablo Escobar <die Fäden> in der Hand . Er soll auch hinter den En
12  itet , schon von Anfang an <die Fäden> in der Hand . Als besondere Attraktion
13  eindevertreter - Börs hält <die Fäden> in der Hand . " Ob politische Konsequenz
14  hen des Ägypters erklärt , <die Fäden> in der Hand zu behalten . Kurz nach sei
15  auben , der der eigentlich <die Fäden> in der Hand haben müßte , wird zum Spiel
16  . Diese hätten schließlich <die Fäden> in der Hand . Ausbildung in Metallbetri
17  mand mehr genau weiß , wer <die Fäden> in der Hand hat und wer wirklich die
18  son dieser Gruppe und habe <die Fäden> in der Hand . cas Tagestip : Eisenbahn
```

Figure 8: KWIC lines for 'die Fäden in der Hand halten'

In certain cases, the evaluation of KWIC lines and bigrams led to an additional word sense missing in the dictionary, or to omitted variant forms of a lemma. In a similar way, possibly with additional terminology extraction tools, we plan to extract from domain–specific corpora terminology specific to a new application domain of COMPASS and add it to the dictionaries.

5.3.4 Formalization of Lexical Context

As the last step in the preparation of a COMPASS dictionary, information on the lexical contexts of the lemmas needs to be formally represented in order for LOCOLEX to take this into account for translation selection. After the relevant lexical patterns have been identified as such, e.g. MWLs, fixed phrases and grammatical collocations as described in the previous sections, the formalization proceeds as follows.

Canonical Form First, the base form or *canonical form* of the multi–word expressions is determined. Together with the classification of the various phrases, this makes up most of the manual work within the formalization procedure. If MWLs were already given in a consistent way in canonical form in the dictionary sources[15], ideally also distinguishing between the various types of multi–word expressions, the formalization could be almost completely automated (assuming that all types of variation are already considered).

The canonical form is a minimal subset of the given phrase where any word that is not essential for the meaning of the multi–word expression is removed. For fixed phrases, the canonical form consists only of the lexically fixed components in the correct ordering. Alternative spellings can be indicated as usual by bracketing single letters, as in (a). Similarly, optional parts are bracketed. For lexical variants, a slash '/' separates the alternatives. If the scope is more than one word on each side, it is indicated with circumflex signs '^' as in (b).

For grammatical collocations, besides the lemma itself the canonical form contains only function words and meta–expressions such as *sb, sth, one's, oneself* for English or *jd, jdn, etw, sich, seinen* for German. The words should be given in unmarked word order and in their base form (i.e. non–graduated adjective, infinitive verb, noun in singular etc.), as in (c) and (d).

(a) wie im Flug(e) (in a twinkling)

(b) ^nicht um alles/um nichts^ in der Welt (not for love nor money)

(c) sich gegen etw abschotten° (to cut oneself off from sth)

(d) das ist ihm eigen (that is typical of him)
→ jdm/etw (ADVP) eigen sein° (to be typical of sb/sth)

[15]The Longman Dictionary of English Idioms (Longman 1979) — though not a bilingual dictionary — is a first step in this direction.

For MWLs, those words requiring a particular morphological (surface) form are of course used in that form, whereas morphologically flexible words are given in their base form, marked with a degree sign '°'. External arguments that are not lexically fixed should be represented by meta–expressions. Possible modifications not systematically allowed for all MWLs of a particular syntactic class can be indicated with variables such as `ADJ, NUM, ADV`. The same variables can be used for external arguments not covered by the meta–expressions, e.g. `NPgen` below. Some examples from the CGE:

(e) das geht über meine Kräfte (it's more than I can take)
→ über jds Kräfte gehen° (to be more than one can take)

(f) er hält die Fäden (fest) in der Hand (he holds the reins (tightly))
→ die Fäden (NPgen) (fest) in der Hand halten°/haben°/behalten°
(to hold the reins)

(g) er hat jahrelang ohne jede Beanstandung seine Pflicht getan
(for years he did his duty without giving any cause for complaint)
→ ohne jede Beanstandung (without giving any cause for complaint)
→ seine Pflicht tun° (to do one's duty)

It is important to consider all possibilities of lexical variation and modification, also those mentioned in other dictionary entries or found in the corpus. Otherwise the corresponding regular expression does not match against all possible sentence contexts. If the canonical form gets too long and clumsy, lexical variants can also be coded as separate items.

Regular Expressions From the canonical form, a corresponding regular expression in IDAREX format is automatically generated[16]. For morphologically flexible words, the program generates their base form and the corresponding IDAREX morphological variable. Morphologically fixed words are represented as surface forms (preceded by a colon). Lexical alternations are correctly represented as disjunctions, both for partial and for full words, and optional elements are bracketed. The lexicographic meta–expressions are replaced by the corresponding IDAREX variables, e.g. *sich, jds, jdm, seine* by `REFL, NPgen, NPdat, POSS` respectively. The current formalization program also covers standard German word order variation, i.e. verb–first/–second and verb–final word order for verbal MWLs. If the verb has a separable prefix (as for example *abschotten* in (c)), this is taken into account for verb–first/–second order. German MWLs with a lexically fixed subject, e.g. *jdm geht ein Licht auf* (to dawn on sb), follow different word order patterns and have to be formalized separately.

[16]This step is implemented for German (program developed by Michael Hipp). The principles and mechanisms also apply for English and other languages.

For the canonical forms above, the following IDAREX regular expressions are generated (a more detailed description how to represent variable MWLs in IDAREX is given in Segond and Breidt (1995)):

(a) wie im Flug(e) →
```
:wie :im [ :Flug | :Fluge ] ;
```

(b) ^nicht um alles/um nichts^ in der Welt →
```
[ :nicht :um :alles | :um :nichts ] :in :der :Welt ;
```

(c) sich gegen etw abschotten° →
```
[ schotten Vfin:  ANY* REFL ANY* :gegen NP ADV* ab Pref:  |
REFL ANY* :gegen NP abschotten V: ] ;
```

(d) jdm/etw (ADVP) eigen sein° →
```
[ sein Vfin:  ANY* NPdat (ADVP) :eigen |
NPdat (ADVP) :eigen ( ZU ) sein V: ] ;
```

(e) über jds Kräfte gehen° →
```
[ gehen Vfin:  ANY* :über NPgen :Kräfte |
:über NPgen :Kräfte ( ZU ) gehen V: ] ;
```

(f) die Fäden (NPgen) (fest) in der Hand halten°/haben°/behalten° →
```
[ [ halten Vfin:  | haben Vfin:  | behalten Vfin:  ] ANY* :die
:Fäden (NPgen) (:fest) :in :der :Hand |
:die :Fäden (NPgen) (:fest) :in :der :Hand ( ZU ) [ halten V: |
haben V: | behalten V: ] ] ;
```

(g) *seine Pflicht tun°* →
```
[ tun Vfin:  ANY* POSS :Pflicht | POSS :Pflicht ( ZU ) tun V: ] ;
```

In IDAREX, systematic structural patterns can also be expressed with macros. This is particularly useful if regular expressions have to be written manually. For example, a macro `WO` to cover verb–first/–second and verb–final word ordering of German verbal MWLs can be defined as follows:

```
[ $1 Vfin:  ANY* $2 | $2 ( ZU) $1 V: ]
```

The macro is used with a set of parameters which are inserted at run–time for the $–place–holders, e.g.

```
WO( fix2( POSS :Pflicht ) tun ) ;

WO( fix11( :die :Fäden (NPgen) ( :fest ) :in :der :Hand ) halten) ;
```

where `fix2` and `fix11` are auxiliary macros necessary to allow for variable length of the fixed component of verbal MWLs. In a similar way, separable prefix verbs in German are automatically formalized, using a macro `PREFV`.

The automatically produced regular expressions can be manually post–edited to include special variations that are not automatically covered by the generation mechanism because they do not hold for all MWLs, such as topicalization or scrambling of a fixed element in German. For example, with the idiom *den Kopf für etw hinhalten* (to take the blame for sth) scrambling is possible: *für etw den Kopf hinhalten*, and topicalization of the fixed element 'den Kopf': *Den Kopf halte ich dafür aber nicht hin* (but I am not going to take the blame for this).

The regular expressions can be understood as *local grammars* specifying valid morpho–syntactic environments for MWLs (see also Segond et al. 1996). This way of describing the variability of MWLs adds a dynamic aspect to the COMPASS dictionaries. The SGML tags that mark the MWLs do not point to static text but to small (finite state) programs.

6 Conclusion

In two cycles of user testing of the COMPASS system, the project has shown that state–of–the–art techniques for natural language processing combined with carefully prepared machine–readable dictionaries can considerably improve the comprehension of foreign–language on–line texts. An interesting finding in the user tests was, that word–steming – a processing step originally played down as a necessary pre–requisite for further processing – was by itself considered as a very helpful feature of the system, particularly for more inflecting languages like German. Extra–linguistic devices such as multiple indexing for spelling variants and access to information for MWLs from every content word of the MWL further improve the assistance quality.

The usefulness of POS–disambiguation techniques for discrimination of dictionary sub–entries varies with the SL. While POS is of little help for German, where it discloses itself frequently by inflectional properties of a word, it plays an important role in English, where a given word form can frequently be assigned to various major POS categories (cf. Feldweg 1995).

The linguistically most ambitious device currently implemented in the COMPASS system is the use of REs to define local grammars for the variability of MWLs, which finds expression in the fact that recognition of REs in the texts is the technically most costly component in the COMPASS system. This pays off however by yielding easy access to extensive information on MWLs contained in dictionaries which the user normally does not fully exploit due to inappropriate sequential access techniques.

A number of extensions to the current system should result in an even better performance with respect to the sense discrimination task:

Collocator lists given in many dictionary entries can be used as additional lexical contexts to match against the sentence network.

Sense disambiguation techniques can be implemented in the LOCOLEX system, along with semantic classifications and selectional restrictions added to the dictionary.

Domain Labels given in the dictionary entries can be explored to prefer accordingly labeled senses for domain specific texts.

Shallow parsing can be employed to explore syntactic patterns for disambiguation according to subcategorization patterns. This would also improve the formal description and recognition of verbal MWLs for languages with relatively free word order (German).

A crucial point, however, remains the availability and technical accessibility of wide–coverage, high–quality lexical ressources building the backbone of lexical knowledge for the system described.

7 References

If you would like to have COMPASS to read any of the German literature, please get in touch with us :-)

Bläsi, Christoph, Heinz-Detlev Koch (1991). Maschinelle Strukturerschließung von Wörterbuchartikeln mit Standardmethoden. In: *Lexicographica Vol. 7, (1991) 1992.* Tübingen: Niemeyer. 182-201.

Bogaards, P. (1990). Deux langues, quatre dictionnaires. In: Dolezal, F.M. et al. (eds.) *Lexicographica, 6.* Tübingen: Niemeyer.

Breidt, Elisabeth (1993). Extraction of V-N-Collocations from Text Corpora: A Feasibility Study for German. *Proceedings of the First ACL Workshop on Very Large Corpora, "Academic and Industrial Perspectives"*, Ohio State University, Columbus, June 1993, ACL.

Breidt, Elisabeth (1994). Terminology for Bilingual Dictionaries in Computational Lexicography. *COMPASS Deliverable 1*, LRE 62-080, University of Tübingen. Also as: *SfS-Report 04-95*, Seminar für Sprachwissenschaft, Universität Tübingen.

Breidt, Elisabeth (1995). Methods and Tools for the Conversion of Bilingual Dictionaries into COMPASS Dictionaries. *COMPASS Deliverable 10a G-E*, LRE 62-080, Universität Tübingen. February 1995.

Breidt, Elisabeth, Christine Thielen (1996). Conversion of bilingual dictionaries: from typesetting tape to dictionary database. *COMPASS Deliverable 23*, LRE 62-080, Universität Tübingen. March 1996.

Christ, Oliver (1993). The IMS Corpus Workbench: The XKWIC User Manual. Institut für maschinelle Sprachverarbeitung. Universität Stuttgart. Internal Report.

Christ, Oliver (1994). A modular and flexible architecture for an integrated corpus query system. *Proceedings of 3rd Conference on Computational Lexicography and Text Research, COMPLEX'94.* Budapest, July 1994. 23-32.

Church, Kenneth W., Patrick Hanks (1989). Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th ACL*, Vancouver. 76-83. Also in: *Computational Linguistics*, Vol. 16/1 (1990). 22-29.

Collins *German-English English-German Dictionary.* Edited by P. Terell, V. Schnorr, W. V. A. Morris, R. Breitsprecher. 2nd Edition. Glasgow: HarperCollins Publishers, 1991.

Cutting, Doug, Julian Kupiec, Jan Pedersen, Penelope Sibun (1992). A practical part–of–speech tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy.

Engelien, Brigitte, Tim Johnson (1991). A study of the market for LT products in Europe. OVUM report, March 1991.

Feldweg, Helmut (1995). Implementation and Evaluation of a German HMM for POS Disambiguation. *Proceedings of the ACL SIGDAT Workshop "From Texts to Tags: Issues in Multilingual Language"*, University College, Belfield, Dublin, March 1995.

Feldweg, Helmut, Michael Hipp (1993). NGRAM – Ein Werkzeug zur statistischen Analyse von Wortfolgen. Seminar für Sprachwissenschaft, Universität Tübingen. Internal paper.

Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery.* Boston/-Dordrecht: Kluwer.

Hauser, Ralf (1993). LexParse User Manual. *SfS-Report 10-93.* Seminar für Sprachwissenschaft, Universität Tübingen.

Hauser, Ralf, Angelika Storrer (1994). Dictionary Entry Parsing Using the LexParse System. In: *Lexicographica, Vol. 9, (1993) 1994.* Tübingen: Niemeyer. 174-219.

Hausser, Roland (1996), ed. Linguistische Verifikation. Dokumentation zur Ersten Morpholympics. Tübingen: Niemeyer.

Ide, Nancy, Jean Véronis (1995). Encoding Dictionaries. Computers and the Humanities, Vol. 29/1–3. Also in: Ide, Véronis (eds.) (1995). Text encoding initiative: background and contexts. Dordrecht: Kluwer. 167-180.

Karttunen Lauri, Kenneth Beesley (1992). Two-level rule compiler. *Technical Report ISTL-92-2*, Xerox Palo Alto Research Center, California.

Karttunen Lauri, Todd Yampol (1993). Interactive Finite State Calculus, *Technical Report ISTL-NLTT-1993-04-01*, Xerox Palo Alto Research Center, California.

Kupiec, Julien, Mike Wilkens (1994). The DDS Tagger Guide. Version 1.1. Internal paper, Xerox Palo Alto Research Center.

Longman Dictionary of English Idioms. Edited by Thomas Hill Long, Della Summers. London: Longman Group, 1979.

Martin, Willy (1993). Wordtranslator: a semi-automatic lexical system for the translation of English texts. *Toegepaste Taalwetenschap in Artikelen, 45.* 102-113.

Maurel, Denis (1989). Reconnaissance de séquences de mots par automate. Thèse de doctorat en informatique, LADL, Université Paris 7, France.

Neff, Mary S., Branimir Boguraev (1990). From Machine-Readable Dictionaries to Lexical Data Bases. *Lexical Systems Project Report*, IBM Thomas J. Watson Research Center, Yorktown Heights, NY.

Oxford-Hachette *French Dictionary*, Oxford University Press, 1994.

Schulze, Bruno M., Oliver Christ (1994). The IMS Corpus Workbench: The CQP User's Manual. Institut für maschinelle Sprachverarbeitung. Universität Stuttgart. Internal Report.

Segond Frédérique, Pasi Tapanainen (1995). Using a finite state based formalism to identify and generate multiword expressions. *Technical Report MLTT-019* Rank Xerox Research Centre, Grenoble, July 1995.

Segond, Frédérique, Elisabeth Breidt (1995). IDAREX: Description formelle des expressions à mots multiples en français et en allemand dans le cadre de la technologie des états finis. *"Lexicomatique et Dictionnairiques"*, Actes des IVe Journ'ees Scientifiques du reseau "Lexicologie, Terminologie, Traduction" de l'UREF, Lyon, Septembre 1995.
Also as: IDAREX: Formal Description of German and French Multi–Word Expressions with Finite State Technology. Internal paper. 1995.

Segond, Frédérique, Elisabeth Breidt, Giuseppe Valetto (1996). IDAREX: Formal Description of Multi–Word Lexemes with Regular Expressions. Submitted to COLING 96, Copenhagen, Denmark.

Silberztein Max (1993). *Dictionnaires électroniques et analyse automatique de textes – Le système INTEX*. Paris: Masson.

Sperberg-McQueen, C. M., Lou Burnard (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange.* TEI P3. 2 Vol. Chicago, Oxford: Text Encoding Initiative.

Storrer, Angelika (1996). Metalexikographische Methoden in der Computerlexikographie. In: Wiegand, H. E. (ed.). *Wörterbücher in der Diskussion II.* Tübingen: Niemeyer, to appear. 239-255.

Tapanainen Pasi (1994). RXRC Finite State Rule Compiler, *Technical Report MLTT-020*, Rank Xerox Research Centre, Grenoble, France.

Wiegand, Herbert Ernst (1991). Printed Dictionaries and their Parts as Text. An Overview of More Recent Research as an Introduction. In: *Lexicographica, Vol. 6, (1990) 1991.* Tübingen: Niemeyer. 1-124.

Wolski, Werner (1991). Formen der Textverdichtung im zweisprachigen Wörterbuch. In: Hausmann, F. J. et al. (eds.). *Dictionaries: An International Handbook of Lexicography.* Part 3. HSK 5.3. Berlin: De Gruyter. 2837-2853.