

Mapping German Word Senses onto a Core Ontology to Create a Multilingual Language Engineering Resource

Richard F. E. Sutcliffe^{*}¹, Oliver Christ[†], Annette McElligott^{*}

Christine Stöckert[†], Ulrich Heid[†], Helmut Feldweg[‡]

University of Limerick^{*}, University of Stuttgart[†], University of Tübingen[‡]

Abstract

A study is presented in which experimental subjects map German word senses onto an English ontology by completing questionnaires over the World Wide Web. The work is intended to establish whether such an approach could be used to create a multilingual concept ontology. While the methods need further refinement, the results of this study, taken together with those of two previous ones, suggest that the method is viable.

Introduction

hyphenationuni-stuttgart uni-tuebingen In order to build language engineering systems for performing tasks such as text retrieval, machine assisted translation or routing, it is necessary to have access to lexical data. Certain types of application require semantic information to be available so that the meaning of textual components can be used in processing. One example is conceptual information retrieval which attempts to understand the meanings of textual components with a view to extracting information which is salient to user queries. SIFT (Selecting Information From Text) was such a system and aimed to help users extract answers to their queries from a software instruction manual (Sutcliffe, Boersma, Ferris, Hyland, Koch, Masereeuw, McElligott, O'Sullivan, Relihan, Serail, Schmidt, Sheahan, Slater, Visser and Vossen, 1995). Underlying SIFT was a concept ontology derived in part from the public domain ontology WordNet (Beckwith, Fellbaum, Gross and Miller, 1992) and extended to cover the technical terminology of the word processing domain (O'Sullivan, McElligott and Sutcliffe, 1995).

¹Address for correspondence: Department of Computer Science and Information Systems, University of Limerick, Limerick, Ireland. Tel: +353 61 202706, Fax: +353 61 330876. Emails: richard.sutcliffe@ul.ie, oli@ims.uni-stuttgart.de, annette.mcelligott@ul.ie, chris@ims.uni-stuttgart.de, uli@ims.uni-stuttgart.de, helmut.feldweg@uni-tuebingen.de (the emails contain hyphens exactly as shown).

WordNet is an extremely useful resource but it is limited to American English. In order to develop multilingual systems, equivalent data is needed for other languages. One approach to the creation of such data is to use a single ontology as the core and to make word senses in other languages point to nodes in that ontology. We have carried out two pilot studies to establish the efficacy of this approach, one working with Russian, the second with Irish. We report here on a third study using German. The essence of this work is that multilingual ontological data can be created automatically by harnessing the power of the World Wide Web (WWW). Rather than using a small team of expert lexicographers, a large group of volunteers is employed. Quality control is ensured not by the guaranteed experience of every contributor, but by statistical characteristics of the data produced.

Two other important projects in this area should be mentioned. The first is the EuroWordNet of Vossen (1996). The second is the GermaNet of Feldweg (1996). Their relationship to this work is described below.

What is WordNet?

WordNet is a public domain lexical database for American English based on a concept ontology of around 70,000 semantic senses, or *synsets*. Each synset consists of a set of word senses which are deemed to be synonymous. A word sense comprises a word spelling and a sense number. An example synset is shown below:

```
file2, data file1 -- (a set of related
records kept together)
```

This synset comprises two word senses, 'file' sense 2 and 'data file' sense 1. The text in round brackets is an English definition or 'gloss' capturing the intended meaning of the synset. There are four types of synset, one each for nouns, verbs, adjectives and adverbs.

Ontological relations in WordNet are defined between synsets. These include antonyms, hypernyms, hyponyms, holonyms, meronyms, attributes, values, co-ordinate terms, entailments and things caused.

A graphical interface is provided with the system which allows parts of the ontology to be viewed. A

programming interface via procedure calls in C is also provided. Version 1.5 of the system contains 70,100 synsets made up of 95,600 word forms.

What is IWN?

WordNet has developed into a very comprehensive ontological lexical database but it is restricted to American English. There is a need for comparable linguistic resources in other languages. At the same time it is desirable for word senses in different languages to be linked. For these reasons we have been investigating whether a multilingual version of WordNet can be developed. There are several possible approaches to the creation of such a system which may be summarised as follows:

- Use the WordNet approach but commence from first principles working in a different language. One very interesting example is GermaNet which is part of the SLD Project (Feldweg, 1996).
- Create concept ontologies from monolingual machine readable dictionaries in different languages and then merge them. This is the approach of Vossen (1996).
- Take an existing monolingual ontology and extend it by making word senses in different languages 'point' to its nodes.

We have decided to experiment with the third alternative, using WordNet as the core ontology. This approach has a number of advantages:

- WordNet is publicly available for research and commercial use. This means that any multilingual extensions which we build can also be made publicly available.
- WordNet has been developed and refined over a number of years. This means that its strengths and weaknesses are well understood.
- The task of creating the multilingual version is fairly constrained and the method is not limited in advance to a fixed set of languages, as is the case with the other approaches mentioned.

Two small demonstration systems have been built and these can be viewed over the WWW (IWN-1, 1995; IWN-2, 1996). Sample output can also be seen in Figure 1.

The key epistemological limitation of an approach based on single monolingual ontology is that there may be concepts in different languages which simply can not be captured by reference to American English concepts. While this is undoubtedly true, many concepts are clearly shared. Until further progress has been made in developing multilingual resources based on both single and multiple ontologies, it will not be possible to determine the extent to which this issue has a material impact on the usefulness of any resources developed.

Strategies for Creating IWN

Having decided on a multilingual extension to WordNet, three possible mapping paradigms were identified for creating the system:

- **Paradigm A:** A word sense is chosen from the source language and the task is to link it to the nearest WordNet synset. For example, if the word is RU.OFIS (OFIS in Russian) then the closest synset might be *office -- (where professional or clerical duties are performed)*.
- **Paradigm B:** A WordNet synset is chosen and the task is to identify the nearest word sense in a given source language. For example, if the synset is *object, inanimate object, physical object -- (a nonliving entity)* and the source language is Russian, RU.OBJEKT might be selected.
- **Paradigm C:** A combination of Paradigm A and Paradigm B.

Whichever paradigm is used, the problem of sense identification has to be addressed. For example, which semantic sense of RU.OFIS is being linked to the office synset? Traditional approaches to the specification of semantic sense use dictionary definitions. However, these suffer from a number of limitations. An alternative method is to specify for each word spelling and sense number a set of example sentences derived from corpora which illustrate the use of the word spelling in the semantic sense intended. This is the approach we have decided to adopt. Indeed, WordNet itself is now to a large extent defined in this manner since it has been used to tag a significant proportion of the Brown Corpus. Thus the semantic sense intended to be conveyed by a particular synset can be taken to be that implied by the word usage in the set of Brown Corpus sentences to which it is linked.

One way in which source word senses can be linked to WordNet synsets is by the use of large populations of volunteer experimental subjects. We have been investigating this approach and two case studies have already been carried out. These are described briefly in the next section.

Previous Mapping Studies

Study 1: Paradigm A

The aim of the first study was to investigate the potential of Paradigm A. The source language was Russian. The task was to map Russian word senses onto WordNet synsets (Sutcliffe, O'Sullivan, Polikarpov, Kuzmin, McElligott and Véronis, 1996). During the mapping task, subjects were shown two sentences containing a Russian word, together with a small portion of the WordNet ontology. The task was to identify the synset which most closely captured the meaning of the Russian word. 10 nouns and five verbs were used. 24 volunteers completed the task, 20 at Smolensk State Pedagogic Institute and 4 at Lomonosov Moscow State

University. The work was carried out by sending a series of electronic mail messages which were completed and returned by respondents.

The main finding of the study was that the subjects could carry out the task despite knowing nothing about ontologies. Subjects also concurred to a large extent in their findings.

Study 2: Paradigm B

The objective of the second study was to establish the viability of using Paradigm B. The source language was Irish. The essence of the task was thus to map WordNet synsets onto Irish word senses. This was accomplished in a two-phase study, carried out by 12 volunteer subjects at Limerick (Sutcliffe, O'Sullivan, McElligott and Ó Néill, 1996). In the first phase, pairs of English sentences with one word highlighted in each were presented. The task was to write down an Irish word which had the same meaning. In the second phase, a set of Irish sentences containing the word written down by the subject was presented. The task was to identify the sentences in which the word was being used in the sense intended in phase one. Ten nouns and five verbs were used for the experiment, chosen from the Irish Constitution (*Bunreacht na hÉireann*, 1990) which is a parallel text in Irish and English. Data was elicited from subjects using printed questionnaires.

Once again, the subjects carried out the task successfully and concurred to a large extent in their judgments. On the other hand, judging semantic senses from examples was found to be difficult and results for verbs were less clear-cut than those for nouns.

German Mapping Study

Aims

Following on from the two studies presented in the previous section, it was decided to carry out a third study. The main characteristics of the study were to be as follows:

- A larger set of volunteers was to take part;
- Data was to be gathered from subjects by the use of forms completed over the World Wide Web (WWW);
- Word senses were to be defined using more realistic corpus data;
- Further information was to be gleaned regarding the efficacy of using statistical analysis as the basis for determining the quality of data produced.

Method

The following steps were carried out:

1. Two lists of words were produced, based on a frequency study carried out on a portion of the British National Corpus (Kilgarriff, 1995). One contained the 40 most frequently occurring nouns while the

other contained the 20 most frequently occurring verbs.

2. The ESCORT system from Princeton (ESCORT, 1993) was used to determine the most frequently occurring semantic sense relative to WordNet v1.4 of each of the words in the two lists, working with 103 semantically tagged files from the Brown Corpus. Where a word did not occur in the appropriate part-of-speech in this part of the corpus, it was discarded.
3. Four example sentences were extracted from the Brown Corpus for each word processed in Step 2. Wherever possible, these were extracted from four different corpus files (and thus were derived from different articles).
4. Two native German speakers who were also fluent at English produced for each English word between one and six possible translations into German, taking into account all possible semantic senses of the word. In other words, the translations chosen were not restricted to the semantic senses of Step 3.
5. For each German word produced in Step 4, 50 sentences containing it were extracted from the Frankfurter Rundschau component of the European Corpus Initiative's Multilingual Corpus I (ECI/MCI) using the IMS Corpus Workbench.
6. 13 English nouns and 7 English verbs were selected from the lists of Step 1 for use in the study (see Table 1).
7. Software was developed to allow a series of Hyper-Text Markup Language (HTML) (1996) forms to be presented to a user. Each form dealt with one English word sense. At the top were listed the four Brown Corpus sentences illustrating its use, as described in Step 2. In each sentence the selected word was highlighted. Under this were listed the corresponding 20 German sentences produced in Step 5. In each one the 'translation' corresponding to the English word sense was highlighted. Alongside each sentence was a check box. The task of the user was to check the box alongside each German sentence which was deemed to show the German 'translation' in a semantic sense corresponding to an accurate translation of the original English word. If no German sense applied, then no check box was to be checked. Having checked all appropriate senses, the user had to press a button which caused the completed question to be sent to the server which then sent the next question in the sequence. Each user completed 20 questions, one for each word selected in Step 6. However, the order in which the questions were asked was random, as were the orders of both the 4 English sentences and the 20 German sentences within each question. Figure 2 shows how each question looks to the user.

No.	Word	POS
1	time	nou
2	world	nou
3	way	nou
4	government	nou
5	life	nou
6	group	nou
7	company	nou
8	development	nou
9	information	nou
10	children	nou
11	women	nou
12	case	nou
13	family	nou
14	want	vrb
15	need	vrb
16	think	vrb
17	know	vrb
18	include	vrb
19	believe	vrb
20	find	vrb

Table 1: Words used in the study

8. A set of volunteer subjects was enlisted from five institutions: University of California at Berkeley International Computer Science Institute, University of Hamburg, University of Limerick, University of Stuttgart Institute for Natural Language Processing and University of Tübingen. All subjects were native speakers of German and fluent speakers of English. They were not paid for their participation in the study.
9. Each volunteer subject was assigned a unique username and password and then used these to complete their questionnaire comprising 20 sense identification questions, working over the WWW.
10. The resulting data was analysed.

Results

Eleven persons completed the study, of whom three were collaborators on the project. The results are summarised in Table 2. Each line of the table deals with one of the twenty questions. Each column refers to one of the twenty German sentences associated with that question. Each figure in the table is the percentage of the subjects who selected that German sentence for that question. Thus a value of 100 means that all the subjects chose a particular sentence for a question, while a value of 0 means that none did.

Where all the subjects in the population have selected a German sentence, we can be very confident that the highlighted German word in that sentence is being used in a semantic sense which is close to that of the corresponding English word sense. Where no subjects have selected a sentence, we can be equally

confident that the semantic sense of the German word is *not* related to the English word sense. Intermediate values indicate partial confidence.

The main conclusions we can draw from this table are as follows. 22% of all German sentences were unanimously judged as relevant to their corresponding English word senses and 9% were unanimously judged not relevant. Thus, all subjects made the same judgement about 31% of the sentences over all.

Secondly, a judgement which is near acceptance or rejection (e.g. above 80% or below 20%) may in fact prove to be reliable. 50% of sentences were judged correct by more than 80% of the subjects with 20% judged correct by less than 20% of subjects. If these thresholds proved acceptable (this is not proved in the present study) then reliable judgements could be made regarding 70 sentences.

Thirdly, only two lines in the table out of twenty contain no 100%. This suggests that twenty sentences containing a particular word spelling gives a good chance of an appropriate sense occurring.

As well as filling in the questionnaires, some subjects provided useful feedback on the task as a whole. The main points were as follows:

- Twenty German sentences for each question is too many and imposes an excessive burden on the subject.
- The network connection has to operate at a reasonable speed for the task to be achievable.
- The instructions should have stressed that subjects were to work quickly and that number of correct sentences per question was not balanced in advance by the experimenters.
- The issue of compounds in German was not properly thought out. Sometimes a compound was the highlighted word in a sentence because it contained the stem of a German word which was a possible translation of the English word. In such a case, the stem itself might or might not be in the correct sense for the sentences to be clicked, but in either case the compound as a whole was not synonymous with the English word. The instructions should have specified what action to take in this situation.
- If each question were to fit completely on the screen, completion of the questionnaire would be much faster. This is because no scrolling would be needed and in addition it would be possible to refer back to the four English sentences easily. This suggests that less German sentences should be used in each question.
- Finally, not all users have access to up-to-date versions of browsers. It was discovered at the last minute that all the subjects at Limerick were using Version 1 of Netscape and could not therefore read the questionnaire because it uses frames. A basic version of the questionnaire in addition to the so-

pshisticated one should be available for future studies in order to get around this problem.

Discussion

The population of subjects used in this study was small and thus any findings would need to be verified using a bigger sample of volunteers. However, the results suggest that the subjects understood the task and were able to carry it out. The fact that 31% of all sentences were unanimously judged either relevant or not relevant implies that the subjects were not responding randomly.

The following are the most important next steps. Firstly, we wish to carry out the same study using more subjects. Secondly, a method for establishing the correctness of the results produced by a given population must be devised. The problem in essence is that while all subjects might agree on a judgement, that judgement may still be wrong. The present experiment has not addressed this issue. One possible approach is as follows. Two populations of subject can be asked to complete a questionnaire. The first population would consist of expert linguists who had made a careful study of the experiment and understood the purpose of the study before answering the questions. The second population would consist of naive volunteers who were not linguists. The responses of the experts could be considered as 'correct' while those of the amateurs were 'possibly correct'. Having computed the analysis tables for each population, these could then be compared with each other. If the two were highly correlated it would suggest that the amateurs' responses were 'correct'. Such a result would then validate the use of amateurs in a real data-gathering exercise.

Thirdly, we wish to determine the size of the population which is necessary to undertake a particular question before the results are reliable. For example, if ten out of ten people all click a German sentence is that enough, or must 100 out of 100 click it before we can trust the data?

Finally, regarding the ontological implications of the present work, we can be fairly optimistic about the viability of the overall approach. The data in this study was significantly more natural than those used in the preceding ones for Russian and Irish. Firstly, the words used varied in concreteness – an attribute which our previous work has suggested tends to affect the viability of the ontology as a data structure for language engineering. Concrete terms are much easier to classify taxonomically than nebulous ones. Secondly, sentences for both English and German were derived from a large number of different documents discussing different topics. Thirdly, subjects had only the written instructions to go on because there was no opportunity for oral clarification. The three factors together suggest that a high proportion of the work involved in carrying out a large scale study could in fact be automated.

We are currently investigating some of the above issues in further studies.

Acknowledgements

We are indebted to the National Software Directorate of Ireland for supporting this research. This work could not have been done without the assistance of Denis Hickey, Tony Molloy and Redmond O'Brien. Donie O'Sullivan of Broadcom Éireann Research and Detmar Meurers of University of Tübingen both made very helpful suggestions relating to the form and structure of the questionnaire.

References

- Beckwith, R., Fellbaum, C., Gross, D., & Miller, G. A. (1992). WordNet: A Lexical Database Organised on Psycholinguistic Principles. In U. Zernik (Ed.) *Using On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bunreacht na hÉireann, (1990). *Bunreacht na hÉireann*. Dublin, Ireland: Government Publications Stationary Office.
- ESCORT (1993). Examine Semantic Concordance Of wRitten Text. Version for WordNet 1.5 is available by ftp from [princeton.clarity.edu](http://princeton.clarity.edu/pub/wordnet) in directory [pub/wordnet](http://princeton.clarity.edu/pub/wordnet) at file [wn1.5semcor.tar.gz](http://princeton.clarity.edu/pub/wordnet).
- EuroWordNet (1996). Building a multilingual database with wordnets for several European languages. <http://www.let.uva.nl/ewn/EuroWordNet.html>
- Feldweg, H. (1996). Information on SLD Project and GermaNet. <http://www.sfs.nphil.uni-tuebingen.de/lsd/english.html>
- HTML (1996). A Beginner's Guide to HTML, Version 2.0, April 1996. <http://www.ncsa.uiuc.edu/General/Internet/WWW/HTMLPrimer.html>
- IWN-1 (1995). First IWN Demonstrator in five languages. http://nlp01.cs.ul.ie/iwn_main_nlp_demo.html
- IWN-2 (1996). Second IWN Demonstrator in eight languages. http://nlp01.cs.ul.ie/nlpd2/iwn_main_nlp_demo_frame.html
- Kilgariff, A. (1995). *British National Corpus Frequency Lists* (a database produced from a sample of the BNC). Obtainable via ftp from ftp.itri.bton.ac.uk in directory [pub/bnc](http://ftp.itri.bton.ac.uk). (See file README.)
- Sutcliffe, R. F. E., Boersma, P., Bon, A., Donker, T., Ferris, M. C., Hellwig, P., Hyland, P., Koch, H.-D., Masereeuw, P., McElligott, A., O'Sullivan, D., Relihan, L., Serail, I., Schmidt, I., Sheahan, L., Slater, B., Visser, H., & Vossen, P. (1995). Beyond Keywords: Accurate Retrieval from Full Text Documents. *Proceedings of the 2nd Language Engineering*

- Convention, Queen Elizabeth II Conference Centre, London, UK, 16-18 October 1995.*
- Sutcliffe, R. F. E., O'Sullivan, D., McElligott, A., & Ó Néill, G. (1996). Irish-English Mappings in International WordNet: A Pilot Study. In *Proceedings of the International Translation Studies Conference, Dublin City University, 9-11 May, 1996.*
- Sutcliffe, R. F. E., O'Sullivan, D., Polikarpov, A. A., Kuzmin, L. A., McElligott, A., & Véronis, J. (1996). IWNRR - Extending A Public Multilingual Taxonomy to Russian. In *Proceedings of the Workshop Multilinguality in the Lexicon, AISB Second Tutorial and Workshop Series, University of Sussex, Brighton, UK, 31 March - 2 April 1996.*
- Vossen, P. (1996). *Proceedings of 7th International Euralex Congress, Gothenburg, 1996, 715-728.*

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s18	s19	s20
1	0	0	64	9	9	18	0	55	9	0	0	9	9	0	18	45	9	9	36	0
2	0	0	0	0	9	9	73	18	82	9	0	0	100	100	73	27	91	9	18	9
3	91	100	100	91	100	91	91	91	91	100	82	91	73	91	91	91	18	100	18	82
4	27	100	100	55	18	100	100	100	100	45	36	36	36	55	100	100	55	91	100	27
5	45	45	27	36	0	100	27	91	91	55	91	91	82	100	73	82	82	91	73	82
6	73	82	82	82	82	82	82	73	82	91	100	100	36	9	9	82	91	27	100	100
7	27	100	27	27	36	27	100	100	36	100	100	0	100	100	100	91	91	18	64	0
8	82	91	91	27	82	36	82	100	64	82	73	45	36	82	82	91	9	9	73	36
9	82	100	100	100	91	100	73	91	73	91	100	73	73	100	91	100	100	91	82	100
10	100	100	100	100	100	100	100	100	100	64	100	100	100	64	55	100	91	100	100	91
11	73	82	82	82	82	91	91	64	100	82	100	91	91	82	82	82	91	91	91	82
12	100	100	100	100	55	73	100	100	73	100	73	100	0	82	0	82	100	55	55	73
13	36	45	91	45	36	27	36	91	36	45	36	82	36	36	45	36	27	36	36	100
14	55	55	91	100	91	82	91	91	100	73	91	100	100	91	0	73	9	18	100	91
15	100	91	91	91	91	100	100	100	91	91	91	100	82	91	91	55	100	73	91	55
16	55	55	36	0	0	27	27	27	45	0	0	45	64	0	91	64	55	9	45	0
17	18	82	100	100	100	27	100	100	9	73	100	36	100	73	73	91	100	9	55	9
18	82	82	27	82	9	82	91	9	27	100	73	45	73	27	91	55	82	0	0	91
19	27	0	82	0	0	0	9	0	0	36	18	9	0	9	64	9	91	9	100	82
20	91	55	9	91	36	64	73	9	45	9	0	27	9	9	36	9	100	91	27	55

Table 2: Results of the study

```

en.university, ir.ollscoil, jp.daigaku
[ agus ansin ta/ an ollscoil ann agus cola/isti/ ann ]

en.establishment, fr.e/tablissement, gr.Stiftung, ir.bunu/, jp.setsuritsu
[ ta/ oifigi/ nua dha/ bhunu/ ar an tsra/id in ]

en.structure, en.construction, fr.structure, gr.Konstruktion,
ir.struchtú/r, jp.koozoo
[ an stoirm bhi/ an struchtú/ir timpeall na ha/ite mar ]

en.artifact, en.artefact, fr.objet fabrique/, gr.Artefakt, jp.kakoohin

en.object, en.inanimate object, en.physical object, fr.objet,
fr.chose, gr.Ding, gr.Gegenstand, ir.rud, jp.mono, jp.taishoo
[ balla no/ cruach no/ rud ar bith mar sin ]
[ ball no/ cruach no/ rud ar bith mar sin ]

en.entity, fr.entite/, gr.Wesen, ir.sla/naonad, jp.jittai
[ gach ceann de na slanaonadai/ le che/ chun an ]

```

Figure 1: Example output from IWN Demonstrator

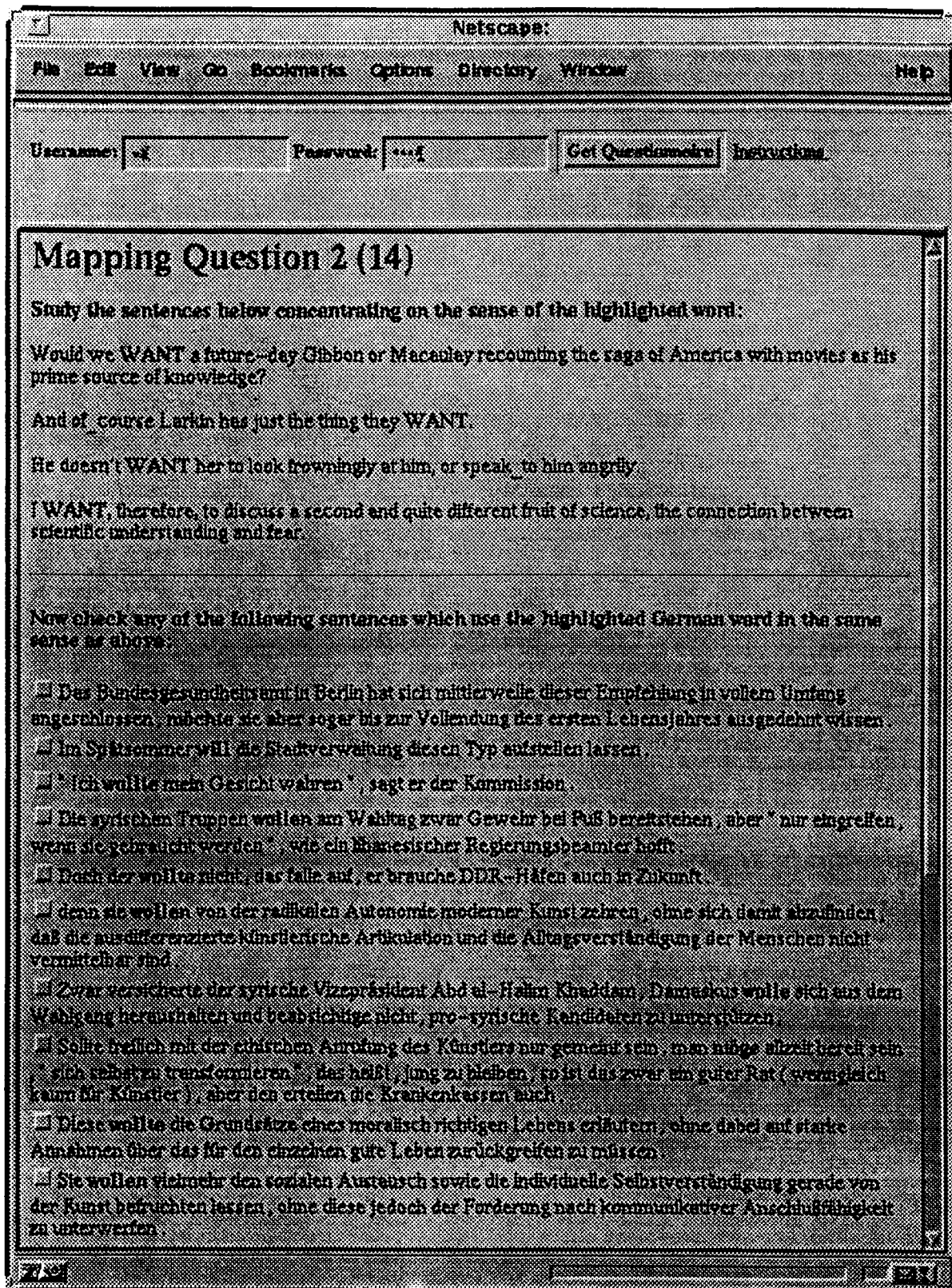


Figure 2: Part of an example question as shown to the user