

A false friends exercise with authentic material retrieved from a corpus

Joachim WAGNER

National Centre of Language Technology
School of Computing, Dublin City University
Dublin 9, Ireland
jwagner@computing.dcu.ie

Abstract

This paper presents a CALL exercise that aims to raise the learner's awareness of false friends. In the exercise, the learner is asked to mark words in a text that are similar in form to a word in his or her native language and then to classify these words according to three levels of meaning correspondence. Text is randomly selected from a corpus and integrated into the exercise. A preliminary evaluation shows that mature students understand the exercise well.

1 Introduction

False friends are words or phrases of two languages that look or sound alike but do not convey the same meaning. They can present difficulties to foreign language learners and translators. The false friend exercise presented here aims at raising the awareness of false friends, as opposed to drawing attention to only those notorious word associations that are well known to cause trouble to many students of the L2. This is a difference to false friend study books, for example (Humphrey 2003 and Dretzke 2001), that address the latter goal.

A few notes on terminology: we use the term “cognate” to refer to orthographic or phonological cognates, although it is used strictly etymologically in linguistics (Warnow 1997). True cognates and false friends are cognates that are either correct translation of each other or not. Of course, this definition is problematic because correctness of translation depends on context. From the point of view of language learning and teaching, false friends are word associations between the language to be learned and another language (often the native language) that are misleading, cause errors and thus not desired.

Section 2 describes what the exercise is about and how it has been implemented. In section 3 we give results from a preliminary evaluation. Section 4 focuses on current work and future plans. Section 5 concludes our findings so far.

2 The False Friend Exercise

This section describes our web-based false friend exercise for German L2. The user interface is in English. However, the L1 of the learner can be any language that has many cognates and false friends with German and can be transcribed in the ISO Latin 1 character set. Of course, the learner must be proficient in English in order to understand the instructions.

2.1 Description of the Exercise

The exercise contains of three steps. First, the learner is asked to identify cognates in an authentic text. The number of words that are available for selection is kept small with help of part of speech (POS) tagging and a stop word list. The learner is also encouraged to annotate his or her associations to the words. In the second step, the learner classifies the identified words into three classes of meaning correspondence: a. similar meaning, b. different meaning, and c. partly different meaning. Finally, the learner reviews his or her findings. Questions about the dominant class of meaning correspondence are intended to draw the learner's attention to the high number of false friends. The exercise is preceded by an introduction explaining its goals, the three steps of the exercise and the implication of the difficulty level of the authentic text. Figure 1 on the next page shows a screenshot of the exercise with colours replaced by dotted and solid lines for better reproduction.

2.2 German Corpus of Book Excerpts

Authentic text presented to the learner is very important in our exercise. If the texts were edited, the learner might doubt that the observed density of cognates is natural. Since we also want to evaluate automatic selection of suitable texts from a corpus, we chose a collection of book excerpts downloaded from a book review website¹. The book excerpts have been automatically extracted from the reviews and manually checked that they do not contain text written by the reviewer. (In retrospect, manual copy and paste would have been faster al-

¹ <http://www.leser-service.de/>

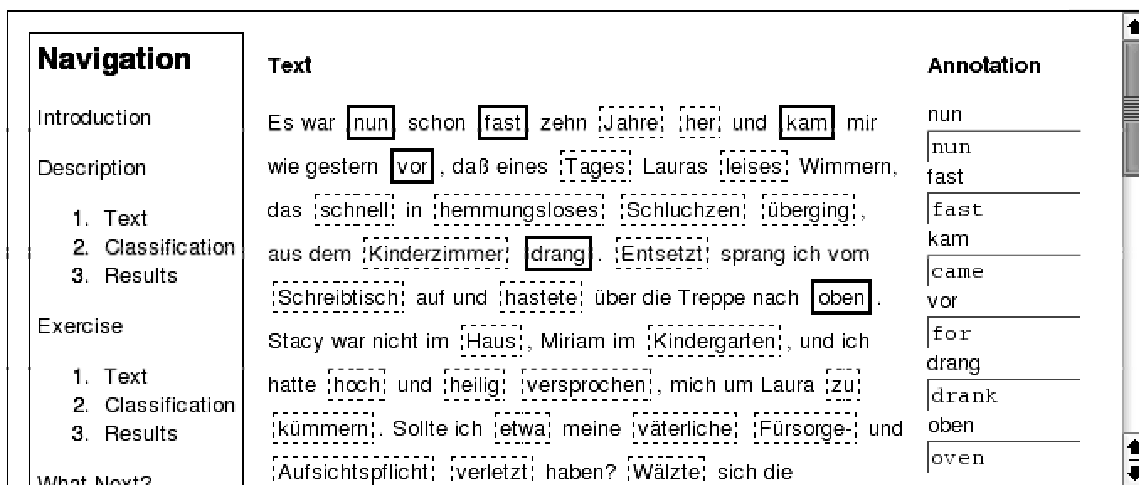


Figure 1: False Friends Exercise – Step 1: Identification of Cognates

beit very monotonous.) In the exercise, we use only the first 180 up to 300 words of each excerpt. In addition, some excerpts turned out to be too short to be used in the exercise. Table 1 summarises quantitative information about our corpus.

Processing Stage	Quantity	Words
downloaded files	331	342,697
extracted excerpts	223	64,356
texts used in exercise	197	37,500

Table 1: A Corpus of Book Excerpts

2.3 Technology Used

All texts have been POS tagged and lemmatised with the IMS TreeTagger (Schmidt 1994). POS tags are used to restrict the words that can be selected by the learner in the identification task to open class words. This considerably reduces the amount of work demanded from the learner.

The user interface is implemented in JavaScript and dynamic HTML. The client-side script generates input fields required to annotate the words identified by the learner as cognates, fills a classification table with these words and maintains a summary panel. The learner can switch instantly between the panels representing the three steps of the exercise and the introduction, since the web server is not contacted. The JavaScript has been tested with recent versions of Internet Explorer, Netscape Navigator, Mozilla, Opera and KDE Konqueror web browsers. Konqueror showed abnormal behaviour of dynamically generated radio button sets whenever the user changed a classification. A note for users of this browser was included in the introduction that changing a decision does not work.

On the server side, a text from the corpus is integrated into the HTML page. In addition, code is inserted to facilitate the evaluation. We use CGI scripts written in the programming language Python² for this. However, the exercise does not depend on CGI support. The HTML pages containing the exercise with a specific text can be generated in advance and served as a static HTML page or placed on a filesystem, for example on CD-ROM.

3 Preliminary Evaluation

We conducted a first evaluation with six subjects that learned German as L2. Four subjects are English native speakers, one is Russian, one Dutch. All subjects are postgraduate research students at Dublin City University. The evaluation started with a 20-minute aural introduction that followed the text of the written introduction within the exercise and explained that it is followed by a questionnaire. After remaining questions had been answered and the local URL of the exercise been provided, the subjects did the exercise and completed the questionnaire. During the evaluation the author checked for technical problems.

3.1 Browser Problems

All subjects could do the first step of the exercise. The annotation input field appeared and could be filled and the classification table was generated. Two subjects could not use the radio buttons to do the classification task. Three subjects did not suffer browser problems but the results send by them to the evaluation script did not contain the first word selected for each class of meaning correspondence. For the purpose of the evaluation, this is not

²pronounced like in "Monty Python's Flying Circus", not like the animal, see <http://www.python.org/>

a problem. However, if a teacher or tutor is expected to give feedback to the learner, incomplete submission of results is unsatisfactory.

3.2 Increase of Awareness

Three questions of the questionnaire deal with the subject's impression of the classification task. The idea was that awareness of difficulties in this task indicates awareness of false friends. The answers varied a lot and did not correlate. Three of six subjects indicated with at least one answer that the classification task was difficult. In contrast, all subjects responded to general questions about the exercise that it was (very) straightforward and not too long. Three subjects seem to be aware of difficulties. However, even if we could conclude that the subjects are aware of false friends, we could not conclude that the awareness has been raised by the exercise. A fourth question asked the subjects whether they have been surprised by the number of cognates and, if yes, what they expected. Two subjects answered "No", three "A little" and one "Yes". (Another option was "Don't know".) Those who answered the follow up question said that they expected fewer cognates. Strangely, one subject who answered "No" is among them. The amount of surprise is lower than expected.

3.3 Expectation of Feedback

Subjects have been asked if they missed feedback on errors. Three subjects responded "A little", one "Yes" and two "Don't know". ("No" was also available.)

3.4 Impact of the Text

The authentic texts contain very difficult vocabulary and grammatical structures. We assumed that this would not cause problems because the learner focuses on surface similarity of single words. In order to verify this assumption, we asked the subjects how much they remember of the content of the text. "Nothing" and "Everything" have been answered twice each and "Parts" and "Words" once each. This seems to contradict our assumption. However, it should be mentioned that one of the subjects who answered "Everything" has a very high level of proficiency of German and the other subject has worked on the exercise for 35 minutes, while the subjects who answered "Nothing" spend only 14 and 8 minutes respectively.

3.5 Impact of Native Language

The English native speakers found between 13 and 26 cognates. The Russian speaker preferred to compare the German words to English words and identified 47 cognates. The Dutch native speaker chose his native language and found 84 cognates,

most of them being true cognates according to his classification.

4 Current and Future Work

At the moment, texts are selected at random and no feedback is provided. The preliminary evaluation showed that this is acceptable for this exercise. However, we would like to know how much the exercise can be improved by offering a choice between a range of automatically detected suitable text and providing intelligent feedback with the help of NLP technology.

4.1 Ranking of Texts

Texts used with this exercise should contain a high number of cognates. We currently develop ranking functions to aid selection of suitable texts. In a first step, we extracted cognates from word lists of both languages with a character based similarity measure. We set a similarity threshold and count cognates of each text according to the list of extracted cognates. With a low threshold, we got a density of cognates from 20% up to 48%. This wide range of density values is quite promising. However, we will have to check if the variance is statistically significant since the texts are very short. Secondly, we will evaluate whether the counts or density values are relevant, i.e. whether the ranking they provide helps to select suitable texts for the false friend exercise.

Next, we want to look at how lemmatisation, POS tagging and direct application of a similarity measure can improve detection of cognates. Bilingual dictionaries and corpora might be used to distinguish between true cognates and false friends. We want to compare the automatic detection and classification with manual annotation and cognate lists that can be found in student's workbooks of educational publishers like Dretzke and Nester 2001.

More long term, we want to identify additional features that text should or should not possess in order to be judged suitable for a specific exercise. For instance, repetitions or other striking stylistic devices can distract the learner's attention. Here, it might be fruitful to have a look at other research areas like information retrieval and genre classification and apply their methods to our task.

4.2 Adaptation to Other Languages

The ranking function and exercise have been implemented for German as a foreign language (DaF) and English native speakers. Since only low-level NLP technology is employed, the processing should be easily adaptable to other language pairs. This way, even for less prevalent languages that

are not covered in commercial training books a false friends exercise can be provided.

In our research group, we have experts on English, German, several Romance languages and Irish. Since these languages use the same script, we can re-use the similarity measure. Therefore, the number of languages and not the number of language pairs (L1, L2) determines the time needed for implementation.

4.3 Artificial Co-Learner

Four of the six subjects stated that they missed feedback in the exercise. Automatic feedback could be generated with help of existing modules and additional linguistic resources, for example a bilingual dictionary to verify classification of translation equivalence. However, these modules make errors like any other NLP technology. In language learning, wrong feedback can result in the learner acquiring wrong knowledge or being frustrated. We hope to overcome these problems by providing feedback in a different way with an artificial co-learner. See also (Greene et al. 2004).

In the false friend exercise, the artificial co-learner should explain its proposal and promote cognate identification strategies. The classification task is very challenging. A bilingual dictionary or an aligned corpus might give some indication. However, an informed classification needs knowledge about the meanings of the two words in a general context. A single translation equivalence between similar words does not constitute a true cognate and absence of such evidence does not always point to a false friend.

4.4 Improved Evaluation Methodology

The Questionnaire has proven to be insufficient to assess the increase of the subject's awareness of false friends. We will try to improve the set of questions. There is room for additional questions since all subjects completed the current questionnaire in less than five minutes. However, it seems more promising to develop, for example, a vocabulary test that contains around ten false friends not well known to the target group of language learners.

5 Conclusion

Even this very simple type of exercise brings up interesting questions in both NLP technology and pedagogy. The challenge of CALL is to combine these two areas. In order to adapt to learner needs, we have to improve our evaluation methodology. The false friend exercise uses NLP technology in order to simplify the first task for the learner. Three ideas to extend the exercise have been outlined.

6 Acknowledgements

Our thanks go to the subjects who spend up to 50 minutes to participate in the evaluation and to the author's fellow researchers of our ICALL group for many useful comments on the exercise and evaluation.

References

- B. Dretzke and M. I. Nester. 2001. *Student's Guide to false friends, old friends and new friends*. 3-464-037109-3 Cornelsen Verlag, Berlin
- C. Greene, K. Keogh, T. Koller, J. Wagner, M. Ward and J. Van Genabith. 2004. Using NLP Technology in CALL. *Proceedings of InSTIL / ICALL 2004*, Venice
- R. Humphrey. 2003. *False Friends, Falser Friends, Flasest Friends*. 3-12-939613-6 Ernst Klett Verlag, Stuttgart, Germany
- H. Schmidt. 1994. *Probabilistic Part-of-Speech Tagging using Decision Trees*. Revised online version, <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>
- T. Warnow. 1997. Mathematical approaches to comparative linguistics. *Proceedings of the National Academy of Sciences*, Vol. 94, pp. 6585-6590, online version used (PS with 19 pages)