http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources

Carolin Müller-Spitzer, *Institut für Deutsche Sprache, Germany (mueller-spitzer@ids-mannheim.de)*

María José Domínguez Vázquez, *Universidade de Santiago de Compostela, Spain (majo.dominguez@usc.es)*

Martina Nied Curcio, *Università degli Studi Roma Tre, Italy (martina.nied@uniroma3.it)*

Idalete Maria Silva Dias, *Universidade do Minho Braga, Portugal (idalete@ilch.uminho.pt)*

Sascha Wolfer, *Institut für Deutsche Sprache, Germany (wolfer@ids-mannheim.de)*

Abstract: In the past two decades, more and more dictionary usage studies have been published, but most of them deal with questions related to what users appreciate about dictionaries, which dictionaries they use and what type of information they need in specific situations — presupposing that users actually consult lexicographic resources. However, language teachers and lecturers in linguistics often have the impression that students do not use enough high-quality dictionaries in their everyday work. With this in mind, we launched an international cooperation project to collect empirical data to evaluate what it is that students actually do while attempting to solve language problems. To this end, we applied a new methodological setting: screen recording in conjunction with a thinking-aloud task. The collected empirical data offers a broad insight into what users really do while they attempt to solve language-related tasks online.

Keywords: DICTIONARY USE, OBSERVATIONAL STUDY, LANGUAGE LEARNERS, ONLINE RESOURCES, SEARCH STRATEGIES, ONLINE DICTIONARIES, AUTOMATIC TRANSLATORS

Opsomming: Akkurate hipoteses en noukeurige lees is noodsaaklik: Resultate van 'n waarnemingstudie uitgevoer op leerders wat aanlyn taalhulpbronne gebruik. In die afgelope twee dekades is al hoe meer woordeboekgebruikstudies gepubliseer, maar die meeste van hierdie studies handel oor vraagstukke wat verband hou met wat gebruikers van woordeboeke waardevol vind, watter woordeboeke hulle gebruik en watter tipe inligting hulle

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

in spesifieke situasies benodig — met die voorveronderstelling dat gebruikers inderdaad leksikografiese hulpbronne raadpleeg. Taalonderwysers en dosente in die linguistiek kry dikwels die indruk dat studente nie genoeg hoëkwaliteitwoordeboeke in hul daaglikse werk gebruik nie. Met hierdie siening in gedagte het ons 'n internasionale samewerkingsprojek van stapel gestuur om empiriese data te versamel om sodoende te kan evalueer wat dit is wat studente in werklikheid doen wanneer hulle taalprobleme probeer oplos. Om hierdie doel te bereik het ons gebruik gemaak van 'n nuwe metodologiese omgewing: skermopnames saam met 'n opdrag wat uitgevoer moet word terwyl daar hardop gedink word. Die versamelde empiriese data verskaf 'n breë insig in wat gebruikers werklik doen terwyl hulle poog om taalverwante take aanlyn op te los.

Sleutelwoorde: WOORDEBOEKGEBRUIK, WAARNEMINGSTUDIE, TAAL(AAN)LEERDERS, AANLYN HULPBRONNE, SOEKSTRATEGIEË, AANLYN WOORDEBOEKE, OUTOMATIESE VERTALERS

1. Introduction

Research into dictionary use has made substantial progress in the past two decades (cf., e.g., Töpel 2014, Welker 2013, Lew 2011; Lew 2015a), especially with regard to online dictionaries (cf., e.g., Müller-Spitzer 2014, Lew 2015b). However, almost all studies in the field deal with the aspects that users value when using dictionaries (e.g. Domínguez Vázquez et al. 2013, Domínguez Vázquez and Valcárcel Riveiro 2015, Müller-Spitzer and Koplenig 2014), which dictionaries or which items in dictionaries are used or required in which situations (e.g. Koplenig and Müller-Spitzer 2014, Nied Curcio 2013), which methods of presenting data are most user-friendly (e.g. Lew 2010, Lew et al. 2013), or which information is most frequently looked up in online dictionaries (e.g. De Schryver et al. 2006, Hult 2012, Koplenig at al. 2014). Therefore, these studies either assume that lexicographic tools are actually used or put the test subjects into concrete situations in which they are asked to imagine what lexicographic tools they would use. At the same time, many language teachers and lecturers in linguistics are under the impression that students do not use a sufficient amount of (good) dictionaries in their everyday life (see, e.g., Frankenberg-Garcia 2011). Accordingly, there is a gap between empirical research on dictionary use and the reality of learners' or students' actual everyday language challenges. We still have too little empirical data to be able to assess the role dictionaries play in day-to-day work. As Levy and Steel put it:

> The study reported here, with data drawn from a large-scale survey, reports on what students *say* they do when using electronic dictionaries. This reportage does not necessarily reflect what students actually *do* […]. Smaller-scale studies are needed to complement and enrich the findings of the present study. (Levy and Steel 2015: 194)

With this in mind, we launched an international cooperation project to collect empirical data with which to evaluate the suggested discrepancy. Our aim was

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

to collect comprehensive and reliable data about what it is that students (starting with German language learners from Romance language-speaking European countries) actually do when they deal with language problems. With the help of this accumulated knowledge about students' actual use of lexicographic resources, these data could then constitute an adequate starting point from which to teach students how to use language resources. Ignoring this aspect can be compared to teaching a language without asking at what level the students currently are.

To get a better idea about what students do during their everyday work, we used a new methodological setting for research into dictionary use: we presented sentences on a notebook computer and the participants were asked to improve these sentences using the online resources of their choice. During this process, we recorded the learners' on-screen actions with a screen recorder and prompted them to think aloud. We collected audio and screen capture data of 42 students from Braga (Portugal), Rome (Italy) and Santiago de Compostela (Spain). All participants were at the A2/B1 level according to the 'Common European Framework of Reference for Languages (CEFRL)'[1]. The collected data include 1,680 minutes of screen recordings and audio material containing more than 2,200 search procedures. The collected empirical data offers a broader insight into what language users today really do when solving language-related tasks. A wide range of questions can be addressed using the data, e.g.: Are our participants aware of the differences between translation systems and dictionaries? Do they adapt the search string to the type of resources used? Does the number or type of resources used have a positive impact on solving the task? How much time do they spend using the various resources? All these questions are addressed in this paper, which is structured as follows: first, we present the study design and our method for collecting the data (Section 2). Then, in the main part of our paper, we describe and explain the results of our study (Section 3). After some general results (Section 3.1), we focus on search strategies (Section 3.2) and on the factors that influence the quality of the corrections (Section 3.3), especially the influence of careful reading and how strongly overall search behavior was influenced by the initial hypotheses. Our article ends with conclusions (Section 4).[2]

2. Materials and method

We employed a mixed-methods design combining (i) a language correction task, (ii) screen recording of all on-screen actions, and (iii) audio recordings to create the participants' thinking-aloud protocols (Ericsson and Simon 1993). We distributed written instructions and a declaration of consent to the participants before the experiment. Both documents were in the participants' native languages. The instructions described the task and the setup on the computer screen. Also, we highlighted that the participants did not necessarily have to find a solution or correction for each and every stimulus sentence. The instruc-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

tions further contained some suggestions for the thinking-aloud task such as "describe what you are doing, why you are doing it, describe your thoughts while solving the task, describe why you are accessing a specific internet site, what you wish to find on the site, tell us why you are choosing a specific correction and whether you are satisfied with the corrections", and so on. Finally, the instructions indicated that the study would only be used for scientific purposes and not to grade the participants[3] in any way. After reading the instructions, the participants were given the opportunity to ask questions. There was a native speaker of the local language (Portuguese, Spanish or Italian) present in the room at all times, along with one or two experimenters. The experimenters could not speak or understand the local languages but explained the experimental setup to the local assistants beforehand. All local assistants also understood and spoke German at a native or near-native level.

The setup consisted of a standard desktop environment on a 15-inch Windows 10 Toshiba notebook with German keyboard layout, a cable-based mouse, a screen resolution of 1920 by 1080 pixels with 8 GB of memory and an Intel i5-6200U CPU. The browser cache and history was cleared after each participant. We used the same notebook for all participants in all locations but adapted the browser language to the respective local language.

2.1 Correction task

Each participant was presented with 18 German sentences[4] containing one error. The errors were constructed in such a way as to satisfy two requirements: (i) the error was typical for early-stage learners of German whose native language was a Romance language; (ii) the error could not be easily resolved by simply searching the web for the stimulus sentence or the part of the sentence containing the error. All sentences were designed by three of the authors of this paper (Idalete Dias, María José Domínguez Vázquez, Martina Nied Curcio) based on their long-term experience as 'German as a Foreign Language' teachers.

For example, one stimulus sentence was "An unserem Forschungsinstitut **ist** Ihnen unsere Bibliothek 24 Stunden **zur Verfügung**" (Eng. "At our research institute, our library is available to you 24 hours"). This stimulus contains an error in the light verb construction "zur Verfügung *sein*". The correct construction is "zur Verfügung *stehen*", hence, one possible correction would be "An unserem Forschungsinstitut **steht** Ihnen unsere Bibliothek 24 Stunden **zur Verfügung**". In Spanish, a correct version of the sentence would be "En nuestro instituto de investigación, nuestra biblioteca está abierta las 24 horas". The German "ist" can be seen as a direct translation of Spanish "está" (accordingly of Portuguese "está" and Italian "è"). The participants had to identify this as an invalid parallelism between Spanish and German and correct the error accordingly. If you search for the original stimulus sentence in Google, you would be faced with several pages of search results related to the libraries of a wide variety of research institutes, but no results dealing with the linguistic

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

properties of the sentence or the error itself.

It may be possible to argue that a correction task is a rather "unnatural" task for learners of German. A more "natural" task might have been to translate sentences from the participants' respective native language into German. However, we chose the correction task because it gave us the opportunity to use the same sentences for all participants from all countries. This, in turn, should reduce noise induced by stimulus sentences from different languages. All stimulus sentences can be found in the Appendix.

In terms of the technical setup, we used a simple Excel spreadsheet that contained the stimulus sentences in one column titled "Satz" (German for "sentence") and an empty column titled "Korrektur" (German for "correction") where the participants were to type their corrected sentence. The problematic parts of each sentence were highlighted in bold face (as indicated above), which was also explained in the participants' instructions. By using standard office software, we hoped to provide the participants with an environment they are well acquainted with. The participants were allowed to use Google Chrome or Mozilla Firefox whenever they wanted to refer to web content. They were not allowed to use any built-in assistance devices in Windows 10. The participants were not given a time limit before the experiment to avoid time pressure. After 30 minutes, each participant was told that they had 15 minutes left to work on the corrections. After 45 minutes, we told the participants that they should finish the sentence they were currently working on and then ended the experiment.

2.2 Screen recordings

The screen recording software ActivePresenter was started by one of the two experimenters in the room. We made sure beforehand that screen recordings did not interfere with the task in any way (e.g., pop-ups, screen flickering or the like). All actions of the participants were captured in the native display resolution.

2.3 Audio recordings

Since we did not want to rely on the notebook's built-in microphone to capture the voice of the participants, we recorded the thinking-aloud protocols with a high-definition external microphone. The audio recordings were inserted as the screen recordings' audio track after the experiments to allow for a synchronized investigation of both the screen recordings and thinking-aloud data. After we completed the data collection, the verbalizations of the participants were transcribed by native speakers of the respective language. German translations of the verbal protocols are also available.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

2.4 Annotations

The corrections that the participants entered were rated by two native German annotators. Five categories were available: "C", correct (all errors have been resolved), "CE", correct with errors (all errors in the stimulus sentences have been resolved but other errors have been introduced into the response), "D", case of doubt (it cannot be determined without a doubt whether the answer is correct or not), "W", wrong (the linguistic problem in the stimulus was not resolved or had been replaced by another), "N", not dealt with (the sentence had not been worked on, no attempt had been made to correct it). One example may illustrate the different categories: The stimulus sentence "Obwohl ich studiere, **wohne** ich noch **mit** meinen Eltern." (English "Although I am a student, I still live with my parents.") contains a wrong preposition. The correct version would be "Obwohl ich studiere, wohne ich noch *bei* meinen Eltern." This solution would accordingly be annotated as correct. An example of a CE-case (corrected with a new error) is the solution of participant R-02: "Obwohl ich studiere, ich wohne noch bei meinen Eltern." Here, the preposition "bei" is correct, but the word order "ich wohne noch bei" is a new error which is not part of the initial stimulus sentence. A wrong solution is, e.g., one made by participant S-09: "Obwohl ich studiere, ich mit noch meinen Eltern whone." Here, the wrong preposition is still there ("mit"), the word order is wrong, and a new spelling error "whone" occurs.

In 712 out of 816 cases (87.3 %), the two annotators labeled the answers of the participants identically. Weighted kappa (Cohen 1968) is $\kappa = .86$, indicating very good agreement between the annotators (we used the weighted kappa value because it penalizes disagreements that are farther apart from each other — e.g., "C" vs. "W" — more than disagreements that are closer to each other — e.g., "C" vs. "CE"). All disagreements were resolved through discussion.

To analyze research behavior, we also annotated the 2,225 search phrases that the participants used during their research. On the top level, three broad categories were distinguished: non-linguistic queries, metalinguistic terms and linguistic queries. (a) Non-linguistic queries are searches for a special dictionary or a general term like "duden wörterbuch", "alemao" or "pons tedesco". Queries were categorized as metalinguistic terms (b) whenever the query contained a linguistic term like "Konjunktiv 2 mit wenn" ("Konjunktiv 2 [a grammatical mood in German] with if"), "coniugazione verbi tedeschi" ("conjugation of German verbs"), "frases com verbos auxiliares em alemao" ("phrases with auxiliary verbs in German"), "deshalb significato" ("sense of 'deshalb'"), "Konzessivsätze mit 'obwohl' und 'trotzdem'" ("concessive clauses with 'obwohl' and 'trotzdem'"). Linguistic queries (c) are searches for words and phrases and are further divided into single-word searches like "beenden" ("to stop") vs. complex queries with multiple words like "ausser Frage" ("out of question") or "Es steht ausser Frage" ("It is without question"). The complex queries in sentence form

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

are also annotated for whether they are "(near-) verbatim" or "non-verbatim" copy-and-paste versions of the stimulus sentences.

3. Results and Discussion

3.1 General results

As we explained in the method section (2.1), our participants were presented with a maximum of 18 sentences for correction. On average, they edited 16 sentences. This number was nearly equal in all three locations (cf. Figure 1.1). The median (mean) number of edited sentences in Braga was 10.5 (11.4), 13.5 (12.1) in Rome, and 14.0 (13.1) in Santiago de Compostela. However, the number of correctly (category "C") improved sentences differed considerably between the three locations (cf. Figure 1.2). The median (mean) number of improved sentences in Braga was 2.5 (2.6), 7 (7.5) in Rome and 7 (7.1) in Santiago de Compostela. This result already points in a direction that is later supported by other results: although we hoped that our participants would reach the same language level in all three universities, the actual language level of the participants in Rome and Santiago de Compostela was clearly higher than of those in Braga.

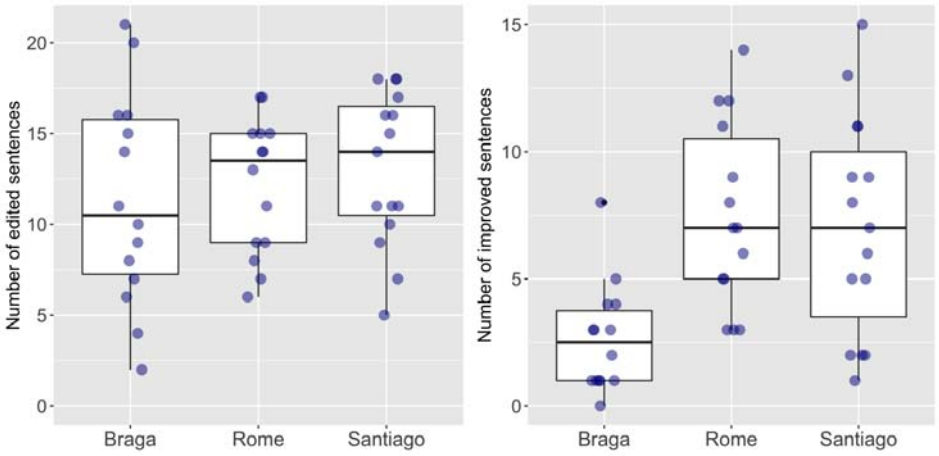


Figure 1: 1.1 (left): Number of corrected sentences in the three locations; 1.2 (right): Number of improved sentences in the three locations. Each dot shows the number of edited/corrected sentences for one participant. A total of 50% of all dots are surrounded by the box. The horizontal line within each box represents the respective median value.[5]

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

How many participants improved a sentence correctly also depends strongly on the sentence itself (cf. Figure 2). A sentence like "Leider kann ich heute nicht Tennis spielen. Ich bin zu **besetzt**." (in English, correctly: "Unfortunately, I can't play tennis today. I'm too busy.", Sentence-ID 2) with a false friend on the adjective position was improved in 70% of all cases, whereas the error in the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly: "No problem. If the sugar is empty, I'll take honey.", Sentence-ID 4) was obviously very hard to identify and transform into a search. In this case, only 17% of the corrections were annotated as improved. Table 1 shows one short excerpt of the search procedures referring to this sentence, illustrating the difficulties the participant had. The example shows that although the participant had the right idea at the end (looking for an adequate way to say "e'finito" in this context), they did not find an appropriate way to search for it. Another excerpt from a Spanish participant shows similar problems (cf. Table 2). The first idea many other students had concerning this sentence was that the participle, i.e. the grammatical form of "beendet," is wrong, but this is not the problem here. However, this initial idea led the students down the wrong path (for more information on the importance of the initial hypothesis, see Section 3.3.3). The combination of these types of quantitative analyses (here: which problems were difficult to solve?) and the closer qualitative inspection of the data (here: what exactly was difficult here and how did it affect search behavior?) is an advantage of the implemented study design. Thanks to this approach, we are able to evaluate exactly those aspects of dictionary use that cannot be identified on the basis of a log file or in a questionnaire study. Recording these difficulties, which leave the dictionary users at a complete loss, is a very useful insight for research into dictionary use.

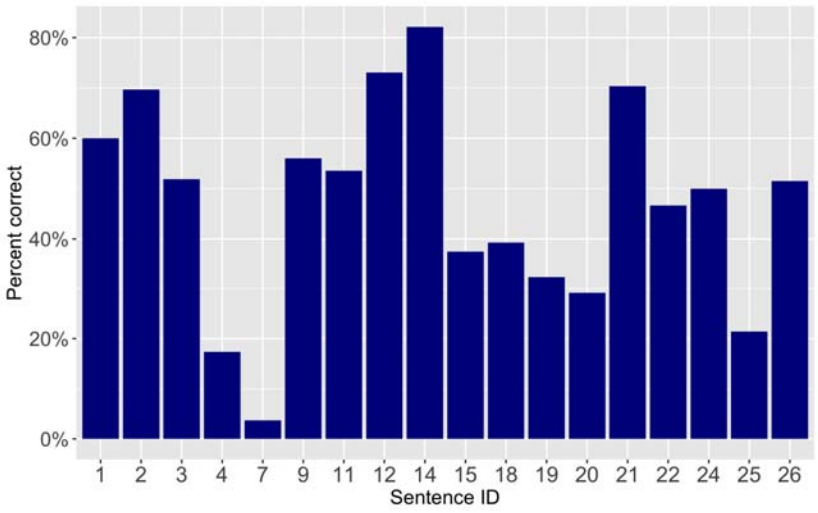


Figure 2: Percentage of improvements per sentence[6]

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Action	**Think-Aloud-Protocol**
Returns to Google results (search string was "beendet")	allora ehm non so cerco esempi perché non mi vengono soluzioni al momento (then ehm do not know I am looking for examples because I don't find solutions at the moment)
opens Deutsches Institut	
opens Bab.la	
returns to Google search results, googles "beendet esempi"	
opens Reverso Context	ehm sto cercando sto leggendo diciamo degli esempi # non ho idea [lacht] (Ehm I'm looking for I'm reading examples # I have no idea [laughing])
opens Excel	
opens Pons Traducao, searches for "e'finito"	sto cercando # (I'm looking)
opens Leo	sto cercando un modo per dire finito ahm (I'm looking for a way to say finished ahm)
opens Google	
opens Excel, no further corrections	okay non mi viene # non mi viene ahm # passo alla frase dopo perché non mi viene (okay I can't think of anything # I can't think of anything to say # I'm going to turn to the next sentence because I don't know)

Table 1: Excerpt from the study data of participant R-01 concerning the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly "No problem. If the sugar is empty, I'll take honey.")

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Action	**Think-Aloud-Protocol**
opens Leo, searches for "beenden"	vale # verbo beenden # que es acabar ## pero no sé si se puede usar para esto ehhm (ok # the verb 'beenden' # that means acabar # I don't know if it can be used for that ehhm)
opens Linguee, searches for "acabar la comida"	voy a mirar acabar la comida (I'm looking for 'acabar la comida')
searches for "acabar el bocadillo"	no # acabar (no # acabar)
searches for "beenden"	vale # miro en linguee beenden (ok # I'm looking for 'beenden' in Linguee)
opens Excel	
opens Linguee	no sé cómo buscar esto (I don't know how to look for it)
searches for "terminar comida"	
searches for "agotar existencias"	igual agotar (maybe 'agotar')
opens Excel, no correction	voy a dejarlo para después (I'm gonna save it for later)

Table 2: Excerpt of the study data of participant S-11 concerning the sentence "Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig." (in English, correctly "No problem. If the sugar is empty, I'll take honey.")

While transferring the screen recordings into analyzable data tables, we also encoded the position of the selected search result on the Google results page. The result is shown in Figure 3: Only the first four hits of the search results list are frequently selected (i.e. almost nobody scrolled because 4 to 5 results were directly visible on the laptop screen, depending on whether the window for the Google Translator was displayed or not). Almost two thirds of all selections (63%) concentrated on the first hit.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

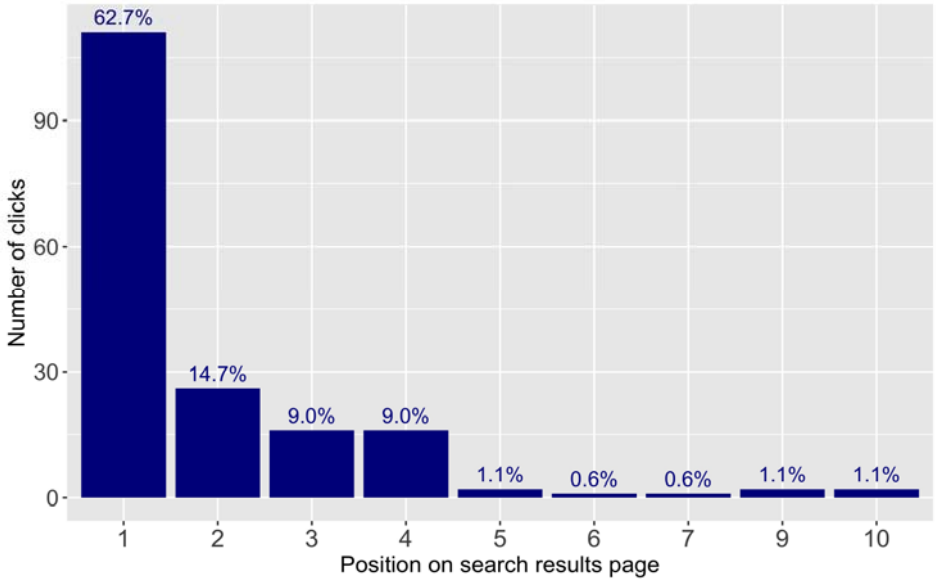


Figure 3: Position of selected search results (only Google was used as a search engine by the participants)

As mentioned at the beginning, we know little about what types of resources students actually use when solving language problems. Some teachers (personal discussion) claim that, nowadays, students use automatic translation programs far more frequently and hardly ever use dictionaries. However, according to questionnaire studies on this topic, online dictionaries are used very frequently (Levy and Steel 2015: 9, Koplenig and Müller-Spitzer 2014: 130). An important question in our study was therefore firstly to find out what types of resources our subjects use, and secondly to see whether they use different search strategies for different resource types or not.

First of all, our study shows that the students used a large number of resources and, above all, many different types (cf. Figure 4): Dictionaries or dictionaries with grammar tables were used the most, followed by search engines (which are, of course, also used to access resources, e.g. by entering "Duden online" in Google). Although 42 subjects is not a large number, the data are valid in the sense that we observed the students directly while working. This means that we do not have to rely on self-reporting, which in the context of language teaching, could be more distorted by some factor of social desirability, since the students usually know that their lecturers like to hear that they do not use automatic translation programs. In this sense, the data collected here may be understood as an encouragement to lexicographic work: indeed, students in our study seem to use dictionaries very often. In the majority (53.8%) of all trials

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

(i.e., sentence edits), one or more dictionaries were used, and these do not include other types of dictionaries, e.g., dictionaries with grammar tables (used in 35.5% of all trials), dictionaries with parallel texts and grammar tables (16.5%) and dictionaries with just parallel texts (11.5%).

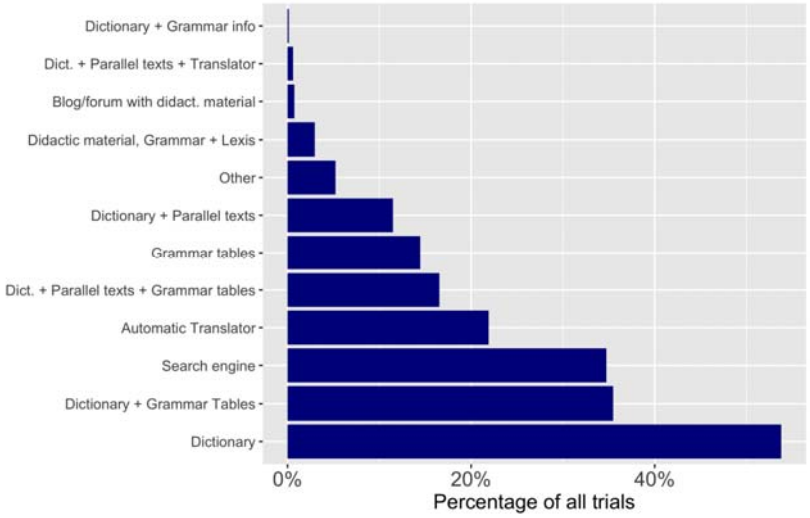


Figure 4: Types of resources used in percentage of all trials. Dict. = Dictionary; Didact. = Didactic

In the following section, we focus on the question whether the students differentiate their search strategies between different types of resources or not.

3.2 Search strategies

The data we gathered contain more than 2,200 search actions. In this section, we focus on the evaluation and analysis of these search actions. Above all, we want to investigate whether students use the various types of resources in different ways.

The language of most search strings is German (aggregated over locations, 69.4% of all search strings are in German), followed by search phrases in the local language (see Figure 5). The use of the local language (aggregated percentage: 22.3%) is remarkable in this study design because students had to conduct an improvement of German sentences, not a translation task. This 'bilingualization' of a monolingual task seems to have to do with the fact that our students want to use their mother tongue as an instance of certainty and/or track down the errors of interference by translating the German stimulus sentence back into their mother tongue and then using bilingual resources. This

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

strategy works very well in some cases. Interestingly, participants in Braga also rarely (but more often than the others) use English as a relay language. In the screen recordings, one can see that this was mainly done upon realizing that the consulted German–Portuguese bilingual resources achieved poor results (e.g. in an automatic translation program), but a translation from German–English as a first step and then English-Portuguese was more promising (see an example in Section 3.3.3). This use of English as a relay language came as a slight surprise for the language teachers involved in our study, but seems to be a viable strategy in some cases.

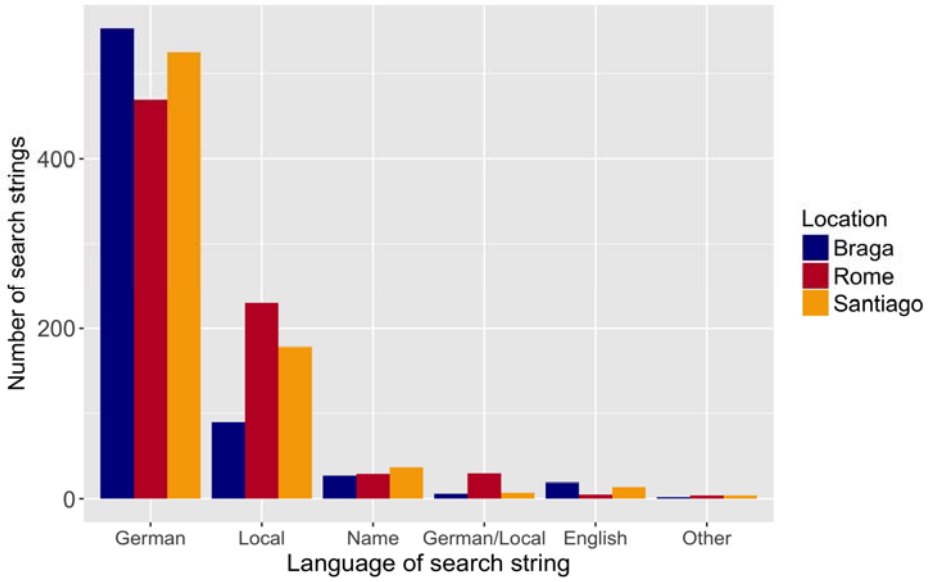


Figure 5: Languages of search strings (Local=Portuguese/Spanish/Italian, Name=Name of a resource)

It is well known that different types of resources are designed for different types of search queries. For example, it is generally not promising to enter entire sentences into the search fields of dictionaries, whereas the more context you have, the better automatic translation programs work. The question is, however, whether students are aware of this and adapt their search strategies accordingly. We wanted to use our data to investigate whether we could prove that our participants are aware of this.

In order to achieve this, we annotated all search strings as explained in the methods section. We see different patterns concerning the complexity of search strings used in different types of resources: although complex queries consti-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

tute the minority in all types of resources, the percentage thereof in automatic translation tools is higher than in all the other types (cf. Figure 6.1): In total, 42.5% of all queries in automatic translation tools are complex. These results may indicate that the participants are basically aware of the different functionalities of automatic translation tools vs. other types of resources. This impression is reinforced by the fact that there is an observable difference between use of the different resources from the same publisher or portal. While less than 5.5% of all queries in the Pons dictionary are multiple word items, there are more than 41.9% complex search queries in the Pons Translator even though both resources are presented on the same website (cf. Figure 6.2). Also the distribution of sentential vs. non-sentential queries points in the same direction: while sentential search queries constitute the majority (58.0%) of all queries in automatic translation tools, our participants almost never (1.9%) used them in dictionaries (Figure 6.3).

A further indication that the students use specific resource types depending on the kind of search query comes from the annotation and analysis of "(near) verbatim" and "non-verbatim" search queries (multiple word queries often seem to be verbatim copies of the stimulus sentences, see Figure 7). Google Translate and the Pons translator are clearly preferred if whole stimulus sentences are copied and pasted, i.e. for verbatim queries. In contrast, a resource like Reverso Context is used for non-verbatim queries.

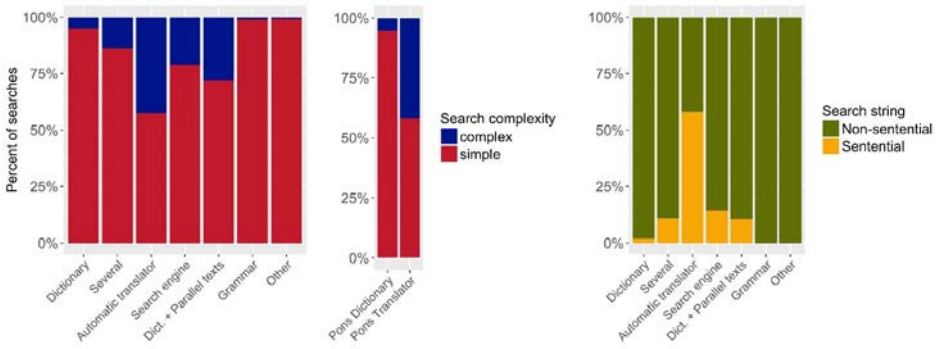


Figure 6: Figure 6.1 (left) Simple (one word) vs. complex (multiple word) queries in different types of resources; Figure 6.2. (middle) Simple vs. complex queries in Pons Dictionary vs. Pons Translator; 6.3 (right) Percentages of non-sentential and sentential search strings in different types of resources

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

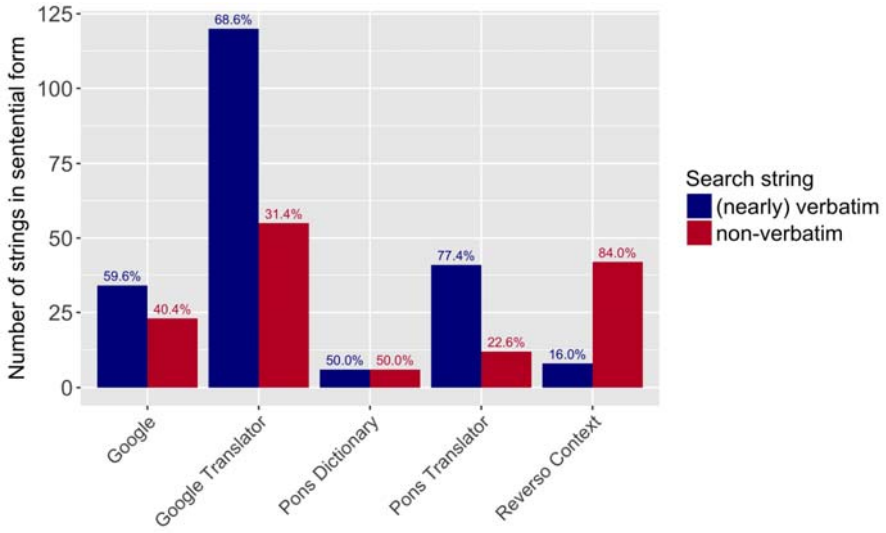


Figure 7: Verbatim vs. non-verbatim queries in sentential form (compared to stimulus sentences) in different types of resources, percentages above bars indicate the distribution within one resource.

Concerning our research question stated at the beginning of this section, we can conclude that the analyses of the search strings have shown that our participants seem to have at least a basic awareness of the different functionalities of the different types of resources used and adapt their search strategy accordingly. In the next section, we will take a closer look at which factors have an impact on the quality of the corrections.

3.3 Which factors influence the quality of corrections?

We now want to investigate whether we can identify systematic factors that influence the correctness of the improvements. We report our results concerning the correlation between types of resources and correction rate (3.3.1), the impact of careful reading (3.3.2) and the importance of initial hypotheses (3.3.3.).

3.3.1 Types of resources

One such systematic factor might be the number of different resources that are used to correct a sentence and whether this has a positive impact on the results. However, this is not evident. The main tendency related to the number of resources consulted is very similar in the case of correct improvements (Mean = 1.76, Median = 2), incorrect corrections (Mean = 1.76, Median = 2), cases of doubt (Mean = 1.92, Median = 2) and in the case of not attempting an improvement (Mean = 1.97, Median = 2) at all (cf. Figure 8). Likewise, the pro-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

cessing time per sentence has no influence on the improvement (no figure). Similarly, the position of the sentence in the study has no influence on the correctness, i.e. it was not the case that the first sentences were improved correctly more often than the last ones. Rather, it seems that there were sentences that were easy to improve even with few searches and in a short time, but others were not easy to correct even with a long overall processing time and many resources used.

In contrast, the type of resources used has an impact on the correctness rate. Two things in particular are striking. First, those participants who used more dictionary resources were more successful. The relationship is presented in Figure 9.1. This correlation is fed, in particular, by the participants from Braga, who revised only a few sentences correctly. A further subdivision of dictionary resources shows that dictionaries with parallel corpus examples such as Linguee tend to produce even better results. However, we must examine this particular connection in more detail before we can draw reliable conclusions. Second, our analyses show that the participants who rely more on automatic translation programs achieved poorer revision results (cf. Figure 9.2). As shown in Figure 9.2, this correlation is mainly influenced by our Portuguese participants who were less proficient in solving the task in general. Also note that the majority of participants used very few automatic translation programs (or none at all). This means that this correlation is driven by the fact that the better students also used more dictionaries and/or the worse ones use more automatic translation programs.

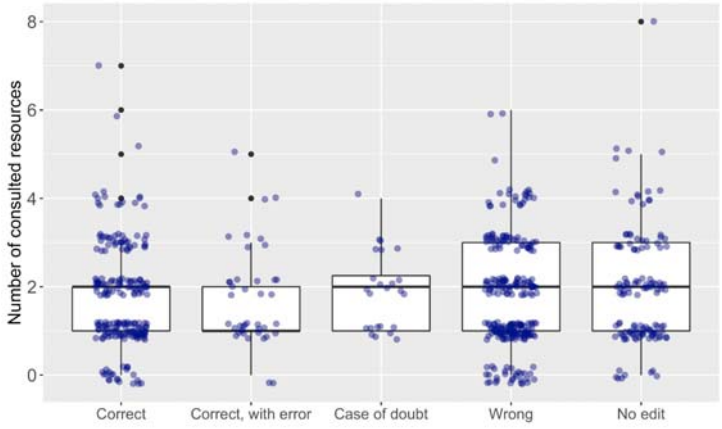


Figure 8: Number of resources used differentiated by correct improvement, correct improvement with new error, case of doubt, wrong corrected or no correction attempt at all (no edit). Each dot represents one sentence of one participant (one 'trial'). The box surrounds 50% of all data points. The horizontal line in each box represents the median value.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

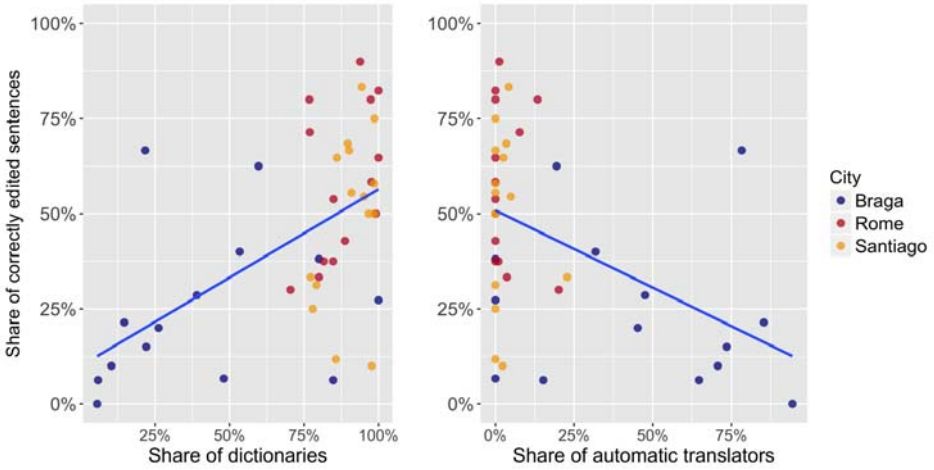


Figure 9: 9.1: Share of dictionaries in all used resources and percentage of improved sentences; 9.2: Share of automatic translation tools in all used resources and percentage of improved sentences. Each dot represents one participant (location is color-coded). The blue line represents the result of a linear regression fitted to the data.

3.3.2 Time spent using the resources and careful reading

Another key factor is time. Looking at the data (Figure 10), we found that there is a relationship between the average time spent using the resources and the correctness of the sentences. The mean difference between wrong and correct outcomes is relatively slight (only 2.4 seconds). However, it should be noted that this difference means that — on average — the time spent on each single resource is 2.4 seconds longer in each sentence edit that results in a correct sentence. During the course of the experiment, this difference may well amount to a much larger overall difference between correct and incorrect sentences. Interestingly, the different performance of the students in the different locations is also reflected in the time spent using the resources (Figure 11): On average, the students from Braga spent less time (Mean = 15.3 sec, Median = 14.8 sec) on the resources than the participants from Rome (Mean = 17.7 sec, Median = 16.1 sec) and Santiago de Compostela (Mean = 18.1 sec, Median = 16.3 sec).

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

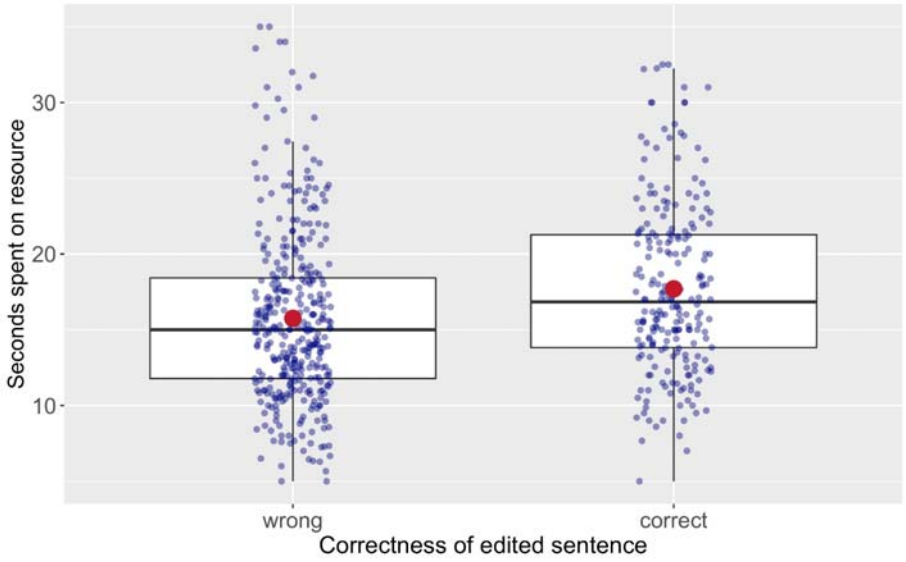


Figure 10: Average time spent using the resources and correctness of the sentences. Each dot represents one sentence edit from one participant (a 'trial'). Boxes are interpreted as in previous figures.

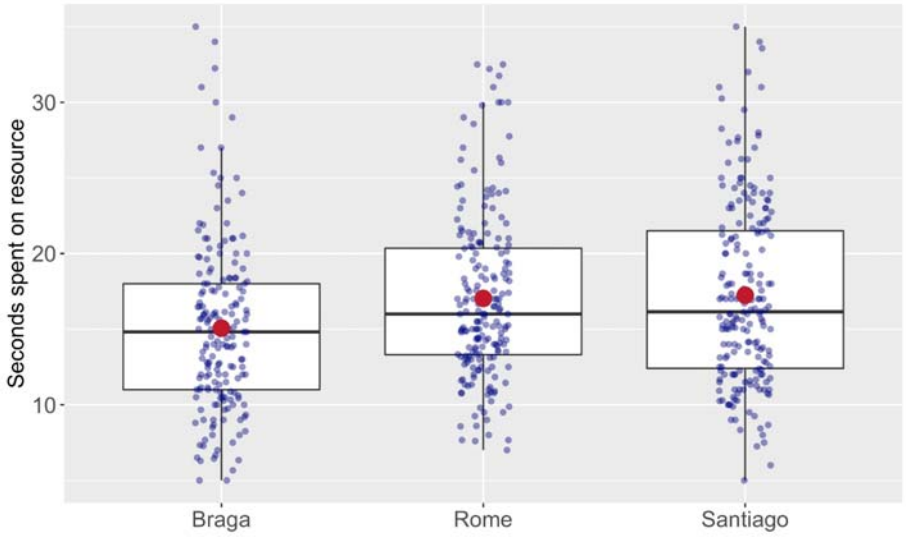


Figure 11: Time spent using the resources at the different locations. Each dot represents one sentence edit from one participant (a 'trial'). Boxes are interpreted as in previous figures.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Obviously, the short time spent using resources implies that students frequently switched between them. One example may illustrate this: Subject R-01 spent a total time of 3.5 minutes on sentence 1, undertaking 25 actions, which means that an average time of 7.65 seconds was spent on a single resource (without correction time). A look at the thinking-aloud-protocols (TAPs) confirms — even on a linguistic level — that subject R-01 takes hardly any time for the individual search queries; they often say "un attimo" or "un attimino" ('a moment'/'a minute'/'just'), e.g., "vado *un attimo* a vedere la costruzione di ehm la coniugazione di enden" ('I will just look at the construction ehm at the conjugation of enden'), "sto vedendo *un attimo* il verbo la coniugazione del verbo per essere sicura" ('I will just look at the verb's conjugation to make sure') or "okay vedo *un attimo* stipendium # okay" ('okay I will just look for stipendium # okay'). It is also very interesting that this student generally gave up very quickly if the solution could not immediately be found and then ascribed blame to the machine by saying "non mi trova niente" ('it doesn't find anything for me') or "cioè non mi sta trovando neanche degli esempi dello stesso verbo" ('so, it doesn't even find examples of the same verb for me') using the 3rd person singular to refer to the computer and/or the resource.

Other participants in contrast spent more time using each individual resource, reflected more upon their actions and achieved better results. These students seemed to solve problems very constructively, were aware of the potential difficulties, had previously developed language awareness, read attentively, and persisted in trying to tackle the same problem from various angles. Another example from Rome (R-07) illustrates this via excerpts from the TAP. Regarding the sentence *Obwohl sich der Junge beeilt hat, hat er die U-Bahn verloren* ('Although the boy hurried, he missed his train') the student was aware of the polysemy of the Italian verb *perdere* ('to lose', 'to miss'), which means that s/he had already developed a certain language awareness; they knew that in combination with a vehicle like *U-Bahn* ('underground train') the German verb *verlieren* (English to lose) was not correct and that a specific verb had to be selected. The student was aware that certain words belong together (collocations) and consequently searched for a specific word in the resources. That is the reason why a word-by-word translation (*perdere* – *verlieren*), which in these selected sentences usually leads to interference errors, could be avoided. In addition, the participant knew about various resources and opened an appropriate resource related to the search query, i.e. in order to find out the meaning of *verloren*, Pons was accessed; for the conjugation of *verpassen,* Reverso Coniugazione (Italian version) was the chosen resource. The student also used linguistic strategies such as searching for synonyms of *U-Bahn,* like *Zug* ('train'), which were considered more prototypical, and synonyms for *verlieren*. It is also interesting that subject R-07 often double-checked, i.e. by changing the search direction and checking the hypothesis, although R-07 was quite sure of the solution. This implementation of multiple strategies was also responsible for the high number of correct sentences. Of course, there is also the willingness to

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

solve the problem or to investigate it more rigorously and the will not to give up, as we can see in the extract in Table 3.

non posso purtroppo non posso andare allora in die Klasse gehen cerco ehm gehen se mi dà qualche utilizzo con Klasse magari se mi dà una frase simile quindi eh # allora (camminare andare a passeggio # andare in una stanza Zimmer in ein Zimmer) quindi allora se devo andare ehm devo usare in più l'accusativo quindi non è sbagliato l'articolo probabilmente il verbo # cerco Klasse se mi dà un un contesto d'uso per esempio no (viaggiare in prima seconda classe) no ehm okay quindi Klasse potrebbe essere anche una categoria forse ho capito male la frase quindi cerco anche Arzttermin (unfortunately I cannot go in die Klasse gehen so I will look up if ahm gehen somehow is used with Klasse maybe it will give me a similar sentence so ahm # so [camminare andare a passeggio # andare in una stanza Zimmer in ein Zimmer] so when I go ahm I must ahm I must use in plus Accusative so the article is not incorrect maybe the verb is incorrect # I will check if Klasse for example specifies a context of use no [viaggiare in prima seconda classe] no ahm okay so Klasse could also be a category maybe I didn't understand the sentence correctly so I will also search for Arzttermin).

Table 3: Extract of the TAP of student R-07 while working on the sentence "Morgen habe ich einen Arzttermin und kann deshalb nicht **in die Klasse gehen**".

However, it must be mentioned at this point that the time factor should not be considered in isolation. Due to the methodological design (including the TAPs), a longer and therefore more detailed, probably more intentional reflection influences the time spent using each resource. As a consequence, we cannot make a clear statement about the direction of the effect: are the more proficient students better at understanding the information in the resources and therefore spend more time using them, or does careful reading alone really lead to success? In other words, language proficiency, time spent using resources, and careful reading of dictionary entries form a complex inter-connected relationship. To allow for inferences, more experimentally controlled studies are required.

Additionally, a rigorous inspection of individual examples such as the ones presented above, incurs a risk of inferring general trends, which may not be confirmed by the overall data set. So, one has to make sure that the importance of individual examples is not overrated. However, as we have seen from the example of time spent on the sentences, the advantage of the data we collected is that these types of qualitative inspections encourage quantitative analyses which can then verify some data or adjust qualitative impressions. And, vice versa, quantitative results can be more closely examined through quantitative analyses (cf. Wolfer et al. 2018).

3.3.3 Searching guided by hypotheses

While analyzing the TAPs and the screen recordings, there seemed to be evidence that students' search behavior might be influenced by the initial hypotheses they formulate when analyzing the stimulus sentence. We will try to show that students tend to focus their initial hypotheses, thereby ignoring

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

relevant information in the online resources. In the following, we will describe this behavior in detail in order to make sense of students' search actions and develop a schema based on the observed search behavior patterns.

We begin our analysis with a description of the search actions carried out by a Portuguese student while trying to improve the following stimulus sentence: "Ich möchte ein Stipendium beim DAAD bewerben" ('I would like to apply for a scholarship at the DAAD'). Correcting the sentence involves identifying that: (i) the verb "bewerben" is a reflexive verb "sich bewerben" and (ii) "sich bewerben" is used with a prepositional phrase introduced by the preposition "um" followed by the object of the preposition in the accusative case: "Ich möchte *mich* beim DAAD *um ein Stipendium* bewerben".

From the TAP it is clear that the student does not know what the verb "bewerben" means. This leads the participant to look up the meaning of the verb in the Pons German–Portuguese Dictionary. The result provided by entering the search word "bewerben" is shown in Figure 12.

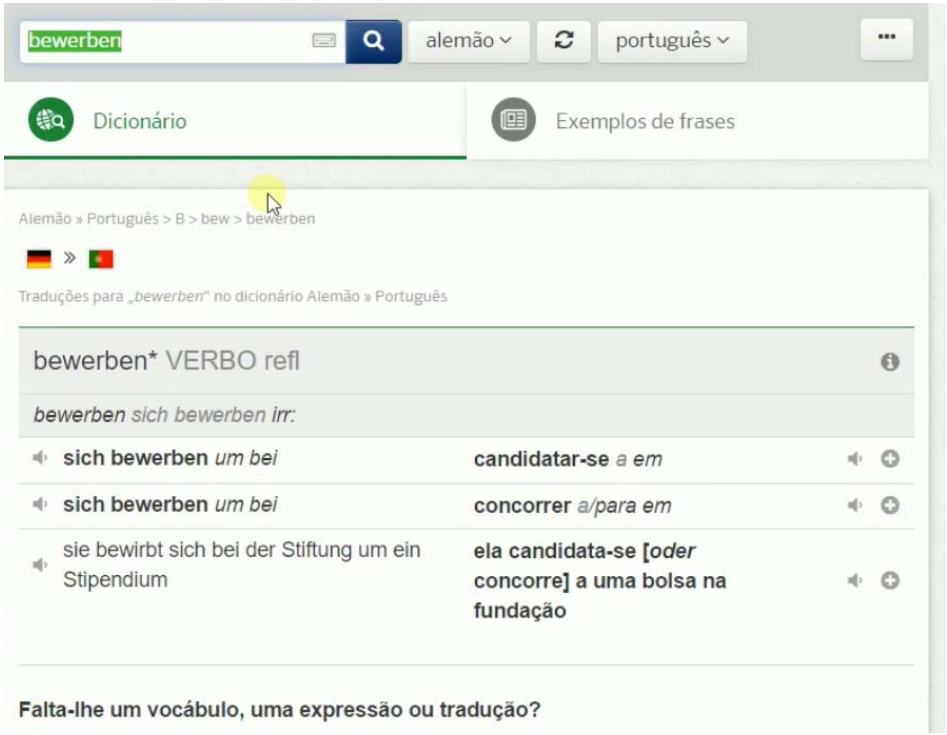


Figure 12: Result of search query "bewerben" in the Pons German–Portuguese Dictionary

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

As can be seen in Figure 13, the entry contains all the necessary information needed for the student to solve the task of correcting the stimulus sentence: (1) "bewerben" is a reflexive verb; (2) it requires the specific preposition "um"; (3) an example sentence is provided; (4) equivalents in the students' native language are provided. In addition, this information appears in the uppermost part of the entry. This means that, in effect, the student does not have to scroll through the entry looking for the answer(s) in order to correct the stimulus sentence. Taking into account studies on patterns of look-up behavior (Tono 1984, Lew et al. 2013), one would expect the student to pay special attention to the central information provided at the beginning of the entry.

Following from the TAP, the student reads the Portuguese equivalent "candidatar-se a", concludes that it is a reflexive verb and all further search actions aim at validating the hypothesis: the verb 'bewerben' in the stimulus sentence is missing the reflexive pronoun. The student focuses on the missing pronoun and does not analyze the entry any further. The information concerning the preposition "um" and the example sentence in the entry go completely unnoticed. To confirm the formulated hypothesis, the student applies the following steps:

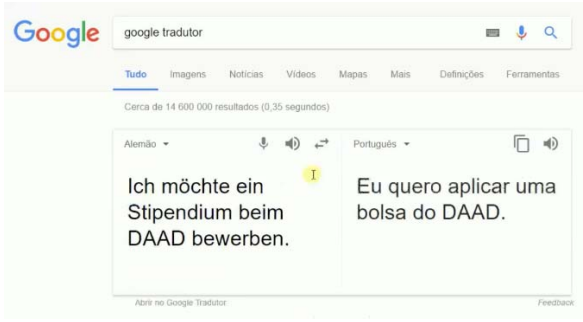
(i) S/he copies the entire stimulus sentence from the Excel file and pastes it in Google Translate (cf. Figure 14.1).

(ii) Since the Portuguese translation equivalent of the stimulus sentence sounds strange ("Eu quero aplicar uma bolsa do DAAD"), the student changes the target language of the translation to English (cf. Figure 14.2) and keeps using English until the task is over. The result is an incorrect German sentence corresponding to a correct English translation equivalent: "I would like to apply for a scholarship at the DAAD".

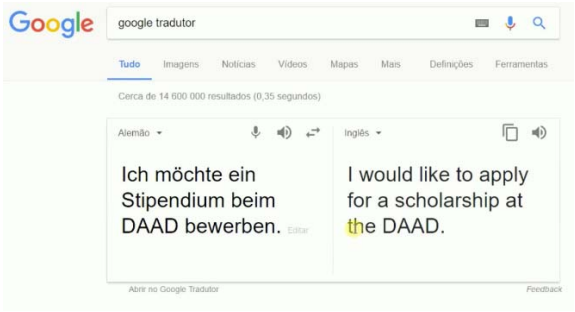
(iii) S/he switches the source and target languages and uses the correct English sentence as the source sentence (cf. Figure 14.3). The result is the correct German translation "Ich möchte mich beim DAAD um ein Stipendium bewerben". So this is an example where including English as a relay language was a promising strategy Interestingly, based on the TAP and the correction proposal ("Ich möchte mich ein Stipendium beim DAAD bewerben"), the student focuses exclusively on the presence of the reflexive pronoun in the correct German sentence, thereby validating the initial formulated hypothesis, and pays no attention to the preposition "um". This example shows how students use Google Translate and switch between languages to confirm their hypotheses.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

13.1



13.2



13.3

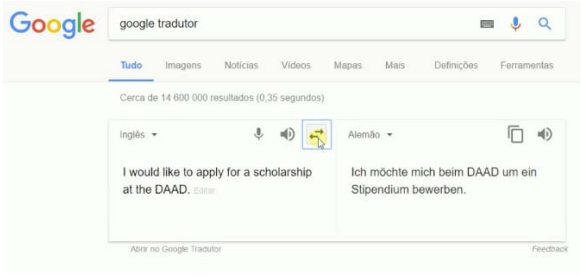


Figure 13: 13.1: Google Translate result for the language pair German–Portuguese; 13.2: Google Translate result for the language pair German–English; 13.3: Google Translate result for the language pair English–German

Based on qualitative observations of the focalization hypothesis in all three participant groups, we arrived at the focalization hypothesis search pattern which can be explained as follows: The students begin by formulating an initial hypothesis (like e.g., in this example: "bewerben" is a reflexive verb), based either on intuition before initiating a search process or on hypotheses formulated on the basis of a specific search action, such as the search for the meaning or translation equivalent of a word. From this point onwards, the whole search process focuses exclusively on the attempt to confirm this hypothesis (see more exam-

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

ples in the Euralex proceedings paper of this study, Wolfer et al. 2018: 109-111). The observational data seems to indicate that students normally focus their attention on the first result they find in the resources that matches their hypothesis and do not search any further. We also observed that an incorrect initial hypothesis in most cases leads to absurd search actions and results. Furthermore, participants who experience difficulties confirming their hypothesis usually cease to make an effort to correct the stimulus sentence.

The focalization hypothesis described above was identified while conducting a qualitative analysis of participants' search behavior. We aim to complement these qualitative findings with quantitative methods in order to compare the datasets in a more systematic manner and gain a deeper insight into students' search behavior.

4. Conclusions

The combination of quantitative and qualitative methods via the examination of verbal protocols and screen recordings has proven to be an effective approach with which to identify search strategies and patterns common to a specific participant or across participant groups. We received empirical data on important questions related to using online language resources and can draw the following conclusions based on our data.

Although our participants' language proficiency levels are not very high, they use a rather broad range of language resources, most of which are accessed via a Google search and not consulted directly, e.g. by typing in the name of a dictionary in the address bar. Our participants are quite aware of the different functionalities of search engines, translation tools and dictionaries. This can be seen in the fact that they adapt their search strings according to the type of tool. Verbatim or near-verbatim parts of the stimulus sentence are mostly looked up in automatic translation tools and not in dictionaries. We identified three factors that influence the correction rate systematically. (i) Participants who use dictionaries more often than other types of tools are more successful in correcting the stimulus sentences. (ii) Participants who spend more time using the language resources are also more successful in correcting the errors. So, careful reading seems to be one influential factor for solving the task in our study. (iii) Another important factor seems to be whether or not the correct hypotheses are formulated before launching the online search. Participants who had the wrong hypotheses did not see the right solutions although they were presented on the screen. One should keep in mind that we do not know whether the students with a higher level of language proficiency also use dictionaries more frequently (because they have more competence in doing so), spent more time using the language resources (because they can gain a deeper understanding from the presented content) and have better initial hypotheses. Or if two students with the same level of language proficiency really perform differently if they vary in their use of dictionaries vs. translation tools, read

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

more or less carefully and spend more or less time reflecting on the initial hypotheses. It may also be the case that some students were particularly motivated and therefore read very carefully. So, this is a classical chicken-and-egg question. But what we can see in our data is that these three factors — using dictionaries, careful reading and starting with the right hypotheses — seem to be indicators of successful user behavior.

This leads us to aspects we would change in further studies. Above all, we would do two things differently in future studies: Firstly, we would conduct a short language test prior to the study because this would allow us to identify whether there is a clear connection between language competence and search behavior. We suspect this for our participants in Braga in contrast to those in Santiago de Compostela and Rome, but are not able to prove this assumption. Secondly, we would use a translation task instead of improving sentences in the foreign language. For this study, one central point was to have the same task for all three locations. However, the data we gained show that the task was quite artificial for the students, especially by jumping back and forth between the native and foreign language. On the other hand, as one of the reviewers of this paper argued, the sentence improvement task had the advantage of demanding specific correct vs. incorrect answers and a translation task would not be as clear-cut. For further studies, this issue must be taken into careful consideration. The methodical structure with screen recording and thinking aloud, on the other hand, worked very well. However, in the future we would practice thinking aloud before starting the test, at least briefly with each test person, in order to facilitate speaking during the study.

Empirical studies such as this one are also important because many of the results of our study were unexpected for the language teachers involved: the use of the local language, sometimes even English as a third language in alternation with German, the differentiation between dictionaries and translation programs, the measurable influence of careful reading and the strong influence of the correct starting hypothesis. All this seems almost predictable in retrospect, but was not so beforehand. In our opinion this is exactly where the teaching of language should begin: instead of making general assumptions about what resources are used by students today, our study data could firstly be used as an opportunity to discuss with own students and language learners what resources they use and what strategies they implement. Secondly, at least according to this study, the basic knowledge of different types of language resources should be used to teach even more strategies that support and develop dictionary usage competence. This teaching approach should always be grounded on students' actual use of lexicographic resources. In our opinion, studies such as this one are particularly helpful in this respect. In a further step, it would be important to collect more data in a similar manner in order to investigate whether these results are also confirmed in other countries, for other languages and with other tasks. As Bowker puts it "the key […] is for lexicographers to listen to users" (Bowker 2012: 396).

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Endnotes

1. See https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions (last accessed 28 June 2018).
2. In Lexicographica 2018/2019, there is a German publication on this study with the title: "Recherchepraxis bei der Verbesserung von Interferenzfehlern aus dem Italienischen, Portugiesischen und Spanischen: Eine explorative Beobachtungsstudie mit DaF-Lernenden" (same authors as this article). This year's (2018) Euralex proceedings also include a more methodologically oriented contribution to this study entitled "Combining Quantitative and Qualitative Methods in a Study on Dictionary Use" (Wolfer et al. 2018). — We would like to thank Alexander Koplenig for discussing the study results, all assistants and contractors involved in the study, as well as the participants of the IDS Colloquium in fall 2017, the EMLex Colloquium in Stellenbosch and the FaDaF Conference 2018 in Mannheim with whom we discussed the study results. Special thanks go to the participants of the study for their cooperation and to the Institute for the German Language for financing the study. Finally, we would like to thank both reviewers for their very valuable comments.
3. We found this especially important because the participants were recruited by a subset of the authors of this paper who were also their university teachers at that time. By including this section in the instruction and due to the fact that the teachers were not present during the study, we tried to make sure that the participants behaved as "naturally" as possible, i.e. that they did not only consult sites of resources that were taught during their university lessons or avoid specific sites.
4. The first three participants from Braga, Portugal, received 26 sentences instead due to human error on behalf of the experimenters. The 18 sentences that were presented to all 43 participants were also included in the stimuli for these three participants. We will mention the biases and the measures we took to control for them throughout the respective sections.
5. All plots in the present paper were created with the ggplot2 package (Wickham 2016) for the R environment for statistical computing (R Core Team 2018).
6. The IDs of the sentences go up to number 26, since more sentences were initially meant to be improved, but the pre-tests showed that this was not feasible in the given time.

References

Bowker, Lynn. 2012. Meeting the Needs of Translators in the Age of e-Lexicography: Exploring the Possibilities. Granger, Sylviane and Magali Paquot (Eds.). 2012. *Electronic Lexicography:* 379-397. Oxford: Oxford University Press.

Cohen, Jacob. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70(4): 213-220. doi:10.1037/h0026256.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe and Sarah Hillewaert. 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.

Domínguez Vázquez, María José, Mónica Mirazo Balsa and Vanessa Vidal Pérez. 2013. Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

lernenden Hispanophonen. Domínguez Vázquez, María José (Ed.). 2013. *Trends in der deutsch–spanischen Lexikographie*: 135-172. Frankfurt a.M.: Peter Lang.

Domínguez Vázquez, María José and Carlos Valcárcel Riveiro. 2015. Hábitos de uso de los diccionarios entre los estudiantes universitarios europeos: ¿nuevas tendencias? Domínguez Vázquez, María José, Xavier Gómez Guinovart and Carlos Valcárcel Riveiro (Eds.). 2015. *Lexicografía de las lenguas románicas II. Aproximaciones a la lexicografía contemporánea y contrastiva*: 165-189. Berlin: De Gruyter.

Ericsson, K. Anders and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data.* A Bradford Book. London: The MIT Press.

Frankenberg-Garcia, Ana. 2011. Beyond L1–L2 Equivalents: Where do Users of English as a Foreign Language Turn for Help? *International Journal of Lexicography* 24(1): 97-123.

Hult, Ann-Kristin. 2012. Old and New User Study Methods Combined — Linking Web Questionnaires with Log Files from the *Swedish Lexin Dictionary*. Fjeld, Ruth Vatvedt and Julie Matilde Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress 2012, 7–11 August 2012, Oslo*: 922-928. Oslo: University of Oslo, Department of Linguistics and Scandinavian Studies University of Oslo. http://www.euralex.org/elx_proceedings/Euralex2012/pp922-928%20Hult.pdf. (Accessed 11 July 2018.)

Koplenig, Alexander, Peter Meyer and Carolin Müller-Spitzer. 2014. Dictionary Users Do Look Up Frequent Words. A Log File Analysis. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 229-250. Berlin/Boston: De Gruyter.

Koplenig, Alexander and Carolin Müller-Spitzer. 2014. General Issues of Online Dictionary Use. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 127-142. Berlin/Boston: De Gruyter.

Levy, Mike and Caroline Steel. 2015. Language Learner Perspectives on the Functionality and Use of Electronic Language Dictionaries. *ReCALL* 27(2): 177-196. doi:10.1017/S095834401400038X. (Accessed 11 July 2018.)

Lew, Robert. 2010. Users Take Shortcuts: Navigating Dictionary Entries. Dykstra, Anne and Tanneke Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6–10 July, 2010*: 1121-1132. Ljouwert: Afûk.

Lew, Robert. 2011. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4.

Lew, Robert. 2015a. Opportunities and Limitations of User Studies. Tiberius, Carole and Carolin Müller-Spitzer (Eds.). 2015. *Research into Dictionary Use / Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie"*: 6-16. (OPAL — Online Publizierte Arbeiten Zur Linguistik 2015(2)). Mannheim: Institut für Deutsche Sprache. http://pub.ids-mannheim.de/laufend/opal/pdf/opal15-2.pdf. (Accessed 11 July 2018.)

Lew, Robert. 2015b. Research into the Use of Online Dictionaries. *International Journal of Lexicography* 28(2): 232-253.

Lew, Robert, Marcin Grzelak and Mateusz Leszkowicz. 2013. How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. *Lexikos* 23: 228-254.

Müller-Spitzer, Carolin. 2014. *Using Online Dictionaries*. (Lexicographica: Series Maior). Berlin/Boston: De Gruyter.

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Müller-Spitzer, Carolin and Alexander Koplenig. 2014. Online Dictionaries: Expectations and Demands. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 143-188. Berlin/ Boston: De Gruyter.

Nied Curcio, Martina. 2013. Der Gebrauch zweisprachiger Wörterbücher aus der Sicht italienischer Gemanistikstudierender. *Lexicographica* 29: 129-145.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/. (Accessed 11 July 2018.)

Tono, Yukio. 1984. *On the Dictionary User's Reference Skills*. B.Ed. Thesis. Tokyo: Tokyo Gakugei University.

Töpel, Antje. 2014. Review of Research into the Use of Electronic Dictionaries. Müller-Spitzer, Carolin (Ed.). 2014. *Using Online Dictionaries*: 13-54. Berlin/Boston: De Gruyter.

Welker, Herbert Andreas. 2013. Empirical Research into Dictionary Use since 1990. Gouws, Rufus H., Ulrich Heid, Wolfgang Schweickard and Herbert Ernst Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 531-540. Berlin/Boston: De Gruyter.

Wickham, Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer.

Wolfer, Sascha, Martina Nied Curcio, Idalete Maria Silva Dias, Carolin Müller-Spitzer and María José Domínguez Vázquez. 2018. Combining Quantitative and Qualitative Methods in a Study on Dictionary Use. Čibej, Jaka, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*: 101-112. Ljubljana: Ljubljana University Press, Faculty of Arts. https://e-knjige.ff.uni-lj.si/ znanstvena-zalozba/catalog/view/118/211/2981-1. (Accessed 11 July 2018.)

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1466

Appendix: Stimulus sentences

ID	***Satz***
1	Meine Nachbarin möchte immer alles wissen. Sie ist sehr **kurios.**
2	Leider kann ich heute nicht Tennis spielen. Ich bin zu **besetzt**.
3	Bist du **bereit?** Wir müssen jetzt los, wir sind sowieso schon zu spät dran.
4	Kein Problem, wenn der Zucker **beendet** ist; ich nehme dann Honig.
7	Ich bin **einverstanden mit dir**.
9	Das erlaube ich dir nicht. Es **ist außer Frage.**
11	An unserem Forschungsinstitut **ist** Ihnen unsere Bibliothek 24 Stunden **zur Verfügung**.
12	Obwohl ich studiere, **wohne** ich noch **mit** meinen Eltern.
14	**Wenn** ich zur Schule ging, habe ich viel Sport gemacht.
15	Morgen habe ich einen Arzttermin und kann deshalb nicht **in die Klasse gehen**.
18	Ich **vorbereite** gerade meine letzte Prüfung.
19	Ich möchte **ein Stipendium** beim DAAD **bewerben**.
20	Ich habe die Hose viel zu klein gekauft. Jetzt muss ich nochmals ins Geschäft zurück und sie **wechseln**.
21	Obwohl sich der Junge beeilt hat, hat er die U-Bahn **verloren**.
22	Er wohnt seit Jahren in Berlin und trotzdem **verliert** er **sich** immer noch.
24	Um beim Kartenspielen zu gewinnen, musst du exakt die Regeln **folgen**.
25	Der Artikel **handelt sich um** die Migranten in Deutschland.
26	Ich möchte **dir** heute über einen interessanten Artikel **sprechen**.