The use of experimental methods in linguistic research: advantages, problems and possible pitfalls

Barbara Mertins

Abstract: In the present paper I will present and discuss several experimental methods used inside and outside psycholinguistic research. The overall focus will be on language production. The methods presented include different elicitation techniques, eye tracking, memory tasks and preference judgment tasks. On the basis of my own experimental data, I will describe the main features of these methods, comment on their suitability for various linguistic research questions, and explore some of their advantages, shortcomings and limitations. The paper addresses general methodological issues and challenges, going beyond the research conducted on Slavic languages. However, all the studies presented and discussed here are based on data collected from native speakers of various Slavic languages. In addition, two studies address language production of Slavic native speakers in a foreign language. The paper concludes with general remarks on the use of experimental methods and statistics in linguistic research.

1 Introduction

The use of experimentally based methods and techniques has become fairly popular in linguistics. Researchers from different linguistic areas employ methods originating in experimental linguistics and psycholinguistics to test various linguistic theories, models and concrete research hypotheses. The choice of a method or a methodological approach always means commitment to a particular experimental design. This choice in turn puts specific requirements on the stimulus material or the selection of participants and has, in the end, consequences for data coding and analysis. It is therefore essential to have some knowledge of the advantages and disadvantages of a particular method before employing it in an experiment.

This paper is structured as follows: Section 2 presents a classification of experimental methods and discusses their advantages and disadvantages. Section 3 focuses on selected aspects of language production research and introduces methods employed in my own research. Section 4 comprises three studies chosen from my own research on the basis of which different methods are explained and evaluated. This section also includes relevant

details concerning the design of an experiment. The article ends with a set of conclusions.

2 A classification of psycholinguistic methods

Experimental methods can be classified in different ways (cf. Höhle, 2010; Müller, 2013; Vanpatten & Jegerski, 2014). In this paper three method types are distinguished (cf. Schmiedtová & Flanderková, 2012): (1) *offline* methods; (2) *online* methods; (3) *true online* methods. I will concisely describe the different types and provide examples for each of them. Then I will elaborate on some pros and cons of the methods and make some general remarks on their suitability for linguistic research.

The terms *offline*/*online* relate to the degree, to which a given method reflects the studied underlying mental and/or neuronal process. The offline methods focus on speakers' linguistic competence, whereas the online methods concentrate more on speakers' performance. (1) *The offline methods* have no direct access to a mental process and reflect conscious decision-making. The tasks are solved with a delay in time. A good example of an offline task is a paper-and-pencil questionnaire (which can also be administered in a more modern manner as a web-based task) or object naming, a method frequently used with special participant groups, such as aphasic patients. It is characteristic for the second group (2), *the online methods* that they offer mediated access to underlying mental processes. These processes are more automatized and unconscious. The participants have to solve an experimental task with only a short time delay. Examples of these methods are reaction time experiments[1] or eye-tracking, both methods frequently used in psycholinguistic research. The last method types (3) are *the true online methods*. These methods have immediate access[2] to the relevant process and can assess highly automatized and unconscious mental and neuronal processes. Functional magnet resonance imaging (fMRI) or electroencephalography (EEG) with the measurement of event related potentials (ERP) are examples for these methods.

---

1 In some other classifications reaction time experiments are considered an offline method. These classifications do not differentiate between online and true online methods. Instead they collapse all behavioral methods in one method type (offline) and keep the online method type only for electrophysiological and neuroimaging methods.

2 Researchers in cognitive sciences and neurolinguistics assume that, even in case of true online methods, the access to the relevant neuronal processes is delayed. This is perhaps true, but not relevant for the purpose of this paper.

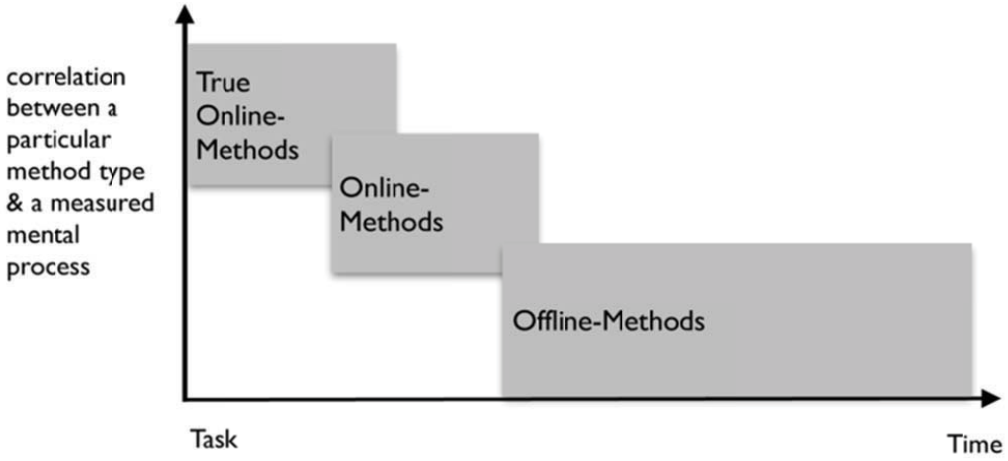

Figure 1: Methods in relation to time

In the following paragraphs I will describe the advantages and disadvantages of the three method types.

Offline methods:

A problem when using offline methods is that the researcher has no or little control over the course of the data collection. This is especially true when the data are collected via an internet-based questionnaire. However, when this approach is employed, it is usually possible to measure the time the participants need to finish the task. When an excessive amount of time is used, participants can be excluded from the data set. A big advantage of the offline methods is that (very) large data amounts can be collected at once. Their use also means easy logistics and almost no costs. The offline methods aim at testing participants' competence and are suitable for getting (first) insights into participants' linguistic preferences or grammatical judgments. The data from an offline task can help to generate specific hypotheses for an experiment[3].

Online methods:

One disadvantage of the online methods is a certain slowdown, due, for example, to the speed of hand movements, when measuring reaction times in a lexical decision task. But also in eye-tracking there is a slowdown

[3] Offline methods can be also used for testing a specific linguistic hypothesis. This depends on the aim and the research question of the individual study.

caused by the movement of the eye. Online methods make relatively high organizational demands (one person per recording session), hence smaller samples. Considering the fact that one needs about twenty participants to run a proper statistical analysis this is not a trivial point, especially in the case of cross-linguistic research. In contrast to the offline methods, the researcher has more control over the experimental procedure and the task execution. The online methods are suitable for testing unconscious and automatized mental processes, with focus on participants' performance. These methods are a good tool for testing concrete linguistic hypotheses.

True online methods:

These methods make considerable organizational and financial demands. This is why studies using true online techniques are usually based on data from only a few subjects. Another disadvantage of these methods is a rather heavy dependency on "hidden" statistical procedures and calculations. This means that various tools used for the data analysis and visualization (e.g., Brain Voyager for functional and structural MRI data sets) operate on many preset defaults that are quite impossible for the researcher (let alone layperson) to retract and understand. Moreover, because of the complexity of the entire experimental protocol, only simplified experimental designs can be employed. Additionally, there are also serious technical restrictions on task execution (e.g., on free language production). An advantage of these methods is the possibility to study highly automatized and unconscious mental and **neuronal** processes. Depending on the type, these methods are suitable for investigating the time course of language processing (e.g., ERP in EEG) or the localization of language skills (e.g., PET, fMRI). In my opinion the true online methods should be only used, when behavioral methods can no longer provide answers.

3 Language production research

Before going into a more detailed description of the various methods I use in my own research, I would like to mention a number of points related to the study of language production in general.
This research area has a long tradition in linguistic research (for an overview see Carroll, 2008). It deals with all phenomena linked to the production of spoken and written language in different populations. There are some challenges to be dealt with, especially when studying spoken language: The logistical and technical requirements are quite high since participants must be tested individually. As mentioned above, a general rule is to have data

from at least twenty subjects to carry out a good statistical analysis. However, the empirical value shows that because of unexpected technical problems and other sources of data loss, one must record about 30% more participants than the minimal number required for a proper statistical analysis.

Another point is the comparability of experimental settings: The entire experiment is almost never run by one and the same person. But even if only one person is in charge, it is very important to keep the experimental protocol across individual recordings as consistent as possible and to minimize variations in the experimental procedure, incl. the instruction and interactions between participants and investigator.

As in any other research the creation of a good stimulus set is an essential prerequisite for a well-designed study. Going into all aspects, which need consideration when creating stimulus material, would go beyond the scope of the present paper. Also, different research questions and experimental designs call for different stimulus material. But in general, the following points should always be taken into account in order to avoid undesirable side effects and biases: a sufficient amount of fillers (a rule of thumb: twice as many fillers as testing/critical items), control over word frequency and word length (optionally also the number of syllables), control over the degree of concreteness/abstractness, and awareness of intercultural aspects.

Last but not least: The transcription of spoken language (audio data) is very time-demanding. This needs to be taken into consideration when planning a language production study.

The methods and tasks I use in my own research are: elicitation, memory tasks, eye-tracking, the measurement of speech onset times (SOT), preference and grammatical judgment tasks. Depending on the study, the methods are either used alone or in combination. Due to the scope of this paper I cannot explain the SOT method and grammaticality judgment tasks. As for elicitation, I will describe this method in more depth because it is widely used in experimental linguistics.

(1) *Elicitation*[4] is in very general terms the act of obtaining specific language data from another person. Depending on the time constraint, it qualifies either as an offline or an online method. There are many areas of linguistic research for which this method is fitting. For example: (a) elicitation of a particular linguistic structure (e.g., *case* — cf. Dąbrowska et al., 2006); (b) elicitation of a phenomenon rarely occurring in spontaneous speech (e.g., *simultaneity marking* — cf. Schmiedtová, 2004); (c) testing of a specific hypothesis (e.g., *determiners before tense marking* in child acquisition — cf. Wit-

[4] In the present paper, the term *elicitation* is confined to eliciting linguistic structures.

tek & Tomasello, 2002). Data can be elicited in different ways, for example by context restrictions, stimulus manipulations, or creation of minimal pairs.

Depending on the research question and the population to be studied, pictures, picture books, audio recordings, written texts or video clips can be employed. Pictures are often used to elicit children's language (cf. Clark, 2009). The well-known picture book "The Frog story" has been a popular elicitation tool over the past decades in many different contexts and all kinds of populations (cf. the CHILDES database: http://childes.psy.cmu.edu). Audio recordings serve as stimuli in phonetics and phonology research. For instance, written texts can be employed to study production in association tasks (cf. Glucksberg & Danks, 2013 for an overview). I use video clips for eliciting spoken language in adult native and second language speakers. (2) Another task I employ in my own research are *memory tasks*. They are well suited to collecting non-linguistic data, an important complement to linguistic data when examining *linguistic relativity* (more detail see section 4). The memory task in my own research was administered to the participants with a time delay, which is why it qualifies as an offline method. (3) *Eye-Tracking* is an online method that makes it possible to study eye-movements and to test hypotheses concerned with allocation of visual attention (e.g., for testing the effects of language on cognition). I combine eye-tracking with (4) measurements of *speech onset times* (SOT)[5], another online method providing an insight into the planning processes taking place just before a participant begins to speak. Two additional methods that can be found in my own research, including (5) *the preference judgment task* and *the grammatical judgment task*[6]. Both tasks belong to the offline group and are suitable for testing a particular linguistic phenomenon in larger or specific populations.

4 Examples from my own research: research questions and the use of different methods

In this section, I will present and discuss a number of studies carried out either by myself or in cooperation with colleagues. The focus will be on how selected methods introduced in section 3 are applied in concrete experimental settings. I will critically discuss different aspects linked to the planning of an experiment (e.g., choice of participants, experimental procedure,

---

5 For more detail on the measurement of speech onset times see Schmiedtová (2011a).

6 Grammatical judgment tasks are commonly used in adult native as well as L2 speakers for testing grammatical acceptability. In my research, this task was employed for testing grammatical knowledge of patients with Broca aphasia (Flanderková, Mertins, et al., 2014).

determining level of proficiency in an L2, and creation of stimuli). When relevant, I will point out possible problems and pitfalls.

Study 1: Elicitation (Schmiedtová & Sahonenko, 2008)

Elicitation has been employed frequently in my own research (cf. Schmiedtová & Sahonenko, 2008; Schmiedtová, 2011, 2011a; v. Stutterheim, et al., 2012; Schmiedtová, 2012, 2013, 2013a). In the majority of these studies, elicitation was used in combination with other tasks (for more detail see below). In the article by Schmiedtová & Sahonenko (2008) elicitation was the only method employed. Because of this, I will present and discuss this study in more detail.

The focus of Schmiedtová & Sahonenko (2008) was to examine the role of grammatical aspect and tense in the encoding of goal-oriented motion in adult native speakers (L1) of Czech, Russian, German and very advanced second language speakers (L2+[7]) of German with L1 Czech or Russian. Based on previous work on German, English, French, and Italian (e.g., Carroll & v. Stutterheim 2002; v. Stutterheim & Carroll, 2003; v. Stutterheim & Lambert, 2005) the research question was posed of how and to what extent core grammatical categories determine how information is selected and structured in dynamic contexts. The related L2 research question was concerned with the restructuring of conceptual knowledge[8], i.e. with the question to what degree are near-native L2+ speakers able to learn to reorganize conceptual knowledge (e.g., encoding of motion events) towards the target language pattern.

We used 40 short video clips depicting different goal-oriented motion events (critical items) and homogenous activities serving as fillers (distractors) for the elicitation of spoken data. The length of the clips varied. The stimulus material appeared automatically on a laptop screen in random order with a five second blank in between. The participants' task was to start to speak as soon as they knew what was happening in the clip. The question in the instruction was presented in present tense (German: *Was passiert*?; Czech: *Co se děje*?; Russian *Čto proischodit*?). In order to ensure comparable conditions across recordings, a standard experimental procedure was developed and set down in written text to be repeated in every session. We also

---

[7] The abbreviation L2+ refers to second languages speakers who speak the target language as their third, fourth or even fifth language. It reflects the fact that European L2 speakers (and participants in our studies) are often multilingual and German is not always their second foreign language.

[8] The explanation and discussion of the terms conceptual restructuring and conceptual knowledge can be found in Schmiedtová (2011a, 2013).

controlled for the effect of language mode[9] (cf. Grosjean, 1998). To make sure that participants were exposed only to the tested language during the recording, only a native speaker of this language (Czech, German, Russian) was present at the recording and interacted with the participant[10]. The audio data were digitally recorded, transcribed and coded by the investigators. The coding scheme comprised the coding of grammatical aspect, tense, and reference to endpoints. In order to calculate intercoder reliability[11] for data from each language, we asked another linguist (who was also a native speaker of that language) to code large parts of the data. This way for each language there were three coders (the authors of the study and an additional linguist). For the data analyses we used a combination of qualitative and quantitative (statistical) tools.

Thirty native speakers for each L1 as well as 30 advanced L2+ speakers of German (15 with Czech L1, 15 with Russian L1) were recruited for this study. All participants were comparable in terms of socio-economical and educational background[12]. The native speaker data were collected in the respective countries. All native speakers were students. All native and L2+ speakers were between 20 and 30 years old (average age 24.6 years). The L2+ speaker data were collected in Germany (Russian L2+ speakers of German living in Heidelberg) and in the Czech Republic (Czech speakers of L2 German living in Prague[13]). All L2+ speakers were either students of German in higher semesters or professionals (e.g., interpreters, translators, German language teachers).

Since our study dealt with language production of advanced and very advanced L2+ speakers we had to ensure that the proficiency level in German was comparable across participants. Accessing the proficiency degree

---

9 A number of previous studies have shown that the choice of language mode can have a great impact on language processing in bi- and multilingual speakers (cf. Soares & Grosjean, 1984; Cenoz et al., 2001; van Hell & Dijkstra, 2002).

10 The control of language mode is very important. It is, however, a question to what extent (and if at all) one can make bi- and multilingual speakers "switch off" the language(s) that is/are not being actively used at a particular moment. For example, in two eye-tracking experiments Marian & Spivey (2003) have demonstrated that the non-active language affects spoken language processing in bilingual speakers.

11 The calculation of intercoder reliability (or intercoder agreement) is in my opinion an absolutely essential prerequisite for any study of linguistic data. I will elaborate this point in the concluding part of this paper.

12 To ensure the comparability of these variables and to create homogenous participant groups we developed a biographical questionnaire that participants had to fill out before the experiment.

13 At that point in time, we were unable to find enough very advanced L2 speakers of German with L1 Czech living in Heidelberg or nearby.

in advanced L2+ speakers is certainly a challenge. In my knowledge, only few studies dealing with topics concerning near-native L2+ speakers (e.g., ultimate attainment issues) have made an effort to lay out their procedures for determining advancedness[14]. I think that this is an unfortunate situation that needs to be changed, as such studies should make sure that the L2(+) participants are near-native in the target language. In our study, we used a combination of linguistic and extra-linguistic criteria for establishing the advancedness of a L2+ speaker. (1) Excellent language knowledge: This parameter was established on the basis of a warm-up interview that was recorded and later transcribed. We qualified only those speakers as advanced who made no grammatical errors in agreement, word order and inversion. Some article errors were tolerated. On the basis of this criterion we excluded three participants from the study. (2) Active use of German in everyday life: We only included speakers who indicated in the biographical questionnaire that they use German as their dominant language in daily life. Dominant was defined as at least 70% of all everyday situations (also for the Czech participants). We did not exclude any participants on the basis of this criterion. (3) An early onset of acquisition: More than 60% of the L2+ speakers in our study started to learn German as a foreign language in primary school, i.e. around the age of 10. (4) A longer stay in a German-speaking country: All L2+ speakers with L1 Russian had been living at the time of the experiment at least four years in Germany. For the Czech L2+ speaker group the criterion was a minimum two-year sojourn in a German-speaking country. (5) Highly tutored L2 acquisition: All L2+ participants learned German at a certain point in their life in school (average length of school tutoring was 4.7 years). Applying these five criteria we were able to put together two very homogenous and comparable L2+ groups with perfect or near-native command of German. The majority of them (80%) were female.

In summary, the online elicitation task with video clips serving as stimuli was a suitable method to study the research questions examined in Schmiedtová & Sahonenko (2008). Furthermore, this study clearly demonstrated that the widely spread notion of "Slavic aspect", which often only includes the Russian system must be further differentiated. This is further supported by the next study discussed in the present paper (v. Stutterheim et al. 2012) that shows that these differences are not only in the linguistic but also in the underlying conceptual system.

---

[14] In some studies self-assessment is used as the only measure of language proficiency. I find this problematic since self-assessment is a very subjective and culturally dependent measurement (cf. MacIntyre et al., 1997 discussing biases in self-rating of language proficiency and the role of anxiety).

Despite all these positives, I would like to point out several problems linked to the design of the study: The choice of stimulus material was suboptimal since the video clips were not controlled for length, type of protagonist (person, animal, object, vehicle, etc.), the direction from which the protagonist appears (left vs. right), or intercultural aspects (e.g., a clip with a typical yellow German mailbox was used which was not immediately recognized by speakers of languages other than German). This clearly disadvantaged those speakers. The number of fillers was too low: Only about a third of the stimulus set consisted of distractors. Yet another difficulty emerged with the instruction text used. As mentioned above, the participants were asked to say what was happening in the clip. We did not, however, instruct them explicitly to concentrate on the event. This imprecision led to some participants producing descriptions of the protagonists, the surrounding environment, etc., rather than the event depicted in the clip. These texts had to be excluded from the analysis because they did not follow the posed *quaestio* (v. Stutterheim & Klein, 1987). A helpful workaround would have been to pilot the instruction before the experiment and adapt it accordingly. Another problematic point was that the two L2+ groups differed in the amount of exposure to German. However, this factor was taken into account when comparing the groups statistically. With respect to the analyzed categories (number of endpoints and the use of tense) no relevant between-group differences were found in terms of the country of residence (Prague vs. Heidelberg). The last point of criticism concerns the number of L2+ speakers in Schmiedtová & Sahonenko (2008). Because it was not possible at the time of the study to recruit more than fifteen L2+ speakers, there were not enough data to perform all statistical analyses. This problem, of course, will always come up when studying participant groups, such as atypical populations (e.g., SLI children) or near-native L2+ speakers, for which there is no "endless" pool of possible subjects one can recruit from (as is the case for native speakers). Nevertheless, as mentioned already, fifteen subjects is not a sufficient number of speakers to do a thorough statistical analysis. In a follow-up study[15] (v. Stutterheim et al., 2012) all these shortcomings were removed and the design was improved. These improvements will be explained further.

[15] The problems of the low number of L2+ participants and different country of residence at the time of testing were removed in another set of studies (Schmiedtová, 2011, 2013) in which elicitation was used either alone or in combination with eye tracking and memory task to study language production of near-native L2+ speakers.

Study 2: Non-linguistic Tasks — Memory Task & Eye-Tracking (v. Stutterheim et al., 2012)

The study by v. Stutterheim et al. (2012) combined the elicitation of spoken data with a simultaneous recording of eye movements and a subsequent memory task. In contrast to the previous study (Schmiedtová & Sahonenko, 2008), this paper examined only native speakers. The general research question was concerned with the effects of language on cognition, i.e. with testing the *thinking-for-speaking hypothesis* (Slobin, 1996) and the *seeing-for-speaking-hypothesis* (Carroll et al., 2004; Schmiedtová et al., 2011). The aim of this study was the encoding of endpoints in goal-oriented motion events in Czech, Dutch, English, German, Russian, Spanish, and Modern Standard Arabic.

Compared to Schmiedtová & Sahonenko (2008) the stimulus material had been improved with respect to the aspects: the number of fillers, standardized video clip length, control of type and appearance of the protagonist, intercultural usability. In total 60 short video clips including 10 critical, 10 control items and 40 fillers were filmed[16] for the purpose of this study. The critical clips showed goal-oriented motion events, in which a potential endpoint was not reached within the duration of the clip (e.g., two persons walking on the pathway, in the background a building). The control items depicted goal-oriented motion events with an endpoint reached before the end of a clip (e.g., a vehicle going along a street, turning and disappearing into a garage). The fillers showed 30 activities with causative events (e.g., a person making a necklace) and 10 static scenes (e.g., a candle burning). The video clips were six seconds long. The number of clips depicting people, animals and vehicles was comparable. The direction of the appearance of the protagonist (left vs. right) was equally distributed across all critical and control items. In addition, all videos were piloted before the experiment with about 100 students with different language and cultural backgrounds to ensure their intercultural transferability.

The task for the participants was to verbalize what was happening in the clip. As in the other study (Schmiedtová & Sahonenko, 2008), the emphasis was on depicting the event and the question was posed in the present tense. The instruction text was improved and included an explicit request not to verbalize any descriptions and to concentrate solely on the event. The text was translated by native speakers into all languages and presented to the

[16] The video clips were filmed and cut over the course of three months by members of a research group at the University of Heidelberg. Twenty clips in total were made and they were all piloted and pretested. Only ten were selected for the experimental stimulus set.

participants first orally and then in written form. The experimenter was a native speaker of the language tested (cf. control of language mode). Each experimental session was preceded with six testing items covering all testing categories. The experimental items were presented automatically from a computer screen, in a pseudo-randomized order, with an eight-second interval in between to give participants sufficient time to finish their verbalization[17]. The elicitation and eye-tracking data were recorded simultaneously. Each recording session took approximately 15 minutes. After that, participants were asked to fill out a biographical questionnaire designed on the basis of the questionnaire used in Schmiedtová & Sahonenko (2008). Subsequently, and without prior announcement, a memory task was administered to the participants (see below for more detail on the design of the memory task). This task took between two and five minutes to finish.

For this study, we recorded data from twenty subjects per language, i.e. from 140 participants in total[18]. For logistical reasons all data had to be collected in Heidelberg. The speakers of Arabic, Czech, Dutch, English, Russian and Spanish were participants in a summer school at the University of Heidelberg, and had no or very little knowledge of German. To minimize the exposure to German, the subjects were recorded in the first five days of their stay in Heidelberg. An utmost effort was made to ensure that all speakers were as "monolingual" as possible, with English being the only foreign language all participants were able to speak (at different proficiency levels). All participating subjects, including native speakers of German, were matched in terms of socio-economical background and were students or postgraduates, aged 20-35 (average age 26.7 years). The groups were balanced for gender and all participants had normal or corrected vision.

The elicited linguistic data and the memory data were transcribed and coded by respective native speakers. The linguistic analyses included the coding for temporal/aspectual categories and reference to endpoints. The transcriptions and the coding schemes for these two tasks were checked for consistency by a second researcher.

---

17 The length of the in-between-clip-interval had also been tested in a pilot study. In a study by Schmiedtová (2013b) the elicitation of spoken data was performed under time pressure so the blank between the presented clips was reduced to three seconds. Such a design aims at eliciting highly automatized responses and presents another good method for studying participants' performance.

18 This number refers to subjects whose data were used for the analysis. The actual number of participants recorded was much higher (see above for a detailed discussion of data loss).

The eye-tracking data included the following measurements: the total fixation count within the area of interest[19], the total fixation duration, and the number of first and second periods of fixation. All eye-tracking analyses were run with average measures across participants as well as averages over items. The memory task comprised fifteen color screen shots in which a specific part was cut off. There were ten critical items in which the endpoint was removed and five control items where a random object was missing. The control items were used to control for general memory performance. The task for the participants was to write down as fast as possible and in only one or few words what exactly was cut out.

Before evaluating the experimental design and the suitability of the chosen methods in v. Stutterheim et al. (2012) I would like to make several general comments on the use of non-linguistic methods and tasks for testing linguistic relativity hypothesis. A question to raise here is: What counts as an effect of language on thought/cognition[20]? In other words, how can it be ensured that the observed effects reflect the influence of language on thought. It is not uncommon in linguistic and anthropological research, from which the linguistic relativity theory has emerged, to assume (or even to claim) cognitive differences solely on the basis of variations in linguistic data. Differences in linguistic form, for instance, are very relevant and may lead to finding differences in cognition-but not necessarily (cf. Lucy, 1996 — an excellent article with relevant methodological thoughts and hints for the study of the relation between language and thought). When linguistic differences are found, usually by means of various behavioral (offline or online) methods, one must employ yet another method to make sure that diversity in language leads to differences in thinking. To this end, a number of methods (e.g., eye-tracking, memory tasks as used in v. Stutterheim et al., 2012) and non-linguistic tasks can be employed (e.g., sorting, matching, classification, or categorization tasks as used in Lucy, 1992 or Levinson et al., 2002). I believe that using data from behavioral tasks with data from non-linguistic tasks and methods is the **only** way to (a) show actual effects of language (or grammatical structure) on cognition; and thus (b) escape the argumentative tautology of claiming language effects on cognition based only on linguistic differences.

---

19 An area of interest (AoI) or a critical region are key terms from the eye-tracking research referring to the part of the stimulus where the eye movement (or gaze movement) is recorded.

20 For a definition and discussion of the terms language, cognition, thought, see Schmiedtová, 2011a.

Overall, the results of v. Stutterheim et al. (2012) have shown that the chosen methods and tasks, especially in their combination, proved to be excellent for testing the effects of language on thought. Compared to the previous study (Schmiedtová & Sahonenko, 2008), the experimental design, including stimulus material, instruction text, and intercultural transferability, was improved and yielded reliable data. The only two minor points to comment on are the recording of native speakers outside their native country and the absence of intercoder reliability calculation. It is obvious that one should opt for collecting data from native speakers in their respective native countries. However, considering the size of the data sample in v. Stutterheim et al. (2012) and the logistics of making recordings of eye-tracking data in seven different countries, it would have been nearly impossible to satisfy this point. The other point of criticism is more serious: Although in v. Stutterheim et al. (2012) a second researcher was asked to check the transcripts and the coding, I am of the opinion that the only way to develop an objective "waterproof" coding scheme is by employing coding of at least two other coders (optimally a mix of linguists and "naïve" native speakers). Based on the coding of several independent "blind" coders, intercoder reliability can be calculated and if necessary the coding schemes adjusted.

Study 3: Preference judgment task (Schmiedtová, 2013a)

Preference judgment tasks have been successfully employed in linguistics for a long time. As pointed out in section 2, they represent a powerful tool to gather large data sets with relatively little effort. However, caution should be exercised when designing these tasks since the selection of the right stimulus material is not trivial or easy. To demonstrate a possible way to design a preference judgment task I selected a study of my own (Schmiedtová, 2013a). In this study an extensive judgment task was designed to test preferences in aspect use in Czech native speakers. The underlying hypothesis was that in contemporary spoken Czech the usage of the *present perfective form* (e.g., *vy-pije*$_{PF}$ "she/he drinks up") has been extended (perhaps under the influence of German, Schmiedtová, 2012, 2012a) from future to *here-and-now* reading. To test this hypothesis, a questionnaire was developed comprising 35 scenarios, 15 critical and 20 fillers, all presented in present tense contexts. The fillers were motion verbs embedded in goal-oriented motion events with a potential endpoint (e.g., somebody riding a bike on a pathway, in the background is the beginning of a forest). The critical items were verbs depicting a situation with a resultant state (e.g., somebody drinking a cup of coffee, somebody throwing garbage into a trash can). Czech verbs are classified into five different conjugation classes. In order to test whether a particu-

lar verb class allows the use of the present perfective form in *here-and-now* reading, three verbs from each class were selected.

In order to avoid priming effects, the target verb did not appear in the prestory. So for instance, in the critical scene "throwing away garbage into a trash can" (the target verb, *vyhodit*$_{PF}$*'/vyhazovat*$_{IMPF}$ "to throw away") the wording was as follows: "Imagine a situation, in which you see a man standing next to garbage containers doing something. He is nearly finished with the activity he has been involved in. How would you most likely describe such a situation?" After reading this text participants could choose from five options in which the target verb appeared in five different tempus/aspect combinations (i.e. present imperfective, past imperfective, present perfective, past perfective, secondary imperfective)[21]. Except the difference in tempus/aspect, these options were identical in wording. The participants' task was to check off the most preferred description of a given situation and if needed, also indicate their second best preference. The 35 scenarios were presented in a pseudo-randomized order, in the form of a paper-and-pencil questionnaire and administered to 256 participants. The questionnaire was piloted with ten native speakers of Czech. Educational level and age of the participants were taken as factors for ensuring homogeneity of the participant group in terms of age and socio-economical background. The subjects were either pupils in the last year of high school or first semester university students (age range 17-30; average 19.3). The gender was not controlled. The questionnaires were filled out in regular classes with a standardized instruction given to the participants orally by their teacher[22]. The participants had twenty minutes to finish the task. To investigate a possible influence of dialectal variations on the use of the present perfective, data were collected in five different regions of Czech Republic. The questionnaires were anonymous and included only information regarding gender, native language, and the origin of the participants. Two subjects were excluded from the sample because they grew up in bilingual families.

Despite the fact that preferential judgment tasks come with some downsides, e.g., the participants may not indicate their real preferences because they find the task odd or boring and make their choices randomly, the task was suitable for the investigation of the questions studied in Schmiedtová (2013a). One may suggest performing a corpus analysis instead, which would perhaps yield (even) more data points. The problem with a corpus

---

21 These are all possible combinations in Czech.

22 Several colleagues of mine kindly did the collection of the data. They received detailed instructions on how to proceed in the collection of the data. In this manner, the procedure was comparable.

study would have been to control the context, in which the tested form was presented (as it had to be present tense).

Before concluding the current paper, I would like to point out several aspects that should be considered when planning a judgment task. (1) Stimuli: In addition to the more formal points listed in section 3 it should be taken into account that the language material in a questionnaire should be natural and not grammatically odd. Also, testing linguistic preferences or grammatical acceptability on isolated items (i.e. without any context) is highly problematic. (2) Fillers and presentation: For a test in written language, the use of a large amount of fitting fillers is absolutely essential. In addition, one has to control for the presentation order since participants have a tendency to connect individual items in a "meaningful way" (e.g., creating some kind of a story) or to select items repeatedly from only one place on the page (e.g., the first choice from the left). (3) General: A lengthy questionnaire will not yield good data due to the attention and interest span of the participants. I would recommend shorter tasks tested on a larger number of participants.

5 Final Remarks

There are many different experimental methods suitable for linguistic research. The focus of the current paper was in regards to the offline and online methods. When planning an experiment, a number of aspects must be taken into consideration in order to come up with a good experimental design and thus usable data. The relevant aspects include the selection of stimulus material, the experimental protocol, the recruitment of participants as well as the coding and analysis of data. Because of ecological validity the calculation of intercoder reliability for the coding of linguistic data is indispensable. For the data analysis, I would always opt for the use of inferential statistics. A prerequisite for this is a proper experimental design and a sufficient number of data points. In my opinion, basing a study only on qualitative analyses does not lead to meaningful and generally valid results, except for case studies in language pathology and child acquisition research. Last but not least: Although the use of experimental methods is crucial for doing linguistic studies, the research cannot be done without a good linguistic theory, yielding interesting and challenging research questions.

References

Carroll, D.W. (2008). *Psychology of language*. Belmont: Cengage Learning.

Carroll, M., & Stutterheim, C. (2002). Typology and information organisation. Perspective taking and language-specific effects in the construction of events. In A.

Ramat (Ed.), *Typology and Second Language Acquisition* (pp. 365-402). Berlin: de Gruyter.

Carroll, M., Stutterheim, C.v., & Nüse, R. (2004). The language and thought debate. a psycholinguistic approach. In C. Habel, & T. Pechmann (Eds.), *Approaches to Language Production* (pp. 183-218). Berlin: de Gruyter.

Cenoz, J., Hufeisen, B., & Jessner, U. (2001). *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Clevedon, UK: Multilingual Matters.

Clark, E. (2009). *First language acquisition*. Cambridge: Cambridge University Press.

Dąbrowska, E., & Szczerbinski, M. (2006). Polish children's productivity with case marking: the role of regularity, type frequency, and phonological diversity. *Journal of Child Language, 33*(3), 559-597.

Flanderková, E., Mertins, B., Bezdíček, O., Baborová, E., & Černá, M. (2014). Posuzování gramatičnosti v Brocově afázii příklad dvou pacientů. *Česká a slovenská neurologie a neurochirurgie 77/110*(2), 202-209.

Glucksberg, S., & Danks, J.H. (2013). *Experimental Psycholinguistics (PLE: Psycholinguistics): An Introduction*. Hoboken: Psychology Press.

Grosjean, F. (1998). Transfer and language mode. *Bilingualism: Language and Cognition, 1*(3), 175-176.

Hell, J.v., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. *Psychonomic Bulletin & Review*, *9*(4), 780-789.

Höhle, B. (Ed.) (2010). *Psycholinguistik*. Berlin: Akademie Verlag.

Levinson, S., Kita, S., Haun, D., & Rasch, B. (2002). Returning the tables: Language affects spatial reasoning. *Cognition, 84*, 155-188.

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.

Lucy, J. (1996). The scope of linguistic relativity. In J.J. Gumperz & S.C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 37-69). Cambridge: Cambridge University Press.

MacIntyre, P.D., Noels, A.K., & Clément, R. (1997). Biases in Self-Ratings of Second Language Proficiency: The Role of Language Anxiety. *Language Learning, 47*(2), 265-287.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition 6*(2), 97-115.

Müller, N. (2013). Transfer in bilingual first language acquisition. *Bilingualism: Language and Cognition 1*(3), 151-171.

Schmiedtová, B. (2004). At the same time: The expression of simultaneity in learner varieties. Berlin: de Gruyter.

Schmiedtová, B. (2011). Do L2 speakers think in the L1 when speaking in the L2? *International Journal of Applied Linguistics, 8*, 97-122.

Schmiedtová, B. (2011a). Wie Sprache unser Denken formt – psycholinguistische Hintergründe. In S. Schulte (Ed.), *Ohne Wort keine Vernunft – keine Welt: Bestimmt Sprache Denken?* (pp. 97-128). Münster: Wachsmann.

Schmiedtová, B. (2012). Vergleich von deutschen und tschechischen kunsthistorischen Texten. In S. Höhne, I. Fiala-Fürst, R. Mikuláš, & B. Schmiedtová (Eds.), *Brücken 2011. Germanistisches Jahrbuch Tschechien – Slowakei; thematischer Schwerpunkt – Sprachwissenschaft* (pp. 221-240). Praha: Lidové Noviny.

Schmiedtová, B. (2012a). *Untersuchung zu Sprache und Kognition am Beispiel von Ereigniskonzeptualisierung und Textkohärenz im Deutschen und Tschechischen*. [Unveröffentlichte Habilitationsschrift]. Ruprecht-Karls Universität Heidelberg.

Schmiedtová, B. (2013). Traces of L1-patterns in the event construal of Czech advanced speakers of L2-English and L2-German. In C.v. Stutterheim, M. Flecken, & M. Carroll (Eds.), *IRAL (51*), 87-116.

Schmiedtová, B. (2013a). Zur Verwendung der perfektiven Präsensform im heutigen Tschechisch. *Journal for Central European Studies*, *(2)*, 125-164.

Schmiedtová, B. (2013b). Zum Einfluss des Deutschen auf das Tschechische: Die Effekte des Zeitdrucks auf die Sprachproduktion. In M. Nekula, K. Šíchová, & J. Valdrová (Eds.). *Bilingualer Sprachvergleich und Typologie* (pp. 177-206). Tübingen: Julius Groos Verlag.

Schmiedtová, B., Stutterheim, C.v., & Carroll, M. (2011). Implications of language-specific patterns in event construal of advanced L2 speakers. In A. Pavlenko (Ed.). *Thinking and Speaking in two languages* (pp. 66-107). Clevendon, UK: Multilingual Matters.

Schmiedtová, B., & Flanderková, E. (2012). Neurolingvistika: předmět, historie, metody. *Slovo a Slovesnost, 73,* 46-62.

Schmiedtová, B., & Sahonenko, N. (2008). Die Rolle des grammatischen Aspekts in Ereignis-Enkodierung: Ein Vergleich zwischen Tschechischen und Russischen Lernern des Deutschen. In P. Gommes, & M. Walter (Eds.), *Fortgeschrittene Lernervarietäten: Korpuslinguistik und Zweitspracherwerbforschung* (pp. 45-71). Tübingen: Max Niemeyer.

Slobin, D. (1996). From "thought to language" to "thinking for speaking". In J.J. Gumperz, & S.C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Soares C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory & Cognition, 12 (4),* 380-386.

Stutterheim, C.v., Andermann, M., Carroll, M., Flecken, M., & B. Schmiedtová (2012). How grammaticized concepts shape event conceptualization in language production: Insights from linguistic analysis, eye tracking data and memory performance. *Linguistics, 4*, 833-867.

Stutterheim, C.v., & Carroll, M. (2003). Typology and information organisation: perspective taking and language-specific effects in the construal of events. In A. Ramat (Ed.), *Typology and Second Languge Acquisition* (pp. 365-402). Berlin: de Gruyter.

Stutterheim, C.v., & Klein, W. (1987). Quaestio und referentielle Bewegung in Erzählungen. *Linguistische Berichte, 108*, 163-183.

Stutterheim, C.v., & Lambert, M. (2005). Crosslinguistic analysis of temporal perspective in text production. In H. Hendricks (Ed.), *The structure of learner varieties* (pp. 1-19). Berlin: de Gruyter.

Vanpatten, B., & Jegerski, J. (Eds.). (2014). *Research methods in second language psycholinguistics*. New York, NY: Routledge.

Wittek, A., & Tomasello, M. (2002). German children's productivity with tense morphology: the Perfekt (present perfect). *Journal of Child Language, 29*, 567-589.