Experimental methods in psycholinguistics

Constanze Vorwerg

1. The rationale behind the experimental approach

To study language production, language comprehension, language representation and language acquisition means largely to ask what processes are involved, what mechanisms underlie and what factors contribute to the phenomena explored. Therefore the experiment is one of the most important methodological approaches within psycholinguistics. Its basic idea is to systematically manipulate certain variables while controlling for the effects of others – as a way of dealing with the problem of scientific explanation, as opposed to the problem of description (which is tackled by other methods). If we are to find out how and why something occurs, we need to compare different conditions with respect to their effects on the phenomena under investigation – and that's what experiments are about.

2. Definitional criteria of an experiment

Experiments study the effects of one variable on another. The main idea is to systematically vary conditions in order to determine their effects onto measures of language processing. Those variables which are measured (in dependence on the experimental conditions) are called dependent. Those which are manipulated directly by the experimenter and constitute the experimental conditions are called independent. The aim is to reveal causal or conditional relations between independent and dependent variables. Accordingly, the idea is to deliberately produce occurrences (cf. already Wundt 1896). That is, the emergence or progression of the phenomenon under investigation should depend on the voluntary impact of an experimenter. Therefore, an important criterion of an experiment is that the experimental manipulation is voluntarily brought about.

Let's say we are interested in factors influencing the resolution of ambiguous pronominal anaphora. Possible factors proposed include gender and number agreement, semantic consistency, first mention, recency of mention, grammatical-role parallelism, subjecthood/topichood, occurrence
in main clause, and many more (cf. Kehler, Kertz, Rohde and Elman 2007; Mitkov 2002; Song and Fisher 2007). To test the possible factor of grammatical-role parallelism, one may compare sentences with a subject pronoun in the final clause, see (1), vs. those with a pronoun in another grammatical role, see (2) – both having two possible antecedents in the first clause (examples from Smyth 1994).

(1) a. Gloria tickled Emma and then she poked Mike. [subject]
   b. Sue gave Maureen a new watch and then she gave Joe a record. [subject]
   c. Richard chased Jim around the corner and then he chased Caroline to Yonge Street. [subject]

(2) a. John pushed Sammy and then Evelyn kicked him. [direct object]
   b. Tommy told Kevin to find the new bikes and then Samantha asked him to find the new tires. [indirect object]
   c. Adam argued with Barry about the new format and then Linda argued with him about the new rules. [prepositional object]

If we ask the participants of such an experiment to indicate whom the pronoun in each sentence refers to, we may compare the proportions of answers referring to antecedent candidate 1 (subject assignment) vs. antecedent candidate 2 (nonsubject assignment) between conditions. This was done by Smyth (1994), who found that the proportion of subject assignment was significantly higher (in fact 1.0) for subject pronouns than for nonsubject pronouns (.12). The grammatical roles of nonsubject pronouns were direct, indirect or prepositional object – in each case corresponding to one of the nonsubject NPs in the preceding clause. So, the dependent measure in this experiment was proportion of subject assignment, and the independent variable was grammatical role of the pronoun (subject vs. nonsubject). Importantly, pronoun assignment was a function of the pronoun’s grammatical role when the sentences were fully parallel (e.g., with a dative-object pronoun having a dative antecedent, or a pronoun within a prepositional phrase having an antecedent within a prepositional phrase, etc., plus constituent structures, thematic roles and attachment site being parallel across clauses); however, this was not the case when they were not fully parallel, with nonsubject pronouns and potential antecedents having different grammatical roles. Thus, the phenomenon studied was brought about by the experiment-
er by constructing fully parallel structures and manipulating the anaphor’s grammatical role.

To make sure, however, that a phenomenon (in our example: pronoun assignment) has really arisen from or been modulated by our experimental manipulation, we need to control other potential influences. And this is the second definitional criterion of an experiment. The experimental manipulation must be sufficiently separated from other potentially influential factors – and not be confounded with them. In our introductory example (Smyth 1994), several aspects were held constant, such as gender and number agreement between the pronoun and potential antecents, type of pronoun, constituent structure and attachment sites (grammatical parallelism), and temporal conjunction. Those variables are called control variables, as these are the variables which we control, i.e. whose influence we eliminate. There are several ways to achieve control over them, which together with the variation and combination of the independent variables as well the way of drawing samples constitute experimental designs (see Section 4).

Finally, replicability is an important feature of experiments, as it allows determining how reliable and precise the conditions of a phenomenon can be specified as a first estimate of the validity of a phenomenon (Sprung and Sprung 1984), which would not be possible with singular observations. In addition, experimental results can be reproduced by other researchers.

3. Research hypotheses and the wheel of science

3.1. Hypotheses, predictions and theories

The purpose of psycholinguistic experiments is to verify hypotheses about human language processing. Hypotheses concern the relation between independent and dependent variables (see 2); they are tentative, verifyable answers to research questions. So, for example, the question how lexically ambiguous words are interpreted in context could tentatively be answered by assuming (i.e., predicting) that only contextually appropriate word meanings are activated (e.g., Glucksberg, Kreuz and Rho 1986). Other examples of research hypotheses are: nine-month old infants are sensitive to phonotactic patterns in their native language (e.g., Jusczyk, Luce and Charles-Luce 1994); the verb used in a placement instruction influences the interpretation of spatial prepositions (Vorwerg and Weiß 2010).
Research hypotheses are at the very heart of the experimental method. Performing experiments means basically testing hypotheses about the effects of independent variables onto dependent variables. Where do the hypotheses to be tested come from? They can arise out of singular observations, practical needs, qualitative research, or pilot studies. However, more often than not they are derived from psycholinguistic theories and models of language processing.

To give an example, the Neighborhood Activation Model of spoken word recognition predicts that word recognition would be inhibited by an immediately preceding phonetically similar word and that low frequency primes would provide relatively more inhibition than high frequency primes (Goldinger, Luce and Pisoni 1989). These predictions derive from the model’s assumption of competition among phonological neighbors (phonetically similar words). Specifically, the model postulates that (1) upon presentation of a word a set of acoustic-phonetic patterns are activated, (2) acoustic-phonetic patterns which correspond to words activate word decision units, and (3) once activated, the word decision units monitor higher-level lexical information, such as word frequency, which adjust the activation levels in the decision units (Luce and Pisoni 1998). For determining whether a particular word has been present, this word's frequency-weighted probability is set into relation to the summed frequency-weighted probabilities of all words in the similarity neighborhood. Therefore, increasing the activity of words in the neighborhood is predicted to reduce the accuracy of stimulus identification. And one way to enhance the activation level of a stimulus word’s neighborhood is the preceding presentation of a phonetically related word (i.e., one of its neighbors), as the activation of a word is known to not dissipate immediately. So, the residual activation of the neighbor’s word should produce increased competition from the word’s neighbourhood reducing identification performance. And if the preceding word (the prime) is of low frequency, it should be identified less accurately and less quickly, producing more residual activation leading to relatively more competition than high frequency primes.

So both predictions derive from the model’s main postulates. Hence, this procedure corresponds to the deductive form of reasoning described (and favored) by Einstein (1919) as a way for gaining scientific knowledge: a set of hypothetical postulates is built which seems suitable to explain a number of facts, and the consequences deductively drawn from those postulates can then be tested empirically providing a criterion for the validity of the proposed postulates. More generally, deductive reasoning is involved in
the experimental method as hypothesis testing usually means putting a logical consequence (entailment) of a research hypothesis to the test. At the same time, inductive reasoning is involved in the experimental method – both in the generation of hypotheses or postulates and in the inferential-statistical analysis of results. The generation of hypotheses (if not derived from more basic postulates) often includes inferring general principles which can explain particular observations, or patterns observed in earlier research. In data analysis, inferential statistics (also called inductive statistics) is used to generalize from small samples to populations (see below).

3.2. Operationalization and measures

To test experimental hypotheses, they need to be operationalized, i.e., transformed into concrete, observable events and measurable response variables. Going back to the anaphora example in Section 2, a particular language has to be chosen, verbs with suitable verb frames must be found and the according arguments, words, names need to be selected and sentences constructed. The sentences have to be presented in a certain form, and measures are needed which serve as the dependent variables.

The parameters used for capturing aspects of human language processing include both offline and online measures. Offline measures register the results of language processing after completion of the process under investigation. Examples include frequency of production or selection of a linguistic form or structure, ratings of suitability or grammaticality, and psycholinguistic test performance.

Online measures in contrast tap into the involved processes and representations in the course of processing, i.e., while language is constructed in real time. More often than not they are time-sensitive and indirect – sometimes supplemented by spatial information. Examples include reaction time measures reflecting aspects of processing such as difficulty or activation, discrimination or confusability data, eyetracking techniques, as well as neuroimaging and neurophysiological measures. In addition to providing insight into processing steps, online measures are more sensitive with respect to group differences (MacWhinney, Feldman, Sacco and Valdés-Pérez 2000) and more specific with respect to the mechanisms involved (cf. Kuperberg 2010; Peelle, Cooke, Moore, Vesely and Grossman 2007; Tompkins and Baumgaertner 1998).
3.3. Data analysis

The results of an experiment performed are subject to data analysis using inferential statistics. The rationale behind performing inferential statistics is to determine whether, in the population, the independent variable (in terms of the values that it takes in the experiment) has differential effects onto the dependent measures. There are two important aspects involved here.

First, the experimental hypothesis predicts a difference between conditions (i.e., levels of the independent variable). This conforms to the “method of differences” proposed by John Stuart Mill (1843: p. 455) as a method of experimental inquiry:

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance save one in common, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or a necessary part of the cause, of the phenomenon.

Second, data analysis refers to the population from which a particular sample was drawn, and the question is whether a particular difference found in the sample can be generalized to the whole population.

As there is usually some variation in language processing and language materials due to circumstances not relevant to and not controlled for by the experiment (such as fluctuations in attention, concepts activated shortly before the experiment, etc.), some difference between conditions can always be expected to occur – even when the experimental manipulation actually has not the predicted effect. So, one way to determine statistically whether there is an effect in the population or not is to calculate how probable the difference observed in the sample might have occurred due to some random variation in the data (i.e., pure chance) – instead of being an effect of the experimental manipulation.

This is called the “null hypothesis”: the statistical hypothesis that there is no real difference in the population with respect to the experimental manipulation studied. On the basis of a statistical model and using an appropriate statistical test, the probability can be computed to obtain a difference like the one observed (or larger) even though the null hypothesis is true, i.e., differences are due to chance. The “alternative hypothesis” – that there is a real difference in the population with respect to the experimental manipulation is judged as being confirmed when the probability of obtaining a difference at least as large as the one observed by chance is very small and
the difference or effect found is then said to be “statistically significant”. Often a criterion of 5% is adopted as significance level.

Null hypothesis significance testing is the statistical method used most frequently for statistical inference, even though it has been frequently criticized, mainly on logical grounds (see Krueger 2001, for a review) and a number of alternatives have been proposed, including “effect size” and “confidence intervals” (see Denis 2003, for a review). However, other researchers defend null hypothesis testing and refute its criticisms; moreover, as pointed out by Krueger, its pragmatic value might be an important reason for its continued use. In addition, the probability of effects being true for a population can be further enhanced by replicating experiments.

When using significance testing, an important aspect to keep in mind is that sampling is twofold: There is a sample of participants and a sample of language materials. Findings should generalize beyond both types of samples, that is, to both the linguistic-community population (such as speakers or learners of a language, bilingual infants, etc.) and the language population (such as words, nouns, sentences of a language, etc.). Therefore, if language materials assigned to participant groups are sampled from a larger population of items, it is not correct to regard the materials variable as a fixed factor and to ignore the sampling variance – a way of analyzing language-related data criticized by Clark (1973) as the language-as-fixed-effect fallacy. Possible solutions are using a statistical procedure which treats language materials as a sample (see Clark), or controlling for linguistic item variance by experimental procedures such as matching or counterbalancing (Raaijmakers, Schrijnemakers and Gremmen 1999; see Section 4 for a description of experimental designs).

An alternative also discussed by Clark (1973) is the “method of single cases of language material” suitable to test hypotheses which can be applied to single words (or other kinds of language materials) in contrast to those which require a comparison of central tendencies. For example the verb used in a placement instruction – and even a single verb particle (German *ein- ‘in’) – can have an impact on the interpretation of perspective-dependent spatial prepositions (Vorwerg and Weiß 2010).

3.4. Interpretation and theory and novel predictions

After performing the data analysis, the findings are interpreted with respect to their bearing on the original research hypotheses, predictions and theory.
New predictions may be derived, the theory may be altered, deductive reasoning plays a part again – closing the circle of science (cf. Wallace 1971).

Typically, psycholinguistic data interpretation, theory and modeling based on experimental findings relate to functional or mechanistic explanations of phenomena in language processing. Whereas *functional explanations* describe causal relations in terms of why a particular psycholinguistic phenomenon (such as mutual adaptation in dialogue) might be useful and what functions it might fulfil, *mechanistic explanations* describe causal relations in terms of the processes and mechanisms involved (see Bechtel 2005, for a discussion of mechanistic explanations; an example is the interactive alignment account by Pickering and Garrod 2004).

An important aim in experiment-related theory development is to also be able to explain previous experimental results; and the plausibility of a theory is enhanced if it can convert a variety of experimental results into a few underlying general principles, even more so if it can reconcile contradictory results obtained in different studies. A next step is then again to derive novel observational predictions and to test them experimentally. For example, a coherence-of-discourse theory has been proposed (Kehler et al. 2007) to explain a number of factors described in the literature for anaphora resolution, including the grammatical-subject preference and grammatical-role parallelism (cf. Section 2). New evidence is then presented in support of the theory, which includes a neutralization of grammatical subject and parallelism preferences by controlling for coherence. So, data interpretation tends to relate to previous findings if possible, and a theory’s explanatory value is relatively stronger if it offers a *single mechanism* for a variety of empirical data, even more so if it can explain otherwise mystifying results.

By the same token, a theoretical account is strengthened by *converging evidences* from independent scientific sources of information. For example, as argued first by Marshall and Newcombe (1973; cf. Vorwerg 2010) in their pioneering analysis, the errors described for clinical patterns of acquired dyslexias are consistent with results from the general word recognition and reading research, such as types of reading errors in children learning to read or adults reading under time pressure, or also word frequency effects. More recent research endeavours have explicitly focused on how converging approaches and sources of evidence for understanding both normal reading and its breakdown can lead to new insights (e.g., Klein and McMullen 1999).

One important aspect addressed in data interpretation is the question of their generalizability and what limitations or qualifications there might be.
For example, the factors of anaphora resolution mentioned above (see Section 2) refer to unaccented (third-person) pronouns, as contrastive stress is known to affect coreferentiality (e.g., Akmajian and Jackendoff 1970). Conversely, the generalizability of an account is corroborated and extended by recurrent findings, which are obtained under different circumstances.

The interpretation of experimental results often involves reasoning over indirectly related effects (especially with online measures, cf. 3.2) in order to be able to gain insight into processing aspects which are not directly observable. For example, differential priming and interference effects with phonologically vs. semantically related items (see 5.1) are used to draw conclusions about the time course of processes in word production. Another example is eye movement data, which are interpreted with respect to different aspects of language comprehension or production (see Section 5).

Reasoning based on indirect observation is also usually involved in attempts to contribute to the fundamental issues debated in psycholinguistics, such as the question of whether (or when) processing is parallel vs. serial, and whether (or when) it is interactive vs. autonomous.

Parallel vs. serial processing refers to the question of whether several elements of input (such as features, letters, words in reading) or competing candidates of internal output (such as words in word recognition, word meanings in resolving lexical ambiguity or sentence structures in resolving structural ambiguity), or sequential elements of output (such as phonemes, syllables, words, constituents in speech production) are processed or represented at the same time, i.e. in parallel, or one after the other, i.e. serially. The term is also used to refer to the question of whether certain subprocesses (stages, subsystems, etc.) occur one after the other or in parallel.

One method of testing parallel vs. serial processing is error analysis. For example, blending errors in speech production, see (3), suggest that alternative forms or speech plans are activated concurrently (example from Harley 1984). Another example is transposition errors in typing, writing or speech, see (4), suggesting simultaneous output activation or display in the output buffer (examples from Bawden 1900, cited in MacKay 1970).

(3) a. Speech Plan 1: THE SKY IS BLUE.
   b. Speech Plan 2: THE SUN IS SHINING.
   c. Actual Utterance: The sky is shining.

(4) a. Wasserflasche → Flascherwasche
   b. Mond und Sonne → Sond und Monne
Other methods of testing parallel vs. serial processing use reaction-time or accuracy measures (see Townsend 1990, for an overview and a discussion). Methods based on accuracy may examine whether delimiting the time available for processing leads, e.g., to a reduction in accuracy (to be expected if one item is processed after the other), or else a large number of confusions between similar items (only to be expected if a number of items is simultaneously in a partially processed state). One example for such a result is the word superiority effect in letter recognition. When participants are asked to report a letter after a very brief presentation, they are better able (i.e., more accurate) if the letter is embedded in a word than when it is isolated (or surrounded by arbitrary letters). A serial processing of letters would predict a better performance for a single letter as the average amount of processing per letter would decrease with an increasing number of letters to be processed in the same, short amount of time. However, accuracy is not only unimpaired by a larger number of letters, it is better for letters within meaningful words. Importantly, this word superiority effect is present at all word positions at once, which means that each letter is, in its recognition, affected by all other letters, while at the same time also affecting them (Wheeler 1970). This is evidence of parallel letter processing and also of an interaction (see below) between letter and word recognition.

Methods based on reaction-time measures are most frequently used. One way to inspect data for parallelism vs. seriality is a comparison of reaction times for a different number of items to be processed. Seriality predicts a linear relationship between “mental load” (number of items to be processed) and response time, as argued by Sternberg (1966). By the same token, flat reaction-time curves are indicative of parallel processing. The problem is that linearly increasing reaction time curves can not only be explained by serial processing but also by other types of limitation in capacity including limited-capacity parallel models (Townsend 1990). However, inferences about parallel processing can validly be drawn, if reaction time remains constant or even gets faster with increasing mental load.

Another way to determine seriality vs. parallelity is the use of factorial designs (see also Section 4). If two or more factors (e.g., word frequency and regularity) known to have effects on reaction time and assumed to selectively affect distinct subprocesses are manipulated at the same time, their effects should be additive if they function serially. Therefore, subadditivity – a kind of interaction between factors, in which the increase in reaction time due to both factors is smaller than the sum of the single effects (see Figure 1) – signals parallel processing (e.g., Townsend 1990).
A further approach is the method of redundancy gain. When one target or one dimension level or one processing route is sufficient for making a response, but a second one equally sufficient for making a correct response is also present or assumed to be available (response “yes”, if either of them or both are present), the question is what effect this redundancy has on response time. Examples include color and color words (Eidels, Townsend and Adams 2010), visually and auditorily presented words (Lewandowski and Kobus 1993), and semantic and asemantic retrieval routes for Arabic digits (Campbell and Metcalfe 2008). Frequently, a reduction in reaction time is observed with redundant targets or target dimensions. This is interpreted as evidence for parallel processing – with either separate activation or coactivation (e.g., Townsend and Nozawa 1995). In the case of separate processes, processing is executed separately within each process producing its own activations, and the gain in response time can be explained by a “race” (or “horse race”) between both processes, which terminates as soon as one of them is completed. As the response time needed varies stochastically, the probability of a shorter reaction time is larger with two targets or processes. Therefore, this purely statistical effect is called “probability summation”. In contrast, in the case of coactivation both processes feed into one decision mechanism before a decision is made. As their parallel activations are summed together in one decision mechanism, a decision can be made faster, leading to the redundancy gain with two entities affording the same response being present at the same time. Another way to explain the redundancy gain is the assumption of a crosstalk (mutual impact on processing) between both channels. If they have mutual access to partial
information and influence each other during processing, this would also lead to shorter responses in the sense of a redundancy effect – both with coactivation or with separate activation (see Miller 1991; Townsend and Nozawa 1995). In contrast to separate activation (without crosstalk), both coactivation in one common decision mechanism and crosstalk before activation summation explain redundancy effects by a cooperation between both processes before a decision is made – some kind of interaction in the broad sense of the world.

Interactive vs. autonomous processing refers to the question of whether the modules, subsystems or subprocesses involved work independently, producing separate results which only after completion can be compared or combined or passed on, or whether their processing works together in that their temporary products of processing before completion combine in some way – either by informing each other or by feeding into one mechanism such that their output is a conjoint result. As can be seen from the preceding discussion, with respect to subprocesses this issue is often intertwined with the issue of serial vs. parallel processing – however, parallel processing is in principle also possible without crosstalk or coactivation of subsystems.

One type of interaction concerns levels of processing. Interaction between levels of processing occurs when there is cascading of activation to a subsequent level before processing is completed at the previous level, or when there is feedback from logically later processes onto those logically preceding them. An example is the above-mentioned word-superiority effect which can be explained by a feedback from the lexical (the word) level to the letter level before letter recognition is completed (as assumed in the interaction activation model of letter recognition by McClelland and Rumelhart 1981). One methodological approach to gain support to interactive processing, followed by Wheeler (1970), is to systematically rule out all conceivable hypotheses for separate mechanisms.

The question of whether there is interaction between levels of processing is also hotly debated for language production. Specifically with respect to lexical access (i.e., accessing a word in the mental lexicon), the question is whether its two components – lexical selection and phonological encoding (see, e.g., Levelt 1999) – occur in discrete, non-overlapping stages or whether there is an interaction in terms of some phonological information already being activated during semantic-syntactic word access and some semantic information being activated during phonological access (see, e.g., Dell and O'Seaghdha 1992). One method used to find evidence on this question is an interference paradigm in which pictures are named in
the presence of either semantically or phonologically related words while varying the relative onsets of pictures and words, allowing for an analysis of the time course of semantic vs. phonological information being activated during word production (Schriefers, Meyer and Levelt 1990; see 5.1).

Another example of findings interpreted with respect to the question of interaction between processing levels in production are so-called "mixed errors". These are substituting speech errors – normal slips of the tongue or aphasic naming errors – which are related to the target word both in form and semantics, see (5). In an interactive production model (e.g., Dell, Chang and Griffin 1999), such a formally and semantically related word gains activation from both shared semantics (via shared semantic features) and shared phonemes (via feedback from phonemes shared with the target). So, if the probability of producing a certain word substitute depends on its being related to the target word both phonologically and semantically, both types of information can be assumed to be activated at the same time, suggesting an interactive account (but see Levelt 1999).

(5)  a. cat → rat  
     b. snail → snake

Another type of interaction concerns processes working in parallel and combining their results. Interaction between those processes occurs when they operate jointly to arrive at a certain decision or representation by either coactivation (feeding into the same decision mechanism), or crosstalk between channels (informing each other), or both. The redundancy gain discussed above as a criterion of parallel processing provides also a means to decide between a race and a coactivation model – as a race model (see above) can only account for redundancy gains up to a certain amount. As such a model explains the redundancy gain as a statistical effect resulting from stochastical variation in response time, according to this account the distribution function for redundant trials cannot be larger than the sum of the contribution functions for trials in which only a single target is present. If this inequality predicted by the race model ("race-model inequality") is violated, race models can not explain the redundancy gain in reaction time (see Miller 1991). Therefore, it can be concluded that some kind of interaction occurs: crosstalk or coactivation or both (see Townsend and Wenger 2004, also for an overview of other RT inequalities proposed in the literature).
A type of data which cannot be explained by coactivation alone are congruence effects (Miller 1991). If reaction time does not only depend on redundancy (presence of two entities each of which licence a "yes" response), but additionally on the relatedness between entities to be judged, this cannot be accounted for by a joint summation of independently achieved activations. If a contingency between both entities, such as a high joint probability or a congruity in content (e.g., correspondence between location of a visual stimulus and pitch of an auditory stimulus, both being HIGH or LOW), leads to faster responses, this can be accounted for in terms of coactivation with crosstalk (Miller 1991) or separate activation with crosstalk (Townsend and Wenger 2004).

4. Experimental designs

The basis for drawing conclusions from experimental evidence is the experimental design used – the method chosen to manipulate independent variables and to control for others. It affects both the techniques which can be applied for inferential statistics and the logical inferences to be drawn (together with other factors, such as coherence of data patterns, convergence of results from different sources, etc., see 3.3).

One aspect is the question of whether unrelated or related samples are used. Unrelated samples are those, which are drawn independently for the different conditions to be compared. Related samples are those, in which the same participants are used in all of the conditions, such that the values measured in the different conditions are related because they stem from the same participants. If there is more than one independent variable involved, there is also a mixed design possible. A design, in which unrelated samples are drawn, is also called a between-subjects design. A design, in which related samples are drawn, is also called a within-subjects design.

Another aspect of the experimental design is the question of how many factors (independent variables) and factor levels are compared. The simplest case is the use of just one independent variable. However, if we are interested in possible interactions between variables, we need to combine several factors. A factorial design is one in which all factor levels are crossed. It is usually described by specifying the number of levels for each of the independent variables used. For example, in a 2x2x3 design, the first variable has two levels as well the second one, and the third has three levels. The combinations of the factor levels used are the conditions of the
experiment to be compared. So, in the case of the 2x2x3 factorial design, we have 12 conditions altogether. In some experiments, one condition is regarded as a control condition relative to which experimental conditions can be judged. For example, in order to establish the effectiveness of a bilingual curriculum in preventing reading problems, an appropriate control group (without bilingual curriculum) is needed. In single-case experiments, there might also be a "control variable" included to distinguish treatment effects from spontaneous recovery or change.

Another aspect of the experimental design is the control of those variables whose influence should be eliminated, sometimes called control variables. One technique used to control for order effects is counterbalancing, which means to use all possible orders of conditions or items (e.g., ABC, ACB, BCA, BAC, CAB, CBA) an equal number of times. If there are too many conditions or items to counterbalance, a frequently used alternative technique is randomizing, which means to use a random sequence each time the experiment is performed. Other options are a latin square, according to which each experimental block or condition occurs at each possible position (e.g., ABC, BCA, CAB) an equal number of times, or an inversion (ABC, CBA). Other variables which might possibly influence experimental results can be controlled by matching them (e.g., age, gender) between conditions or by eliminating them (e.g., background noise).

The combination of factor levels, control techniques and sampling yields the experimental design. An equal number of participants and/or language materials is assigned to each cell in the experimental design. In a true experiment, participants and/or language materials are randomly assigned to conditions, as it is a central feature of experiments that the factor levels are under the control of the experimenter (cf. Section 2). However, sometimes such a random assignment to conditions (i.e., arbitrary control of the independent variable) is not possible; if instead independent variables are used whose levels cannot be controlled (e.g., native language, bilingualism, etc.; for a review of experimental psycholinguistic methods for studying language disorders see Vorwerg 2010), this is called a quasi-experiment. The interpretation of quasi-experiments is more difficult, because it is often unknown what other variables might be confounded with a quasi-experimental variable, and also because a causal relation is not easily established. Therefore, it can sometimes be useful to have several control groups in order to enhance the probability of a causal relationship (e.g., a comparison of children with reading disorder with both an age-matched and a reading-level-matched control group).
5. Specific experimental paradigms

5.1. Experimental methods used in studies of language production

There is a multitude of experimental paradigms the use of which depends on the particular question investigated. In order to study the conditions under which a certain variant is produced, frequencies of use are measured (e.g. in studies of object naming or a verbal localization). These can also be compared and related to other types of data, such as a rating of typicality, goodness, validity or usability of a linguistic expression and also response time for verification (see Vorwerg 2001).

In language acquisition research (see Eisenbeiss 2010, for an overview) children may also be asked to imitate spoken language; from the way they reconstruct the utterance, conclusions about their grammar can be drawn. In other studies, the production of questions, negated sentences or plural forms may be elicited by according questions, sometimes including novel words (e.g., This is a wug. These are two ...?; Berko 1958; Chan, Lieven, and Tomasello 2009). A further example is speeded production experiments, in which learners produce complex forms as fast as possible and onset time is measured in order to get insight into whether forms are holistically retrieved from mental lexicon or assembled during production.

A different method, which is frequently used, – both with children and adults – is structural priming. This term refers to speakers’ tendency to repeat linguistic structures heard or produced before (see, e.g., Branigan 2007; Vorwerg 2009, for self-priming). So, presenting a participant with a certain linguistic structure may possibly influence the probability that the same type of structure will be produced. This type of priming effect can be used to study interaction in dialogue, the mapping from message to syntax, levels of grammatical processing, the representation of linguistic knowledge, bilingual processing, and relations between written and spoken language as well between production and comprehension (see Pickering and Ferreira 2008, for an overview).

Furthermore, phonological and semantic priming or interference can be used to study the time course of word production (see also 3.4). In word-picture interference tasks, a word presented with a picture slows down picture-naming latencies – depending on the relation between word and picture name. In priming studies related to production, a previously presented stimulus (e.g. a word) affects the subsequent picture or word naming, again depending on the relationship between prime and target (e.g., semantic,
Experimental methods in psycholinguistics

The time course of facilitation and inhibition is used to draw conclusions about activations and units used in language production.

Another methodical development used to study language production is eyetracking, the continuous registration of gaze and eye movements and their analysis in relation to speech produced. In word production related to visual objects, gaze is assumed to reflect word preparation and therefore used to study aspects of this, such as the locus of (self-corrected) speech errors (see Griffin 2004).

5.2. Experimental methods used in studies of language comprehension

Experimental methods used for examining language comprehension vary widely, depending also on the area of study. Examples of methods used to study word recognition include gating paradigms (a repeated presentation with increasing presentation time of a word to be “guessed”), presentation in noise, phoneme monitoring, shadowing of continuous speech and detection of mispronunciation (see Grosjean 1980, for the gating paradigm and an overview of other paradigms).

Phonological, orthographic, morphological, semantic and associative priming, including crossmodal priming, can be used to study aspects of word recognition (specifically factors influencing it) and the time course of activation (e.g. the activation of word meanings of ambiguous words with different semantic contexts used as prime). Priming and interference paradigms do also offer a possibility to examine aspects of bilingual processing.

In reading research, reading times and eyetracking data may be analyzed with respect to difficulty of processing, the visual input in reading and specific aspects of sentence or text processing, such as parsing (analysis of syntactic structure). Several eyetracking parameters have been developed referring to either single words or regions, some of which are first-fixation duration, gaze duration, skippings, regressions, first-pass reading time, second-pass reading time and total reading time (e.g., Rayner, Slattery, Drieghe and Liversedge 2011).

Syntactic processing can also be studied by syntactic priming, grammaticality judgments, measures of sentence comprehension or the word detection procedure (in which participants press a button as soon as they hear a certain target word in a sentence – a response which is delayed after gram-
mathematical-agreement violations; e.g., Grossman, Lee, Morris, Stern, and Hurtig 2002).
Syntactic as well as semantic and pragmatic processing are further addressed by a number of other experimental techniques, many of which are also used in *language acquisition research* (see Schmitt and Miller 2010), e.g., truth-value judgment tasks (judged generally or in relation to particular situations) and verifications times, sentence-picture matching tasks, enactment paradigms (see also Vorwerg and Weiß 2010) and semantic or structural priming.

The *visual-world paradigm* (e.g., Dahan and Tanenhaus 2002) is another way to use *eyetracking* for the study of language comprehension (see Sedivy 2010, for a discussion of the assumptions behind this use of eyetracking). In this paradigm, eye gaze to objects in a display is analyzed during the processing of speech. As eye movements and speech processing are tightly coupled in situations where language refers to the visual surrounding, a number of issues can be addressed, including word recognition, syntactic analysis and the lexical activation of perceptual features. Typically, appropriate distractor objects (relative to a target object or word) are present in the display, such as those with a similar shape, shared word onset in the name or constituting referential ambiguity, such that gaze trajectories and fixation probabilities of distractor objects allow insights into speech processing mechanisms.

These are just a few examples of experimental paradigms used in psycholinguists. Others, which cannot be addressed here for limitations of space, deal with aspects of prosody, phonotactics, nonverbal communication, pragmatics, figurative language, discourse processing and dialogue, etc. Moreover, a number of different experimental approaches have been developed in order to tackle the question of interfaces and the interplay between different levels of processing discussed in 3.4.

### 6. Final remarks

The experimental approach provides a valuable tool for studying language processing, and a large number of elaborate and sophisticated techniques have been developed to address research questions which are difficult to tackle. The results of this type of research can fruitfully be compared with research from other methodological approaches and related to studies from other research areas.
The experimental approach may also directly be combined with other methodological approaches – by informing them (e.g., computational modeling, clinical research), by using their data as a basis (e.g., computerized lexicons, word frequency data from corpus linguistics), or by doing joint research and a convergent data analysis (e.g., discourse analysis, see Vorwerg and Tenbrink 2007).

While the experiment plays a central role in much of the psycholinguistic research – benefiting also from its interdisciplinary basis –, there are other areas of linguistic enquiry, in which the experiment is also of importance. Specifically, phonetics is a well-established experimental discipline; but experiments are also used in sociolinguistics (e.g., Auwärter 2005) and experimental pragmatics (cf. Noveck and Sperber 2004). Other linguistic disciplines are very much related to psycholinguistics, such as clinical linguistics, computational linguistics and neurolinguistics. Some research areas, such as the relation between language and thinking, as exemplified by the hypothesis of linguistic relativity (cf. Werlen 2002), language attitudes (see Werlen 2007), or bilingualism and language competences (e.g., Werlen, Rosenberger and Baumgartner 2011; Werlen, Tunger and Frei 2010) lie thematically at the intersection of psycholinguistics with other linguistic disciplines. And currently there also some endeavours for joint research programmes with respect to questions of language variation and language change between psycholinguistics, sociolinguistics, historical linguistics, perceptual dialectology, and cognitive linguistics (cf. Vorwerg, in prep.).

References

Akmajian, Adrian, and Ray Jackendoff

Auwärter, Manfred

Bechtel, William

Berko, Jean
Branigan, Holly

Campbell, Jamie I. D., and Arron W. S. Metcalfe

Chan, Angel, Elena Lieven, and Michael Tomasello

Clark, Herbert

Dahan, Delphine, and Michael K. Tanenhaus

Dell, Gary S., Franklin Chang, and Zenzi M. Griffin

Dell, Gary S., and Padraig G. O'Seaghdha

Denis, Daniel J.

Eidels, Ami, James T. Townsend, and Daniel Adams

Einstein, Albert

Eisenbeiss, Sonja
Goldinger, Stephen D., Paul A. Luce, and David B. Pisoni

Glucksberg, Sam, Roger J. Kreuz, and Susan H. Rho

Griffin, Zenzi M.
2004 The eyes are right when the mouth is wrong. *Psychological Science* 15: 814–821.

Grosjean, François

Grossman, Murray, Christine Lee, Jennifer Morris, Matthew B. Stern, and Howard I. Hurtig

Harley, Trevor A.

Jusczyk, Peter, Paul A. Luce, and Jan Charles-Luce

Kehler, Andrew, Laura Kertz, Hannah Rohde, and Jeffrey L. Elman

Klein, Raymond M., and Patricia McMullen (eds.)

Krueger, Joachim

Kuperberg, Gina

Levelt, Willem J. M.

Lewandowski, Lawrence J., and David A Kobus
Luce, Paul A., and David B. Pisoni  

MacKay, Donald G.  

MacWhinney, Brian, Heidi Feldman, Kelley Sacco, and Raúl Valdés-Pérez  

Marshall, John C., and Freda Newcombe  

McClelland, James L., and David E. Rumelhart  

Mill, John Stuart  

Miller, Jeff  

Mitkov, Ruslan  

Noveck, Ira A., and Dan Sperber (eds.)  

Peelle, Jonathan E., Ayanna Cooke, Peachie Moore, Luisa Vesely, and Murray Grossman  

Pickering, Martin J., and Simon Garrod  

Pickering, Martin J., and Victor S. Ferreira  

Raaijmakers, Jeroen G. W., Joseph M. C. Schrijnemakers, and Frans Gremmen  


Townsend, James T., and Michael J. Wenger

Vorwerg, Constanze


in prep. Cognitive mechanisms and social factors in language variation.

Vorwerg, Constanze, and Thora Tenbrink

Vorwerg, Constanze, and Petra Weiβ

Wallace, Walter

Werlen, Iwar


Werlen, Iwar, Lukas Rosenberger, and Jachin Baumgartner

Werlen, Iwar, Verena Tunger, and Ursula Frei

Wheeler, Daniel D.

Wundt, Wilhelm