Usage statistics, semantic transparency and segmentability in the selection, access and (de)composition of complex words

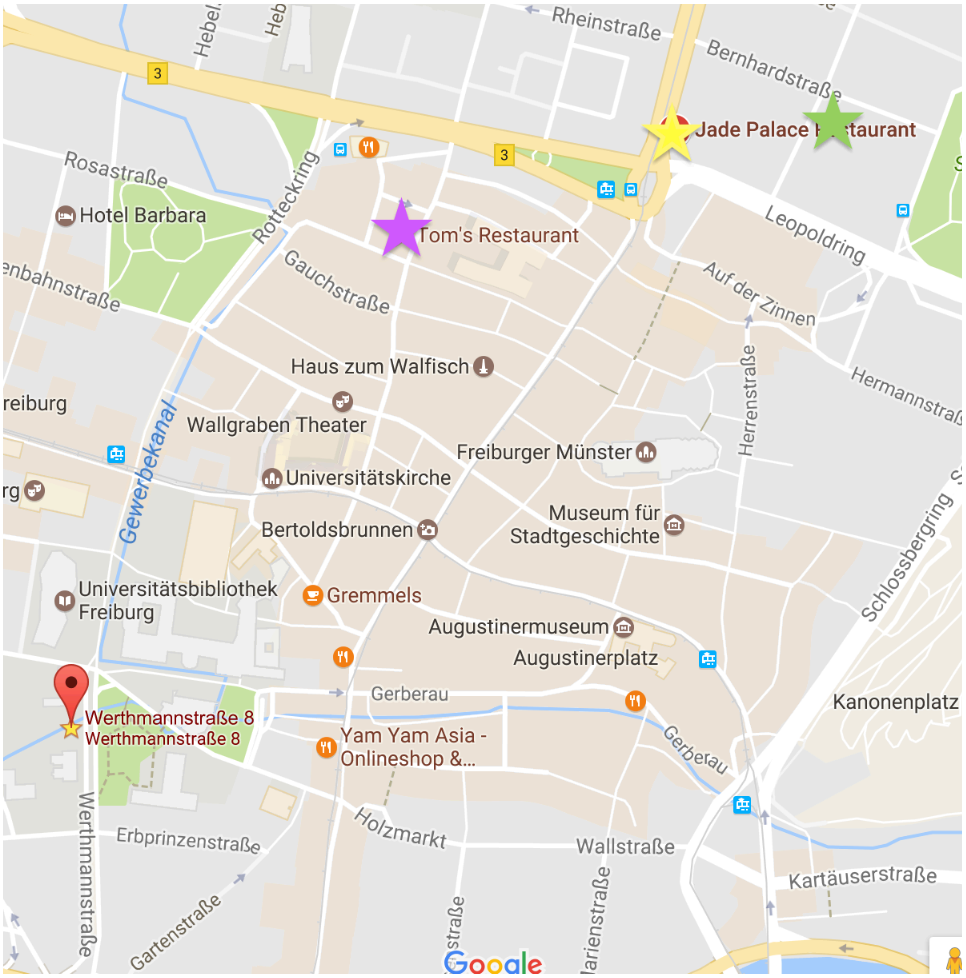
workshop booklet

University of Freiburg, May 4–6, 2017
Großer Sitzungssaal, Werthmannstraße 8

UNI FREIBURG

frequenz
effekte
graduiertenkolleg 1624

Important locations



— Workshop venue: **Großer Sitzungssaal** (Werthmannstraße 8)

— Guest accommodation: **Stadthotel Freiburg** (Karlstraße 7) ★

— Dinner / drinks (May 4, 19:30): **Jade Palace** (Habsburgerstraße 133) ★

— Dinner / drinks (May 5, 19:30): **Tom's** (Unterlinden 3) ★

The organizers would like to thank the DFG and the committee of the GRK 1624 "Frequency Effects in Language" for making the workshop possible.

Workshop abstract: In recent years, the study of morphological processing has been invigorated by an increased use of neurobiological methods, as well as an increased focus on typological variation. Many key research questions, however, remain controversial at least four decades after their systematic investigation began. Some of these questions include:

- Whether complex words are represented and accessed independently or as 'compositional' combinations of stems and affixes in the mental lexicon; whether both independent and compositional representations coexist, and, if so, which factors influence the choice of one route over another in lexical access
- Whether lexical access proceeds compositionally in the case of regular or default inflection but not in the case of irregular or non-default inflection, and which factors govern overregularization or analogical levelling
- Whether similar principles apply across in the processing of inflection, derivation and compounding, and whether, in the case of derivation and compounding, semantic opacity obliges full-form storage
- How quickly semantics can be activated relative to perceptual input onset
- Why processing phenomena such as morphological priming differ along typological lines

This workshop seeks some consensus on the questions outlined above. What is the best model that we currently have of the organization of the mental lexicon and the addressing of its units in production and comprehension? To what extent must such a model integrate the neurobiological evidence for two systems (a left-hemispheric 'combinatorial' network and a bilateral full-form retrieval network) in morphological processing, if at all? Finally, to what extent must the model be able to incorporate continuously-valued usage statistics (e.g. token and type frequencies, semantic transparency), in both individual variation and linguistic typology, to account for the array of effects known from various modalities and experimental paradigms?

Day I (Thursday, May 4)

from 10:00	Registration, welcome
10:30–11:00	Introduction
11:00–12:00	Luke Bradley (University of Freiburg) *Frequency and semantic transparency effects on word recognition in analytic and synthetic languages*
12:00–13:30	*Lunch*
13:30–15:00	Eva Smolka (University of Konstanz) *Who nicks the nickname?—The influence of frequency and semantic transparency on the processing of complex words*
15:00–15:15	*Coffee*
15:15–16:45	Mirjana Bozić (University of Cambridge) *Morphological functions in their neurobiological context*
19:30	*Workshop dinner (optional)*

Day II (Friday, May 5)

9:30–11:00	Tal Linzen (LSCP & IJN, École Normale Supérieure Paris) *Information and representations in the neurobiology of morphology*
11:00–11:15	*Coffee*
11:15–12:45	Adam Albright (Massachusetts Institute of Technology) *Pair-finding, segmentation, and morpheme convergence*
12:45–14:15	*Lunch*
14:15–15:45	Péter Rácz (University of Bristol) *Learning the English past tense with robots*
15:45–16:00	*Coffee*
16:00–17:30	Harald Clahsen (PRIM, University of Potsdam) *What language processing reveals about constraints on word formation*
19:30	*Dinner / drinks (optional)*

Day III (Saturday, May 6)

10:30–12:00 João Veríssimo (PRIM, University of Potsdam)
Generalisation and processing of 'pure morphology':
Evidence for two systems

12:00–13:00 *Lunch*

13:00–14:00 Carmen Pietropaolo (University of Freiburg)
Modelling and testing the productivity of morphophonological rules:
The case of Italian mood

14:00–15:15 Roundtable, close

Chairs: Alice Blumenthal-Dramé (University of Freiburg)
Verena Haser (University of Freiburg)

Abstracts (Day I)

Luke Bradley
Frequency and semantic transparency effects on word recognition in analytic and synthetic languages

I describe a series of lexical decision experiments from Vietnamese and German, focussing in particular on psycholinguistic claims that have been explicitly anchored in typological features of these languages. Specifically, I discuss a) anti-frequency effects of frequent syllables in Vietnamese, claimed to result from the high semantic burden of the typical syllable of this morphologically impoverished language; b) semantic transparency and opacity effects in German derivational and compound priming, whose apparent indiscriminability has been claimed to result from the morphological richness of this language; and c) regularity effects in German inflectional priming, which have been claimed to embody a classical dual-mechanism distinction rather than providing evidence for graded similarity effects across both regulars and irregulars.

Eva Smolka
Who nicks the nickname?—The influence of frequency and semantic transparency on the processing of complex words

Consensus about the processing and representation of complex word formations remains unresolved in psycholinguistic research. How are the meanings of complex word combinations like *verstehen* (‘understand’) and *Standpunkt* (‘standpoint’) stored and processed—as a whole or via the single constituents? Do we therefore access the meaning of *stehen* in the course of processing *verstehen*? The present talk investigates this issue for German word formations with a special view on the influence of lexical frequency and semantic transparency.

I will review the findings of a series of behavioral and electrophysiological experiments that examined the different degrees of morphological complexity by using different types of meaning units, as exemplified here in increasing order: stems like *steh* and *stand* in verb inflections like *stehen* (‘stand’) and *gestanden* (‘stood’) respectively, stems like *stehen* (‘stand’) in prefix and particle verbs like *verstehen* (‘understand’) and *anstehen* (‘stand in line’), and stems in whole-word combinations as in *Standpunkt* (‘standpoint’).

Semantic association tests, intra-modal and cross-modal priming with lexical decision tasks gauged the degree to which the stems are processed and represented, irrespective of whether they are considered to be ‘regular’ or ‘irregular’, and irrespective of the meaning of the whole formation. The behavioral and electrophysiological findings indicate that processing of complex word formations necessarily entails activation of the stem. The present work presents a “stem-based frequency” model—an account that integrates these new findings for the mental lexicon in German.

Mirjana Bozić
Morphological functions in their neurobiological context

Language comprehension engages functionally distinct large-scale brain networks in both hemispheres. I will present a series of studies investigating how this neural architecture supports the processing of inflectional and derivational complexity. Across languages, our results suggest that the two types of morphological complexity differently engage the language processing network: the processing of regular inflectional complexity selectively activates the left-lateralised peryslvian system associated with combinatorial processing of grammatically complex forms. In contrast, derivational complexity primarily engages a distributed bilateral system, argued to support general perceptual and semantic interpretation of whole words. These bilateral effects are however significantly modulated by the semantic compositionality of derived words, and show a degree of cross-linguistic variation. I will discuss the implications of these findings for the theories of processing and representation of morphologically complex words.

Abstracts (Day II)

Tal Linzen
Information and representations in the neurobiology of morphology

Probabilistic language processing is occasionally seen as incompatible with the abstract representations favored by formal linguists. This is a false dichotomy: it is increasingly clear that humans track the frequency of linguistic elements across the entire spectrum of levels of representation, ranging from phonemes to syntactic structures. This detailed sensitivity to probabilities allows listeners and readers to form accurate predictions about upcoming linguistic material. Far from being incompatible with abstract representations, then, probabilistic prediction can be leveraged to investigate precisely what the mental representations are whose probability we track.

The process of forming and evaluating predictions is often characterized using two quantities drawn from information theory: surprisal, which is related to the predictability of the current linguistic element; and entropy, which quantifies the uncertainty over the probabilistic hypotheses that are currently being considered. In my talk, I will present a series of MEG studies that apply this framework to investigating the processing of morphologically complex words. These studies demonstrate that listeners make use of morphology over and above sequential phonological information to make predictions during auditory word recognition. At the same time, our findings do not provide evidence that morphology serves as a intermediate level of representation between the lexicon and the syntax; rather, morphological and syntactic processes show a similar neural and functional profile.

Adam Albright
Pair-finding, segmentation, and morpheme convergence

Grammatical models of morphological productivity typically conceive of productive affixation as a function that takes an input (a derivational base), and produces an affixed form as its output. Under this conception, morphology has information about properties of the base form, such as its phonological form, syntactic category, and meaning, and determines properties of the affixed form, such as the derived category and meaning. From the point of view of learners, who must reverse engineer the morphological processes and their productivity from linguistic data, this effectively requires finding pairs of forms (base, derived), so that shared properties can be extracted, competition can be assessed, and properly restricted rules can be formulated. In other words, this type of learning appears to require a supervised model, in which the learner is given pairs of forms that have been predetermined to be morphologically related, and labeled with syntactic and semantic properties. This is obviously an unrealistic requirement for natural language learners, who have incomplete knowledge of the lexicon and properties of words; in fact, learning some morphology would be a useful step in decoding the category and meaning of words. To this end, numerous unsupervised models of morphological segmentation have been developed using distributional techniques such as MDL or Bayesian inference; however, these models focus almost exclusively on segmenting words into morphemes, and very little on establishing productivity or selectional restrictions.

In this talk, I report on the results of on-going project to apply a supervised morphological learning model to data with incomplete or missing information about pairs, with the goal of discovering the pairs and their labeling simultaneously. I first review the learning model that I assume (the Minimal Generalization Learner; Albright and Hayes 2003), and show how it learns when given omnisciently labeled data. I then show how the assumption of labeled pairs can be relaxed, given some loose pair-finding heuristics that can create hypotheses about potential pairs. In the complete absence of information about syntactic or semantic properties, the model is very similar to other distributional learners, discovering recurring affixal pieces, but it is also able to learn phonological contexts in which the pieces occur most readily. This model is useful as a bootstrapping model: the learner can discover those pairs that are most likely to be morphologically related, which in turn helps focus attention on those aspects of function and meaning that are most likely to be associated with the affix. However, the model also makes an interesting and novel prediction about the morphological grammar that is learned at early stages: it frequently favors rules that represent several different morphological processes that happen to be phonologically similar or identical. In other words, affixes receive a productivity advantage if they are homophonous with other affixes. I conjecture that this factor may be responsible for the high degree of morphological convergence that one typically sees in morphological systems: many functionally and etymologically distinct affixes end up with the same phonological form, reused throughout the system.

Péter Rácz

Learning the English past tense with robots

The linguistics of human-machine interactions is poorly understood. Do we treat a speaking computer the same way as a human conversation partner? If not, what are the specific differences in our stance and framing, and how do these manifest in our perception and processing of linguistic information? These are important questions not only because of the increasing ubiquity of devices with verbal interfaces but also because simple versions of such interfaces are used extensively in linguistic research. As linguists, we use artificial languages, artificial interlocutors, and artificial settings to emulate those aspects of a linguistic interaction that we want to study.

In this talk, I look at three extreme examples of such interactions; one, in which we compare the effect of a robot peer group to a human peer group in a convergence study, one, in which the participant engages in a language game with the computer, and one, in which the distinction between a computer opponent and a human interlocutor is emphasised. The results suggest that humans do not treat talking machines the exact same way as human conversation partners. And yet the distinction is not clear-cut – many aspects of our linguistic behaviour hinge more on the situation as we perceive it than on whether the conversation partner is carbon- or silicon-based.

Harald Clahsen
What language processing reveals about constraints on word formation

Morphological systems are constrained in how inflectional, derivational, and compounding processes may interact with each other. Derivational suffixes, for example, typically appear inside inflectional ones indicating that derivation can feed inflection and not vice versa. One case that has been widely studied in the psycholinguistic literature is the avoidance of plurals inside compounds in English and other languages, the so-called plurals-in-compounds effect. Compounds with singular non-head nouns (*mouse eater*) typically sound better than those with plurals, and should the non-head appear in plural form, regular plurals generally sound worse than irregular ones (**rats eater* vs. *mice eater*). Several previous studies have shown that both adult and child speakers are sensitive to this contrast, but the question of how this contrast is to be interpreted has remained controversial.

My presentation will review findings from a number of experimental studies on the plurals-in-compounds effect in English and German. We will consider results from (i) different modalities (production, judgment, comprehension), (ii) different experimental techniques (offline studies, online techniques, e.g. eye-movement monitoring during and reading and listening, event-related brain potentials) and (iii) different populations (children and adults, native and non-native speakers), and it will be shown that the contrast between regular and irregular plural non-heads inside compounds is remarkably consistent across (i) to (iii). I conclude that the results can best be understood in terms of morphological and semantic constraints on word-formation processes that become operative at different points in time during processing. Alternative proposals sans grammar that attribute the plurals-in-compounds effect to surface-form properties or to exposure-based learning will be shown to be less successful.

Abstracts (Day III)

João Veríssimo
Generalisation and processing of 'pure morphology': Evidence for two systems

The 'classical' approach to morphology ascribes productivity to knowledge of rules: categorical, context-free operations which create structured representations. Alternatively, within analogical, connectionist, and stochastic approaches, it has been proposed that the mechanisms that generalise and process complex forms are inherently graded, as well as frequency- and similarity-sensitive. In this talk, I will review work conducted in Romance and Semitic languages aimed at adjudicating between these two broad theoretical positions.

We have made use of a number of experimental techniques (elicited production and judgement tasks, masked and cross-modal priming, and computational simulations) to examine the generalisation and processing of verbal conjugation classes, by both native and non-native speakers, in three different languages (Portuguese, Italian, Hebrew). Conjugation classes are instances of 'pure morphology', abstract features that do not express meaning or syntax beyond their morphological properties. As such, this phenomenon is particularly suited to examine knowledge of morphology beyond the sound-to-meaning mappings that characterise inflectional and derivational morphemes.

The results suggest that native speakers partition the space of conjugation classes by distinguishing between: i) a default class, which generalises in a context-free manner and forms structured stems; and ii) 'exceptional' classes, which are generalised in a graded manner on the basis of phonological similarity and form stems that are not internally structured. In contrast, non-native speakers do not make a principled distinction between conjugations classes, in that they generalise and process forms of both productive and non-productive classes via a single similarity-based mechanism.

Carmen Pietropaolo
Modelling and testing the productivity of morphophonological rules: The case of Italian mood

The study of morphological productivity and processing has sought answers to the question of whether and how abstract representations of morphological rules are formed. Usage-based models have highlighted the idea that every instance of processing is a concurrent instance of learning. Speakers' individual experience with complex forms influences the mechanisms according to which they are generalized and processed. Knowledge of grammar is probabilistic and emerges from associations made among words related at different levels of representation from phonetic, to structural and semantic. Importantly, it is influenced by frequency.

There is a lack of consensus on the frequency measures that speakers capitalize on when generalizing complex forms. Type frequency is the only important consideration according to Bybee's network model, while token frequency also improves the productivity of morphological patterns according to connectionist models. In my talk, I will explore the generalization patterns of verbal conjugation classes by native speakers of Italian. The work presented here compares verbal forms, such as the infinitive and the subjunctive, which exhibit different distributional properties in language use. These are shown to impact generalizations speakers make, such as the identification of a default morphological class, which may be a mood-specific phenomenon.