Conference

Data-Rich Approaches to English Morphology: From corpora and experiments to theory and back

Victoria University of Wellington
Wellington, New Zealand
July 4-6, 2012
## Timetable (at a glance)

45 minutes per paper (30 mins presentation/ 15 mins discussion)

<table>
<thead>
<tr>
<th>Day 1</th>
<th>Wednesday</th>
<th>Day 2</th>
<th>Thursday</th>
<th>Day 3</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.20</td>
<td>Opening</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.30</td>
<td>Gagné &amp; Spalding</td>
<td>9.30</td>
<td>Beliaeva</td>
<td>9.30</td>
<td>Lieber</td>
</tr>
<tr>
<td>10.15</td>
<td>Bell</td>
<td>10.15</td>
<td>Arndt-Lappe &amp; Plag</td>
<td>10.15</td>
<td>Bauer</td>
</tr>
<tr>
<td>11.00</td>
<td>Break</td>
<td>11.00</td>
<td>break</td>
<td>11.00</td>
<td>break</td>
</tr>
<tr>
<td>11.30</td>
<td>TUTORIAL</td>
<td>11.30</td>
<td>TUTORIAL</td>
<td>11.30</td>
<td>TUTORIAL</td>
</tr>
<tr>
<td></td>
<td>Albright</td>
<td></td>
<td>Keuleers</td>
<td></td>
<td>Kuperman</td>
</tr>
<tr>
<td>12.15</td>
<td>Lunch</td>
<td></td>
<td></td>
<td>12.15</td>
<td>Lunch</td>
</tr>
<tr>
<td>14.00</td>
<td>Tarasova</td>
<td></td>
<td>free afternoon</td>
<td>14.00</td>
<td>Albright</td>
</tr>
<tr>
<td>14.45</td>
<td>Kuperman</td>
<td></td>
<td></td>
<td>14.45</td>
<td>Nagano</td>
</tr>
<tr>
<td>15.30</td>
<td>Break</td>
<td></td>
<td></td>
<td>15.30</td>
<td>break</td>
</tr>
<tr>
<td>16.00</td>
<td>Keuleers</td>
<td></td>
<td></td>
<td>16.00</td>
<td>Hay</td>
</tr>
</tbody>
</table>

### Tutorials

**Adam Albright**  
The Minimal Generalization Learner

**Emmanuel Keuleers**  
Memory-based Learning

**Victor Kuperman**  
Eye-tracking in morphological research
Timetable (long version)

Wednesday, July 4, 2012

9.20 Opening

9.30 Christina L. Gagné & Thomas L. Spalding, University of Alberta
Conceptual productivity: The role of relational meaning construction in the comprehension of novel and established noun compounds

10.15 Melanie Bell & Sabine Arndt-Lappe, Anglia Ruskin University & Universität Siegen
English compound stress in an analogical model of word formation

11.00 Break

11.30 Tutorial
Adam Albright, MIT
The Minimal Generalization Learner

12.15 Lunch

14.00 Elizaveta Tarasova, Victoria University of Wellington
Position-bound and/or relation-bound? The productivity of nouns as constituents of compounds.

14.45 Victor Kuperman, McMaster University
Spelling of English noun-noun compounds and its implications for word production and recognition

15.30 Break

16.00 Emmanuel Keuleers, Ghent University
From data-driven linguists to data-driven models: Some issues in building computational models of inflectional morphology

Thursday, July 5, 2012

9.30 Natalia Beliaeva, Victoria University of Wellington
The power of slanguage: Conceptual integration on the word formation level

10.15 Sabine Arndt-Lappe & Ingo Plag, Universität Siegen
Phonological variability in English blends

11.00 Break

11.30 Tutorial
Emmanuel Keuleers, Ghent University
Memory-based Learning

free afternoon (Excursion)
Friday, July 6, 2012

9.30  Rochelle Lieber, University of New Hampshire
      The case of the missing nominalization: an exploration of how far data-rich approaches can take us

10.15  Laurie Bauer, Victoria University of Wellington
       Grammaticality, acceptability and the notion of possible word

11.00  Break

11.30  Tutorial
       Victor Kuperman, McMaster University
       Eye-tracking in morphological research

14.00  Adam Albright, MIT
       Measuring the unobservable: quantifying paradigm gaps and morphological retreat

14.45  Akiko Nagano, University of Tsukuba
       Doing morphology with the OED: A data-rich approach to English affixation.

15.30  Break

16.00  Jen Hay, University of Canterbury
       Spoken Morphology: Investigating the implementation of plural /s/ in a large corpus of historical New Zealand English
Beyond linguistics

Hotel
Trinity Hotel: http://www.wellingtonnz.com/accommodation/trinity_hotel_wellington

Pre-conference get-together
Tuesday, 7 p.m., at Bistro 169, right opposite the hotel, on the corner of Willis and Dixon

Conference dinner
Wednesday, 7 p.m. (venue to be specified)

Excursion (Thursday afternoon)
Weather permitting, one option is a nice walk (about two hours) through the green hills with great views of Wellington and Wellington harbor (see below), ending in the Botanical Garden. Bring good shoes, the hiking trails can be muddy.

Useful links

Public transport in Wellington (bus routes, timetables, etc.): http://www.metlink.org.nz

Weather forecast: http://www.metservice.com

Tourist information
http://www.wellingtonnz.com
http://www.doc.govt.nz/
Abstracts
(in alphabetical order of author’s last name)

Measuring the unobservable: quantifying paradigm gaps and morphological retreat
Adam Albright, MIT

Observing the non-occurrence of ‘expected’ forms is a well-known problem for language acquisition and corpus studies (the problem of negative evidence). In morphology, expected forms may be missing for a variety of reasons, including paradigms gaps, selectional restrictions, and decreasing productivity of an affix. For example, for many speakers of American English, the verb ‘stride’ has a paradigm gap in the past participle (‘He has *stridden/*stroden/*strode’). Synchronically, this fact is reflected in the low token frequency of (etymologically expected) ‘stridden’ relative to other forms (e.g., present ‘stride’, past ‘strode’); however, merely observing low token frequency does not prove that stridden is rarer than expected, let alone explain why it is rare. In this talk, I consider how diachronic data can help document the existence of paradigm gaps, and offer insight into how gaps and selectional restrictions arise. I show that we can observe at least two different sources of restrictions: (1) failure to generalize an affix beyond its original domain of application, and (2) ‘retreat’ of an affix to a smaller phonological domain. I argue that both patterns are predicted by a conservative inductive learning model, such as the Minimal Generalization Learner (Albright and Hayes 2003).

The first goal of the talk is empirical: using data from the Corpus of Historical American English (COHA; Davies 2011) and the Google n-grams corpus (Michel et al. 2011), I show that it is possible to observe the gradual erosion of previously attested forms such as ‘stridden’. A basic obstacle to observing underattestation is that we must estimate the expected frequency of affixed forms. In English, the token frequency of a verb’s past participle is related to the verb’s lemma frequency, but it depends on many additional factors as well (argument structure, collocations, real world factors), some of which are difficult to estimate. However, these factors appear to be relatively stable across time, such that in general, the relative frequency of the participle of a given verb tends to remain fairly constant. Fig 1 shows this for the proportion of ‘sung’ (etc.) tokens relative to present/infinitive ‘sing’ in the American English Google n-grams corpus. Not all verbs show such constancy, however. In particular, many of the verbs in the ‘stride’ class show a gradual decrease in the ratio of participle forms over the past 100 years, shown in Fig. 2 (normalized for lemma frequency). A loglinear model confirms that participles of this class have been waning over time, affecting especially (but not only) low frequency verbs like ‘stride’ and ‘strive’.
The next question, of course, is why certain forms cease to be used. In previous work (Albright 2009), I have argued that the *stridden gap is due to the sparseness of data concerning participles of verbs with [oo] pasts: there are not many such verbs, and those that do exist are limited to specific phonological contexts, and are split between [i] participles (wrote, drove, rose) and [oo] participles (wove, froze) (competition). In order to explore the relative contribution of type frequency, phonological consistency, and competition from other processes, I turn to English derivational suffixes, which provide a larger and more diverse set of cases to compare. Data from the Google n-grams corpus and COHA show clear effects of competition and contextual productivity. For example, in the past 40 years, -ness forms have experienced a gradual decline in frequency after -ive, -ic, -al, etc., under competition from -ity. Other affixes show radically different rates of productivity in different phonological contexts, as predicted by the MGL and many other models of similarity-based generalization. A more difficult challenge is to establish expected frequencies of existing suffixed forms, in order to identify cases of underattestation. I discuss several strategies, along with their relative merits and drawbacks.
Phonological variability in English blends
Sabine Arndt-Lappe & Ingo Plag, Universität Siegen

In spite of a number of studies in this domain (see, for example, Bat-El 2006 for an overview), blending remains a somewhat enigmatic prosodic-morphological process. Examples of English blends are given in (1).

(1) | blend | base word 1 | base word 2 |
---|---|---|---|
brunch | breakfast | lunch |
stagflation | stagnation | inflation |
smog | smoke | fog |

The pertinent literature is strongly divided over the issue of how predictable the structural properties of blends are. Whereas the traditional descriptive literature has tended to stress the variability and, hence, unpredictability of the process (cf. e.g. Marchand 1969), there is a growing number of studies that has postulated constraints on variability (e.g. Kubozono 1990, Kelly 1998, Plag 2003, Gries 2004, Bat-El & Cohen to appear). These constraints pertain to the questions of how many and which segments of the base words survive, when and how segmental overlaps occur, where cut-off points can be located with non-overlapping forms, and how stress and length determine the shape of the output. It is, however, largely unclear, how the phonological variability is principally restricted. In particular, there is to our knowledge no empirical work on English that systematically investigates productive variability for a single set of base words.

The present study sets out to remedy this situation, presenting the results of a production experiment with more than 1,700 observations, in which 30 native speakers of English formed blends on the basis of 60 word pairs which systematically elicited specific constellations of structures.

It turns out that in general, blend structure is surprisingly uniform in terms of which portions of the base survive, the location of cutoff points, and stress assignment, lending support to earlier approaches working on the predictability assumption. However, we also identified areas of systematic variability which call into question some of the existing proposals.

We will quantify the respective types of variation in our data, and discuss implications for a theoretical model. In particular, we will argue that our data are in principle compatible with an optimality-theoretic approach to blend structure. In such an approach, systematic variability can be modelled as variability in ranking that is part of the co-phonology of the morphological process. However, a major challenge for an optimality-theoretic model of blend formation lies in the formalisation of the intricate faithfulness patterns that emerge from an interaction of constraints on length, syllable structure, and stress.

The power of slanguage: Conceptual integration on the word formation level
Natalia Beliaeva, Victoria University of Wellington

The “concept-making power of neologism” (Leech, 1969, p. 44) has been noted by linguists for a long time. In the light of contemporary cognitive linguistics findings it is essential to investigate in what way the phonological and graphical form of a neologism can reveal the
underlying process of conceptual integration. Such creative and attention-catching neologisms as blend words can serve this purpose perfectly. However, the interaction between the form of blends (i.e. the way the words *flirting* and *texting* come into the blend *flexting*) and the conceptualisation of their meanings has never been brought out in linguistic studies despite the vast literature on the phonological (Kelly, 1998; Kubozono, 1990) and semantic (Kemmer, 2003; Lehrer, 2007) properties of blends. The present paper suggests an integrated approach to the problem and aims to analyse a corpus of recently coined (2000-2011) English blends and to investigate the interaction between the meanings of their constituents in relation to their form.

Being a part of a broader research considering blends as a word formation phenomenon, this study focuses on novel blends in order to find out how conceptual integration is realised in them. The material collected from on-line sources such as WordSpy and Urban Dictionary includes a large number of colloquial and slang words (*sinlaws, bragplain*) that reveals a slice of contemporary English operating in everyday life. The selection of the source words for blends tends to be subject to the semantic (i.e. coordinative or determinative) relationships between them. Formally, blends can be approached as consisting of the initial part of the first constituent and either the initial (*digicam*) or the final (*Wikiality*) part of the second constituent. The interaction between the formal and semantic integration of blend constituents is considered.

REFERENCES


**English compound stress in an analogical model of word formation**

Melanie Bell & Sabine Arndt-Lappe, Anglia Ruskin University and Universität Siegen

It is well known that while many English noun-noun compounds have the leftward stress typical of compounds in other Germanic languages, not all English compounds conform to this pattern, and rightward stress is far from exceptional. For example, while *ápple juice, window washer*, and *téabag* usually have stress on the lefthand constituent, *apple pie, glass dóor*, and *car rádio* are normally spoken with stress on the righthand constituent. A number of recent corpus studies have shown that, inter alia, a compound’s semantics and the identities of its constituent nouns, are significant probabilistic predictors of stress placement (Plag et al. 2007, 2008, Bell 2012 on semantics; Plag 2006, 2010, Arndt-Lappe 2011 on constituents). Furthermore, English compound stress is subject to within-type variation, so that many com-
pounds are attested with both right and left stress patterns, although some compounds show more within-type variation than others (Bell 2012, Kunter 2011).

This paper shows that these empirical facts of English compound stress assignment are naturally accounted for by an analogical theory of word-formation, as implemented computationally in the analogical algorithm AM::Parallel (Skousen & Stanford 2007, cf. Skousen 1989 et seq. for the theoretical basis). The evidence for this is based on a set of 1,000 compounds from the British National Corpus, experimentally elicited from multiple speakers and carefully rated for stress placement (Bell 2012). In a series of simulation experiments using this data, AM is extremely successful at predicting compound stress assignment, correctly predicting over 90% of both left and right stresses in the portion of the database that does not show within-type variation. Furthermore, in the database as a whole, AM is highly successful at predicting the extent to which different types are variably stressed.

An analogical theory of word formation assumes that new words are formed on the basis of similar items already in the lexicon. For compound stress assignment, this means that stress in new compounds is assigned on the basis of the stress pattern of similar compounds that have previously been encountered. In the AM model presented here, degree of similarity is calculated on the basis of the compounds’ constituents and semantic properties, and close analysis of the model reveals how these different factors interact in an analogical system. Although both types of information contribute significantly to the predictive power of the model, the nature of their contributions is different. A number of compounds with very specific semantic properties tend to cluster into a large, strong ‘analogical gang’ in the lexicon, influencing stress assignment in new compounds with the same semantic properties, and overriding potential constituent effects. Because a large number of compounds with similar semantics have the same stress behaviour, the result resembles the effect of a categorical rule. For many other compounds, however, stress assignment is based on more local analogues in the lexicon: in these cases, the identities of the constituents are key features in the computation of similarity.

Based on these findings, it is argued that an analogical theory provides a grammatical model in which the major types of effect seen in English compound stress assignment are expected. These include the variability found in stress placement, the rather local effects based on particular constituents (e.g. compounds ending in street are usually left stressed), and the more general semantic effects that resemble abstract rules. In an analogical model, all of these effects arise naturally from a single mechanism.

REFERENCES
Grammaticality, acceptability and the notion of possible word

Laurie Bauer, Victoria University of Wellington

Early Chomskyan grammar distinguished between grammaticality and acceptability in syntax, and it was explicit that the two might not be co-terminous. Acceptability was determined by speakers, grammaticality by the grammatical system. The notion of possible word is in some ways a continuation of the notion of grammaticality as applied in derivational morphology. According to Aronoff (1976), it is the job of a grammar to specify the possible words of a language (so they are grammatical in the Chomskyan sense), but notes that not all actual words can be generated by the grammar (and thus are not possible?). Productivity deals with what happens in the gap between the actual and the possible.

All of this assumes a vision of grammar where there is a determinable outcome from the series of rules which make up the grammar. But more and more, it is being accepted that there are probabilistic aspects to grammar (perhaps specifically in areas such as the prosodic morphology of blends and hypocoristics). The more this is accepted, the less clear the notion of possible word becomes.

Corpus study allows us to attack this thorny problem from the other end. A large corpus gives us a body of ‘actual’ words (for some value of ‘actual’), from which we wish to deduce a system. Some of the actual words we find by this method may appear ungrammatical. Yet it is dangerous to reject corpus evidence on the grounds that it does not fit with a pre-established notion of grammaticality, especially if we wish to acknowledge some potential variability in outputs. At the same time, accepting all attested words on an equal footing may allow speech-errors to count as data.

In this paper I look at the extent of the problem by considering some attested forms and what they tell us about English morphology. The intention is to air the issues rather than to present a definitive solution, but I shall also suggest that the theoretical dangers involved in our practices are not as great as individual examples may suggest.

Conceptual productivity: The role of relational meaning construction in the comprehension of novel and established noun compounds

Christina L. Gagné & Thomas L. Spalding, University of Alberta

At first glance, word compounding seems relatively straightforward: It is the creation of new words through the combination of existing words. However, this process is actually quite
complex in that it involves the coordination of several types of information including lexical, morphological, semantic, and conceptual knowledge (see, for example, Downing, 1977, Gagné, Spalding, Figuerado, & Mullaly, 2009; Levi, 1983, Štekauer, 2005). Similarly, a maximization-of-opportunity approach to the processing of morphologically complex words (Libben, 2006; 2010) suggests that the language system is set up to maximize the opportunity for meaning creation during the processing of complex words, and that it does so by drawing on multiple representations, including both whole-word and constituent representations (see also Ji, Gagné, & Spalding, 2011; Kuperman, Bertram, & Baayen 2010). Although many theories of compound processing assume the existence and activation of decomposed constituent representations, it is not fully known how the decomposed representations are integrated. The research on word compounding suggests that such integration requires information outside the constituent representations themselves, including semantic and conceptual information.

We will begin by presenting a theoretical framework aimed at accounting for meaning construction via relations that link the constituent concepts of complex expressions (Spalding, Gagne, Mullaly, & Ji, 2010), and then will present empirical evidence aimed at testing this framework. Our particular focus will be on the role and nature of relational competition during the processing of both novel modifier-noun phrases (e.g., vegetable cube) and established compounds (e.g., necklace and hogwash). We will discuss empirical evidence for a semantic composition process that involves the use of relational structures during the processing of novel compounds as well as for established compounds (even for semantically opaque established compounds).

Spoken Morphology: Investigating the implementation of plural /s/ in a large corpus of historical New Zealand English
Jen Hay, University of Canterbury

The Origins of New Zealand English corpora (ONZE), based at the University of Canterbury, contain hundreds of hours of time-aligned spontaneous interview data, spanning 150 years - the entire history of New Zealand English. The recordings have been manually aligned at the utterance level, and then force-aligned at the phoneme level. The corpus contains many layers of annotation, and is housed in an interactive platform which enables efficient search, browsing, analysis and export of multi-layered information. The availability of such large, time-aligned corpora enables us to undergo large-scale analysis of morphology-as-it-is-spoken, investigating the interplay of influential global factors (e.g. lexical frequencies of words and component parts), with more local dynamics (e.g. the degree to which particular morphemes are salient in the local discourse). This talk examines the dynamics of ‘spoken morphology’ in a set of over 40,000 tokens of plural /s/ drawn from the ONZE corpora. I examine the range of factors which are predictive of the plural /s/ duration, and consider the data alongside a comparable set of monomorphemic word-final /s/. The analysis provides clear evidence that plural word forms are stored in the lexicon. The phonetic implementation of the plural suggests varying degrees of morphological analysis, and is affected by both global and local factors.
In the past three decades, the study of inflectional morphology has evolved from a mostly descriptive linguistic endeavor to the design of computational models. In a sense, these two extremes can both be seen as data-driven. In the first case, it is the linguists who are data-driven: they use more or less systematic linguistic data to characterize inflectional systems by carving out spaces where different inflectional patterns apply. In the second case, it is the models that are data-driven while it is the linguists' task to design computational models which, given more or less systematic linguistic data, are able to produce behaviorally relevant results.

In this talk, I will outline some issues which I think are important in building data-driven computational models (DDCMs) of inflectional morphology. Some of the most important issues are described below.

1. Similarity metrics: for most DDCMs, the way in which similarity between words in the lexicon are computed is of crucial importance. However, different metrics can give different result. I will give an overview of some currently used metrics, their characteristics, and consequences of their use.

2. Pre-computation vs on-line computation: Do data-driven models require an intermediate representation of knowledge (rules) or can knowledge be used on-line (memory-based)?

3. Deterministic vs inductive mapping: Most DDCMs work by mapping different inflected forms to each other (e.g., singular-plural). Hence, they assume that the different cases are known. Can models be built without using such a mapping and how?

4. Frequency: DDCMs typically do not take into account the token frequency of the forms in the model in tasks such as generating novel inflected forms. Is form frequency necessary to influence inflectional productivity and how should it be used?

5. Input data: DDCMs are crucially dependent on their input data. What are current practices for selecting the data we feed to models and how careful should we be with this data selection.

Spelling of English noun-noun compounds and its implications for word production and recognition
Victor Kuperman, McMaster University

The present study explores linguistic predictors and behavioral implications of the orthographic alternation between a spaced (bell tower), hyphenated (bell-tower) and concatenated (belltower) format observed in English compound words. On the basis of two English corpora (Wikipedia and New York Times), we model the evolution of spelling for compounds undergoing lexicalization, as well as define the set of orthographic, phonological, distributional, and semantic properties of the compound's constituents that co-determine the preference for one of the available realizations. We explore iconicity and economy as competing motivations for both the diachronic change and synchronous preferences in spelling. Observed patterns of written production closely mirror the demands and strategies of recognition of compound words in reading. Orthographic choices that go against the reader's economy of effort come with a high recognition cost, as evidenced in inflated lexical decision and naming latencies to concatenated compounds that occur in other spelling formats. Likewise, ortho-
graphic preferences affect production of compound words in spontaneous speech, such that a stronger bias towards concatenation comes with shorter acoustic realizations. Implications for models of morphological processing are discussed.

The case of the missing nominalization: an exploration of how far data-rich approaches can take us

Rochelle Lieber, University of New Hampshire

English has a wide range of nominalizing processes. We have affixes that form agent and instrument nouns (‑er,‑ant), sentient patient nouns (‑ee), and event/result nominalizations (‑ation,‑ment,‑al,‑ure,‑ing, etc.). We also use conversion to form event/result nominalizations. What is somewhat odd, and only rarely remarked on in the literature (Booij & Lieber 2004, Lieber 2004) is that we lack an affix or process that forms inanimate or minimally-sentient patient nouns, in other words nouns that mean ‘thing or stuff which is Xed’. Other languages have nominalizers that create inanimate patient nouns – what I will henceforth call ‘patient-thing nouns’. The Algonquian language Nishnaabemwin (Valentine 2001, 503-5) does. Surely speakers of English sometimes have need to talk about the thing or stuff which is Xed, but we do not have a dedicated word formation process that creates patient-thing nouns. Morphologists have long studied word formation processes that are well attested. But to my knowledge no one has tried to study a word formation process that logically ought to exist but apparently does not. The question that I raise in this talk is what happens when there is no word formation process? What do native speakers do?

One reason that no one has raised such a question until now is that until now we have had no coherent way of answering it except perhaps by forced production experiments. What I will try to explore in this talk is whether large-scale databases and enormous corpora like COCA will now give us the means to go about studying questions like this in a different way. As corpora like COCA get larger and larger, they offer the possibility of capturing rare word formation events on the fly, and thereby studying what really happens when the real-time need for a patient-thing noun arises. There are several logically possible answers. One possibility is that existing nominalizations are repurposed on the fly, as in the case below:

Outdoor Life 2005: I had taken bears before and had been hunting for several years for a truly outstanding bear, and here one was standing broadside at 20 yards. I didn’t have to think twice about this bear. It was a shooter.

If so, this leads to a further question of which affixes are repurposed, why those affixes, and how far the prototypical reading of the affix can be stretched. The second possibility is that speakers typically just say something like ‘the thing or stuff Xed’. Finally, it may be that speakers do nothing; they simply make do with circumlocutions.

In this talk I will try to develop a methodology for answering such questions. I will first discuss patient-thing nouns that are item-familiar that I have culled from a database of 3000+ verbs and their nominalizations. From this data-base of 3000+ verbs, I will also cull those verbs that might be expected to have patient-thing nominalizations (specifically, verbs that are transitive and that have objects that tend to be inanimate) but appear not to have
item-familiar ones. I will then discuss various ways of searching COCA for the missing nominalizations and the outcome of these searches. The paper will report on the current status of this work-in-progress.

Doing morphology with the OED: A data-rich approach to English affixation.
Akiko Nagano, University of Tsukuba

I believe that one of the main aims of this conference is to examine how far the bi-directional assessment between data and theory is possible. That is, by moving back and forth between a language corpus and a theoretical hypothesis, we want to calculate both the theoretical profitability of the former and the empirical validity of the latter. In this presentation, I will try to do such bi-directional assessment by using the *Oxford English Dictionary Second Edition on CD-ROM* and *Oxford English Dictionary Online* as my data sets. General usefulness and problems of the OED in linguistics (diachronic linguistics in particular) have already been pointed out in several studies, but my aim is to examine the profitability of the OED in more concrete terms, focusing on its possibility as a tool for morphological investigation.

The theoretical hypothesis I will take up in this assessment is the semantic, output-oriented approach to derivational morphology proposed in Plag (1999, 2004), Lieber (2004, 2006), and Trips (2009), among others. The main claim of this approach is that the input category specification is unnecessary in derivational morphological rules because it is predictable from the output semantics. Assessing the validity of this hypothesis is the second aim of my presentation. I will examine this hypothesis using the OED data of a number of English derivational affixes, including (i) -ness and -ity, (ii) de- and un-, (iii) out- and over-, (iv) pre- and post-, and (v) “zero” affix. Careful examination of these affixes will show that there are certain empirical and theoretical problems with the hypothesis. Also, using the OED as a corpus makes it possible for us to discuss the semantic, output-oriented approach to English affixation from a diachronic point of view. That is, does this approach provide an accurate prediction for the diachronic changes of category-selection properties of derivational affixes? For this question, I will look at the development of the derivational suffixes of the form -ful and what the OED tells us about it.

REFERENCES
Position-bound and/or relation-bound? The productivity of nouns as constituents of compounds
Elizaveta Tarasova, Victoria University of Wellington

Nominal compounding is often taken to be the most productive method of word-formation in English. In the case of N+N compounding, free lexemes are utilized as building blocks for creating new senses. Although the productivity of the compounding process is not in doubt, the productivity of the elements is yet to be considered. One recent study of lexicalised modifier-head compounds in English (Baayen 2010) shows that in the majority of cases the constituents of lexicalised compounds are position-bound, i.e. they are used either as a head or as a modifier. The current study looks at this issue on a larger scale, investigating the hypothesis that the same trend holds true for elements of non-lexicalized items. Another hypothesis tested in the presented research is whether morphological productivity is connected with semantic productivity, i.e. the productive use of the element in one position may correlate with the productive use of a limited number of semantic relations that connect the elements. Levi’s (1978) set of semantic primitives was used for the analysis.

The study is based on the corpus of 197 head and modifier compound word families for 100 key nouns extracted from the BNC (Davies 2004-). The study demonstrates that productivity of an element in head position is negatively correlated with the productivity of the same element in modifier position, though the degree of the negative correlation is variable. The analysis of the semantic relations within the analysed compound word families also demonstrates a strong effect of the family size on semantic productivity. The effect of family size is stronger for the analysed head families than for the modifier families. It is suggested that although head and modifier families demonstrate similar patterns in terms of semantic productivity, the underlying conditions that yield such results are inherently different and are connected with the roles that the head and modifier nouns play in the formation of compound nouns.

REFERENCES
Davies, M. 2004-. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.